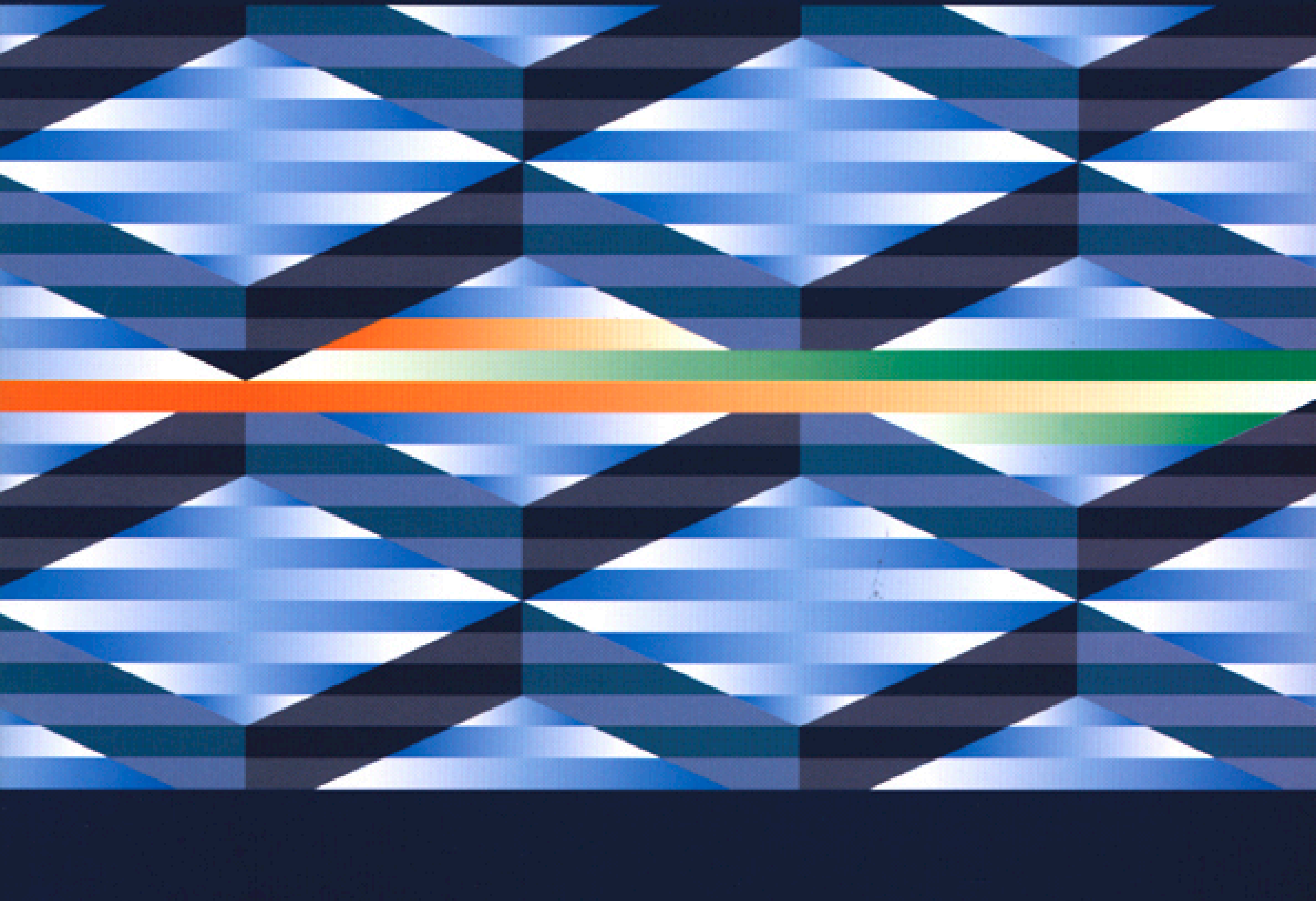
Telektronikk 1.97

Quality of Service in telecommunication



Contents

Editorial, *Ola Espvik* 1

Quality of Service / Network Performance analysis – an integrated part of the network design, *Ola Espvik* 2

Elements of a QoS measurement framework for bearer services, *Bjarne E. Helvik* 7

QoS frameworks for open distributed processing systems, *Finn Arve Aagesen* 26

QoS specification of ODP binding objects, *Arnaud Février, Elie Najm, Guy Leduc and Luc Léonard* 42

A user's perspective on Quality of Service activities in world telecommunications, *Albert Lee* ... 50

Network interworking concepts for the competitive multiple service deliverer environment, *Peter Gerrand* 56

Measuring Quality of Service in the public switched telephone network, *Eddie Ulriksen* 59

Measurement-based software process improvement, *Tore Dybå and Øystein Skogstad* 73

Tool based in-process software quality analysis, *Syed R. Ali* 83

EIRUS – a user group for quality measurements, *Rolf Johansson* 86

Why EIRUS? – 5 years quality assurance by Betatest NKÜ 2000, *Helmut Imlau* 89

Glass-BOX – more than a transparent black box, *Chris Aldenhuijsen* 96

Review of international activities on software process improvement, *Øystein Skogstad* 101

The Error Propagation Phenomenon – an introduction, *Bjarne E. Helvik* 109

Identification of operational modes of distributed systems by cluster analysis, *Bjarne E. Helvik and Sven Arne Gylterud* 118

Plant and people – the influence of human factors in telecommunications network performance, *John Mellis* 128

Dependability in telecommunication networks, *Stein Hagen and Kjell E. Sterten* 136

A traffic model for calculating the effect of reduced circuit capacity on the availability of the ISDN/ telephone network, *Arne Østlie* 156

QoS differentiation in ATM networks – a case study, *Bjarne E. Helvik and Norvald Stol* ... 161

ATM network performance parameters and methodology for deriving their objectives, *Inge Svinnset* 169

EURESCOM and QoS/NP related projects, *Henk Groen, Amardeo Sarma, Tor Jansen and Ola Espvik* 184

A presentation of the authors 196

Telenor

Telektronikk

Volume 93 No. 1 - 1997
ISSN 0085-7130

Editor:
Ola Espvik
Tel. + 47 63 84 88 83

Status section editor:
Per Hjalmar Lehne
Tel. + 47 63 84 88 26

Editorial assistant:
Gunhild Luke
Tel. + 47 63 84 86 52

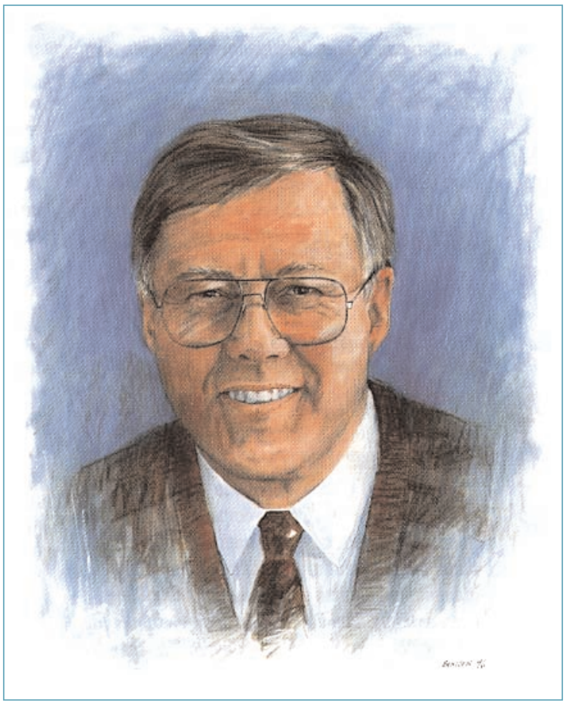
Editorial office:
Telektronikk
Telenor AS, Telenor Research & Development
P.O. Box 83
N-2007 Kjeller, Norway
Telefax: + 47 63 81 00 76
e-mail: telektronikk@fou.telenor.no

Graphic design:
Design Consult AS

Layout and illustrations:
Gunhild Luke, Britt Kjus, Åse Aardal
Telenor Research & Development

Editorial

BY OLA ESPVIK



Using telecommunication, any subscriber can in principle make a connection with any other subscriber no matter their location in the world.

Communication knows no cultural, religious or political divisions, but spans the globe, space oceans, and land masses through cooperation. In war or peace, using diverse equipment and networks, no matter the language, procedures, or age of equipment, the connections are made. We live in an age of social development dependent on a comprehensive and intricate telecommunication infrastructure. We have come to expect telecommunications services to be available when we need them and to perform to a quality level that fulfils our expectations. Few people consider the enormous amount of systems, functions and hardware/software technical devices that have to function properly to make a single connection possible. But things go wrong even if telecommunication equipment is amongst the most reliable in the world.

It is the quality of service (QoS) engineers' job to prevent things going wrong and to design the network and procedures to a necessary robustness that will supply the customer with an adequate service even if some of the parts should fail. Robustness means redundancy, redundancy means more costly systems than might be considered necessary at first glance. And we know that in a competitive world the budget concerned customer will choose the service provider who can deliver the services to the lowest possible price.

Agreeing upon the right QoS is fundamental to finding the right price for both the customer and the service provider. When deciding upon QoS the customer and the service provider alike have to think in terms of risk – what is the probability of something not being up to expectations now and in the future – and what are the consequences. The network operator has to make a delicate balance between what is commercially feasible at the moment and what has to be invested in expected quality demands for the future.

In this feature section we try to educate the reader about valid basic ideas about how to structure the various aspects of quality of service and network performance starting with the ideas over the years having been developed by ITU up to the constructive research results having emerged from the TINA work and project programs, especially within EURESCOM.

Very important to all work on QoS are the opinions of the reflected customers – in this edition appearing as the experience gained from work within INTUG.

We also present a set of papers addressing the performance analysis of SDH and ATM networks. Without data no quality control. Together with an intersting example of how to correlate external data with network performance, we also present the ongoing data collection and processing necessary to ensure the QoS of an operating network.

A particularly interesting set of articles describe the important evolving work of harmonising the various interfaces between the public network operators and their suppliers to ensure a standardised quality understanding of supplied systems and services. Having already been implemented in the US these harmonisation procedures are now being developed to a European version under the support of an increasing number of major suppliers and network operators. In turn a prerequisite for the serious network operators' long chain of actions necessary to provide the customer with a standardised quality handling of services and a trustworthy basis for making a service level agreement at the right price.

Quality of Service / Network Performance analysis – an integrated part of the network design

Highlights from more than two decades of Norwegian research indicating challenges for the future

BY OLA ESPVIK

1 Introduction

Network Performance research, founded in the beginning of this century, has a prominent position in Scandinavia. Erlang's results are well known to everybody working in the area of communication, but important pillars of knowledge have been established by Palm, Engset and others.

At the threshold of the seventies a relatively strong build-up of network performance research started in Norway. First, professor Arne Myskja and his colleagues in Trondheim took up traffic measurements together with studies of subscriber behaviour – later expanding into a wide range of traffic research areas. Then the teletraffic research group headed by Richard Solem at the newly established Norwegian Telecom Research Establishment (NTR) at Kjeller entered many of the same areas – soon becoming a force centre of network performance research cooperating with other research institutions, universities, and industry all over the country. A major reason for that was of course the then monopoly status of the Norwegian Telecom (later Telenor) as an operator – with its money and national research responsibility – together with enthusiastic researchers trying to solve communication problems of the present as well as paving the way for the future.

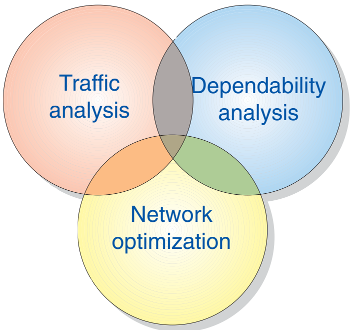


Figure 1 Dependability and traffic analysis must be considered together as a basis for the optimization process

Although anybody with enough money can wire together the latest equipment on the market and start offering communication services, it will always be a competitive advantage of an operator to know the optimal way of doing it. It was therefore completely clear from the very start of organized research at NT – and even more so to-day in a competitive environment – that teletraffic analysis, in traditional combination with dependability analysis and network optimization, by nature has to play a crucial role in the development of any network and the revenue of the operator.

2 The seventies

Entering the seventies we began to see the first fragments of what was later to be called "the information age". The first ideas of what is now known as the Internet matured into a first data network – ARPANET in the USA – at the beginning of the decade – as early as 1975 connected to Kjeller. Discussions of a new kind of network integrating all services began in many places – also at NTR – this network later being established as ISDN in the nineties. It became clear that communication in the years to come had to be digital and that networks and perhaps the rest of the communication activities some day were to be controlled by computers. Suppliers were already developing computer-controlled analogue exchanges, and long-distance transmission could be done via communication satellites. The future was very interesting and open for all kinds of research ideas – however, the present situation was more pragmatic. A few things simply had to be taken care of!

- The trunk network was heavily underdimensioned and severe congestion problems had to be solved
- The Oslo network had both dimensioning and dependability problems
- Data collected for both traffic and maintenance purposes were inadequate and of a very incoherent nature all over the country
- The waiting list for telephone subscriptions was long and totally unacceptable amongst the public.

In other words, the network around 1970 was in a bad shape and new technology was just around the corner. Network Performance research rescue teams were needed immediately!

2.1 Traffic analysis

Very important to successful improvement and further development of the network was to establish a well organized data collection system as well as taking advantage of promising methods now being possible with the new mainframe computer. Based on promising results from military environments and the very advanced usefulness of the programming language SIMULA large resources were invested into building up very near to reality like huge simulation models of the telecommunication network. But alas, the bigger the model, the smaller was the result. Of course! None the less, what we did in those virgin days of simulation was in many ways a success: We got a crash training program in how to do teletraffic simulation – and how not to do it. And it is fair to say that about every major simulation project since then reflects the lessons learned in the early seventies. Within a short time we were able to put SIMULA based simulation and theoretical methods into productive analysis of the Oslo network, the bottlenecks of which little by little were revealed and actions taken to give the subscribers the network they deserved. How well the network performed in practice could still not be quite ascertained due to inadequate measurement methods and technology. Based on the ideas from professor Myskja and his team in Trondheim we therefore launched, together with IBM, what was probably at the time the most ambitious development program of realtime measurements of traffic in operational analogue exchanges. The new equipment gave all the traffic data we wanted. It also included our first solution to controlled corrective maintenance of the network leading up to a nationwide specification of the same.

At the end of the decade the time was ripe for starting the first simulation experiments with traffic control, the possibilities already being demonstrated in computer networks but fully utilised in our telephone network in the next decade.

2.2 Dependability analysis

Our dependability studies in the seventies had a focus on hardware. We established a laboratory for testing hardware components and systems as a quality evaluation of what we received from our suppliers. The big problem, however, was how to specify dependability to the various parts of the total network. An important work was therefore to write a handbook describing how to perform dependability planning, measure and evaluate dependability data, as well as how to specify the dependability of new equipment. The work was complemented by a simulation system addressing dependability of specific network expansions and a method for utilizing split routing. Looking back to the seventies, we have to admit that too little attention was given to theoretical training of our planners giving them the possibility to understand what was being handed over. The result was quite a lot of good research results not being put into practice soon enough.

2.3 Network optimization

Computer development had in the early seventies made it possible for us to address selected areas of network optimization. Programs for doing structural planning in rural areas at a certain point in time ahead were developed as well as programs for optimal logical and physical routing complete with traffic dimensioning of the trunk network. Although limited, these programs were utilized all over the country. At the time we wanted to address more advanced aspects of network planning but had to accept the computer limitations, remaining patient and building up mental power for the next decade with new potent computer generations.

3 The eighties

The eighties were dominated by numerous activities leading up to a fully digital national network. Although NT had not been in the front line during the previous years of European digitization we came relatively strong when we finally got started. A massive study – DIGSTRAT – was started to work out how to construct the new network ending up with the first ITT System 12 exchanges being operative by 1986. At the time a fairly novel construction System 12 was not without complications during the first years. NT therefore decided to include the latest AXEs into the network, thus confirming the balance that had for many years existed between our two major suppliers of switching machines, namely STK (now Alcatel) and EB (now Ericsson).

Another important development was the CCITT Signalling System No. 7, the network of which being studied and first made operational at the end of the decade, thus paving the way for the major challenges now being investigated, namely ISDN and IN.

3.1 Traffic analysis

Traffic investigations played the key role in establishing a sensible traffic solution to all to the brand new technology being offered. ITT System 12 was completely new and the capacity characteristics so far hardly understood even by the supplier. In the midst of the turmoil of the first testing and trial installations massive traffic studies mainly based on simulations were undertaken to find out the real capacity characteristics necessary for proper dimensioning of the switching systems complete with measurements to confirm the results as soon as real traffic data became available. The new systems also made possible a long awaited opportunity to optimize traffic flow throughout the network by using advanced traffic control mechanisms. Fortunately, we had invested much money and knowledge into building up a strong simulation capacity in that area around the turn of the decade – the advantage of which now being utilised to decide how the traffic should be routed and controlled in the new network.

In the eighties our researchers established the traffic platform for the ISDN network finally made operational in the nineties. In parallel, capacity of the new signalling network was investigated.

Simulation systems were also developed to establish good traffic performance characteristics of both our DATEX and DATAPAC networks.

3.2 Dependability analysis

The introduction of new switching systems initiated the analysis of the dependability aspects of distributed systems – special concern being given to the spread of errors throughout the network. The nature of the new mobile systems and the first structuring of IN networks also made necessary an increased effort in finding feasible solutions to fault tolerant database systems. The new public data networks being established in the eighties introduced high dependability demands, the analysis results of which were completely reflected in the structure of the operative networks. Spurred by a couple of serious accidents in the network a nationwide dependability plan was established. We also started the first serious discussions about the economic revenue of costly dependability actions. During this period a greater understanding of the quality of service to the end-user emerged internationally and nationally, in the next decade becoming a vital area of concern as monopoly operators transform into competitive companies.

3.3 Network optimization

Network optimization will always have a strong dependency upon computer power. Realizing the technological possibilities at hand and more to come, we combined the digitization studies of the project DIGSTRAT with a large investigation of what was available in the way of proper computer planning tools internationally. Our ambition was to acquire what could be purchased or borrowed, adapt it to our conditions and start our own development program of key modules not being open to us from elsewhere. Our own research in the eighties resulted in program systems optimizing the physical routing and grouping of the PDH based trunk network as well as the subscriber network. When used these network optimizing programs showed very promising results giving investment cost reductions in the range of 10 to 20 % compared to the old manual planning methods. In addition, a network planning computer centre was established at Lillehammer, the objective of which was to take care of new program modules arriving – some of them research prototypes – and make them user friendly to the planner around the country complete with the necessary hands-on training.

Giving our network planners a sound theoretical knowledge had for a long time been neglected. A very ambitious basic training program was therefore established, partly in cooperation with the ITU TETRAPRO project to give our planners the necessary platform for basic theoretical understanding.

4 The nineties and beyond

So far into the competitive nineties, the most obvious features are the extensive use of mobile communication, the ISDN network finally being made operational and producing services at acceptable

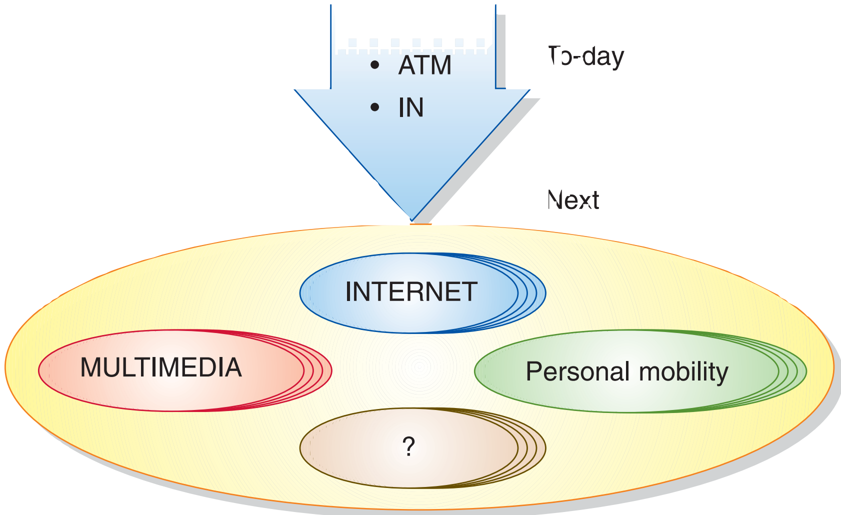


Figure 2 Moving from today's traffic analysis focus on to the next

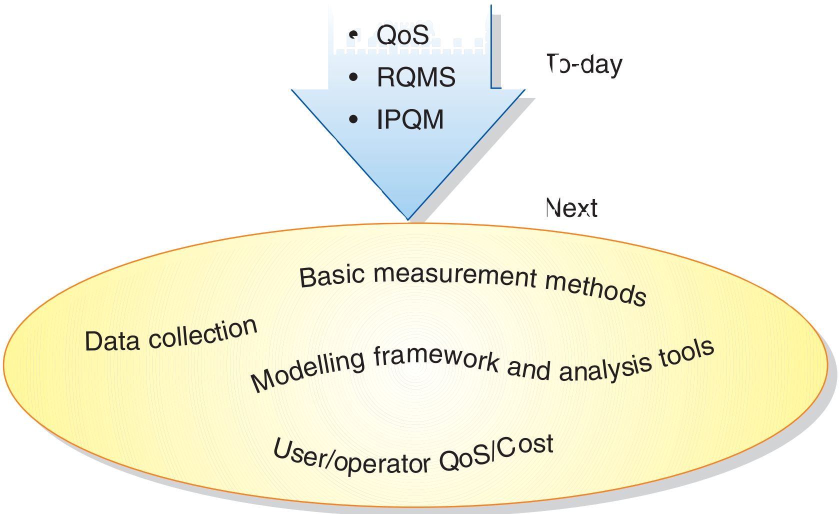


Figure 3 Today's dependability issues and important ones from now on

QoS. Fibre is also being more and more common combined with ring structures in important areas giving a very good basis for reliable communication. Another important challenge is found in the introduction of the first commercial ATM services and the market dominated, almost out of hand use of the Internet.

4.1 Traffic analysis

The coming public B-ISDN has so far been considered to be based on ATM – a fact to a large degree reflecting our research efforts into the nineties. Much of the research has been in cooperation with other institutions and bodies like SINTEF, EURESCOM, RACE, ACTS and COST. Transport protocol analysis and performance measurements of ATM networks have been investigated to determine good protocol parameters finally leading to maximum possible QoS to the end user. A laboratory system has been developed consisting of ATM switches and ATM traffic generators making it possible to experiment on traffic models and traffic characteristics of ATM networks, the results of which already being utilized in our first commercial ATM services. As part of a EURESCOM project the control part of the IN network has also been investigated.

Our ambitions to provide Internet solutions to all our subscribers may represent a fundamental challenge to the network dimensioning. A way of meeting required QoS and hence network performance has to be developed together with solving the obviously never ending problem of having an adequate system for measuring traffic and maintenance condition of the network.

Summing up

ATM has become the principle for establishing flexible user connections in B-ISDN networks. A lot of development work remains, however, in order to achieve good traffic control mechanisms. This will have consequences for both network utilization and quality which will be the responsibility of telecom operators.

Looking into the future we see a demand for communication solutions to networks with more intelligence, subscribers with greater mobility requiring access to a variety of media wherever they are and whatever the time. And specified QoS is all the time expected to be fulfilled. The fate of any serious operator caring about the costs is of a very eternal nature: Establish the adequate traffic distributions and dimensioning rules for these new demands, or combinations of demands, quickly and specify a cost effective network and service solution. With enough time to develop basic methods a lot can be done, however, quickly changing conditions will always be a problem to research intensive disciplines like traffic dimensioning.

4.2 Dependability analysis

In the competitive environment, QoS (Quality of Service) – and hence network dependability – is becoming increasingly important to all serious public network operators. QoS is now studied in all international organizations. So far into the nineties, our activities have mainly been concentrated around projects in EURESCOM where special attention has been given to establishing the first modelling framework being service and technology independent and thus paving the way for future quantitative modelling work. Another aspect is to establish basic methods for measuring QoS in the various networks. Our present research concern is to establish a way of proving the feasibility of measurement methods using our ATM laboratory network. Important to all public operators are the EURESCOM activities adapting to European operators and suppliers the Bellcore developed, and already in use in the USA, RQMS (Dependability and Quality Measurement System) and IPQM (In Process Quality Measurements). These activities have so far resulted in a new organization – EIRUS (The European IPQM and RQMS user group) – with major operators in Europe being members and the activity of which keeping close contact with EURESCOM research.

Investments and operational costs in order to achieve sufficient dependability represents a substantial fraction of the total network investment cost, tentatively 40%. Hence, the ability to perform, in a wide sense, a correct dependability dimensioning is a major factor in our economy.

The new generation of public telecommunication systems has a number of architectural and technological features for service handling, O&M (Operation and Maintenance) and transport, which require a proper dependability dimensioning to ensure an optimal cost performance trade-off. Methodology and tools for doing this have to be further developed.

Increased "centralization" (e.g. larger switches, IN service handling and TMN based O&M, high capacity transmission systems) may invite severe dependability pitfalls and demand a careful design and a proper dimensioning. Network wide logical interdependencies is an element in this picture. The major outages experienced with SSN 7 internationally are indicators of the even worse situations that may arise when all parts of an operator's network become logically tightly coupled by the IN and TMN functionality.

The objective of the QoS provided to the end-user is currently settled with a minor regard to the end-user's valuation of the QoS, i.e. how much the end-user is willing to pay for various levels of QoS. Since the QoS is closely related to the network dependability, it becomes essential to obtain insight into this valuation to set correct dependability dimensioning criteria in a competitive environment.

Most operators are probably still lacking the sufficient data collection and -preparation tools necessary for dependability handling and planning, e.g. provisioning of dimensioning criteria, input of failure rates, restoration times, network optimization, etc.

Summing up

To establish a set of tools necessary for network dependability dimensioning, quantitative relations of how the end-user valuates QoS, etc., a considerable effort is still necessary.

QoS versus cost has to be the big issue to all operators on all services in the years to come. Everybody will profit from a common understanding of both definitions and an accepted way of measuring it, especially the customers. We have to «teach» the customers and ourselves to have the same understanding. Research must be put into establishing basic analysis tools that can easily be adapted to new technological and market situations giving the operator a possibility to find out what is wrong if the QoS level is reduced, and do something about it at optimum cost. The same goes for tools enabling us to design new networks and services to a given QoS at optimal cost. The QoS data collection problem has to be solved. The QoS situation for providers of Internet services is at the moment very troublesome, the solution to which has to be found on an international basis.

4.3 Network optimization

Optimization research in the nineties has had a major focus on the transition from PDH to SDH transmission networks. Program modules have been developed and put into operation addressing the trunk network as well as the access network – in the latter special concern being given to ring structures applying both radio and optical fibres as physical transmission media.

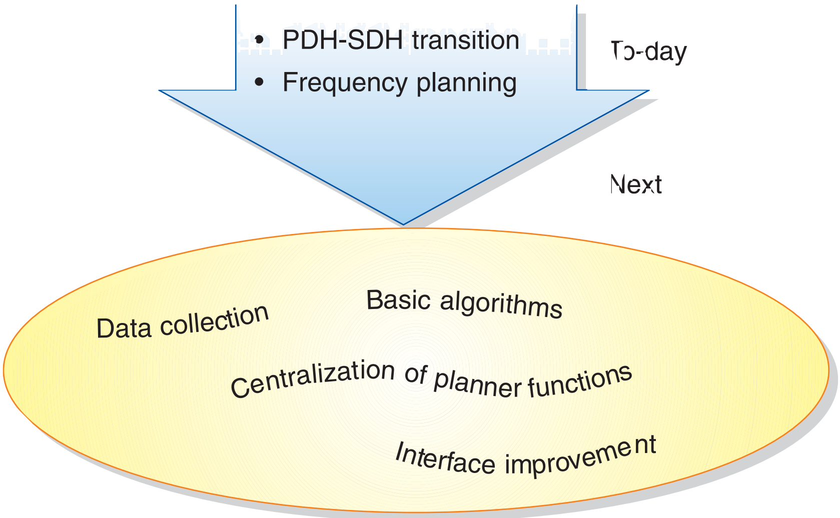


Figure 4 Basic tasks to be started up in our network optimization activities

Both the eighties and the nineties have seen an enormous increase in people's use of mobile telephony. Much attention has therefore been given to find optimal solutions to frequency planning both in the NMT and the GSM networks.

So far, network planning even when assisted by advanced research program tools, has relied heavily on the practical experience and local knowledge of each planner around the country. Into the nineties with organizational restructuring and downsizing, we see a centralization of the planning functions, the result of which is that the network planners have longer distances to practical activities. The planners' "fingerspitzgefühl" may decrease in the years to come but their proficiency in using advanced optimizing tools will surely increase together with their dependency upon basic theoretical knowledge. However, as far as we can see into the competitive future, network optimization tools will hardly ever become on-the-shelf products from the nearby computer store. The network planners will therefore constantly have to adapt to new prototype modules from the research departments adjusting the program modules at hand to ever changing market demands. Thus, to make their job manageable we will have to increase our research efforts into making program systems more and more user friendly.

Summing up

Traffic- and dependability analysis are to some extent inputs to the optimization process. The result of the network optimization process is the foundation of the operator's profit. Market and technology change very rapidly while advanced optimization modelling and result verification nearly always are research intensive and thus time consuming. We want to be ahead of the problems but often see that we address the problem too late. Our ambition is thus to speed up the optimization process in the competitive market. Again, the data collection issue has to be addressed, much money is involved and a dramatic improvement of input data quality has to be done together with improvement of the interfaces both to our databases and the planners themselves.

5 Conclusion

All considerations so far have relied heavily on advanced research. And research takes time – time constants so far are experienced to be seldom less than 3 years, often 5 or 10 years. However, technology and the market have far shorter time constants. What constitutes the operators' profit to-day stems from research ideas going even as far back as a couple of decades.

Based on good research so far we are happy to offer the society a good network with good QoS at moderate cost. The future competitive world is going to be far more complex than to-day with enormous increase in services, multimedia and network solutions. The number of pitfalls and the possibility of losing money are high. However, it is assumed that all serious operators are more determined than ever to continuously being trustworthy toward the customer, giving good QoS at the right price through correct dimensioning and optimal solutions.

To all serious operators – I believe the following is completely clear: No sensible network and service provision without traffic considerations. No satisfied customers without QoS considerations. No satisfied shareholders and operators without economic optimization.

Elements of a QoS Measurement Framework for Bearer Services

BY BJARNE E. HELVIK

1 Introduction

1.1 Objective

This paper discusses a reference model or framework for the measurable properties of the QoS (quality of service) delivered to the end-user. It is an objective that the framework shall, as far as possible, be independent of the specific services and shall enable identification of QoS measurements and measurement methods common to all or a wide range of services. In this context both currently available and future services are considered. The framework shall have a structure and a content which enables it to be a basis for unambiguous measurement definitions. How the end-user receives the QoS, including QoS failures, and eventually triggers an event observable by the service provider or network operator, e.g. a service complaint, is an important aspect. The focus will be on those quality aspects related to the technical performance and characteristics of services, more precisely the aspects encompassed in the serveability performance as defined by ITU-T [1]. (A brief introduction will be given as item E later in the paper.) The framework will be based on previous results towards making generally applicable QoS models/frameworks. Some of these are outlined in separate information boxes.

The paper does not claim to cover all aspects end-user related QoS measurements, and the discussion deals primarily with bearer services.

A End-user opinion/satisfaction

End-user opinion polls and customer satisfaction surveys on the quality of telecommunication services (and the network provider delivering them) is an important approach to get feed-back on the end-user satisfaction [9]. Note that opinion polls and customer satisfaction are fundamentally different:

- *Opinion polls* are where anyone is asked for an opinion
- *Customer satisfaction* is where you ask users shortly after they experience a service about their satisfaction.

It must not be forgotten that quality refers to customer needs and/or expectations. Therefore, the comparison between the quality of services provided by network operators as perceived by them and their perception by the customers is an essential dynamic and ongoing process that must be permanently analysed and compared.

This means that it is important to agree upon recommended QoS indicators between operators and customers. This to establish a higher level of competition, on the operator side, and at the same time offer the customers attractive commercial contracts.

The only “common framework” for QoS end-user investigations known by the author, is the comparable performance indicators (CPI) collected and published by Oftel in UK. See [9], [10] and [11] for further information. It is important in opinion polls and customer surveys related to QoS (eventually concerning a specific service) to reach a common position related to the skeleton of surveys. An important element in this is to agree on a common approach related to measures and analyses of the surveys and hence, to be able to correlate and compare results between operators. As mentioned, such surveys are now ongoing in the UK.

A salient issue is how the concepts and measures used in this kind of framework and corresponding investigations correspond to the QoS concepts and terms in ITU-T E.800 and other standards, and how firm quantitative relationships between opinion poll type of measurements and measurements of the technical QoS parameters are.

1.2 Outline

The next subsection of this introduction gives a brief review of the basic “recursive” quality of service concept which is used for system design and evaluation. The remaining part of the introduction gives a brief review of QoS frameworks which are taken into account when the one presented here is developed.

The remaining chapters discuss the framework itself. The first, Chapter 2, discusses the reference point in the network for the measurements in terms of interface, protocol layer and plane. Next, in Chapter 3, it is discussed how the user experiences QoS. This discussion forms the basis for a simple way to relate measurements of the technical aspects of QoS parameters to observable end-user/customer reactions, for instance in terms of complaints. The composition of bearer services is introduced in Chapter 4. In the next chapter QoS parameters associated with these are introduced, before the result is compared to some of the previous frameworks. Chapter 6 concludes the paper.

1.3 Basic concepts

It is useful to consider the basic concepts concerning a service and the provisioning of this service. There are two entities involved, the user of the service and the provider. The service is characterised by the function it provides to the user and it is described by its primitives. The quality/performance of the service is characterised by its QoS parameters. Figure 1 illustrates this basic relationship.

In the public telecommunications domain, the term user is often implicitly perceived as the end-user (customer, subscriber) and the service as the function(s)

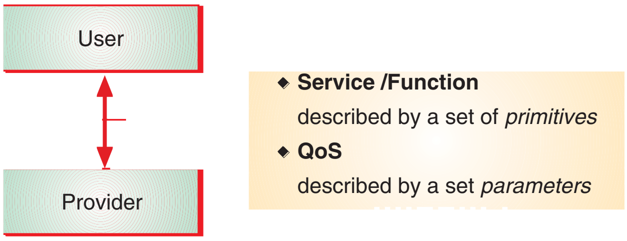

User
Provider
◆ Service /Function
described by a set of *primitives*
◆ QoS
described by a set *parameters*

Figure 1 Generic service relationship

B ETSI framework

ETSI's technical committee on network aspects has prepared the report ETR 003 which covers General aspects of Quality of Service and Network Performance [2]. This report is closely based on the work of the FITCE Study Commission, The Study of Network Performance Considering Customer Requirements [3].

Figure B-1 shows how the various concepts used in these documents interrelate and how these again relate to the various actors in the telecommunication marketplace. For details, see the referred documents.

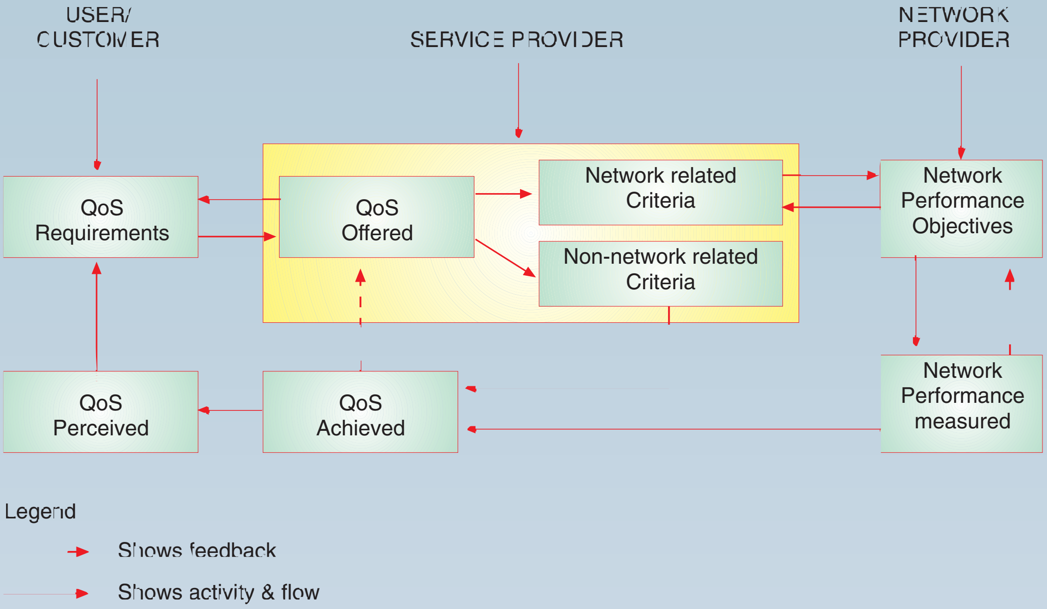


Figure B-1 Inter-relationship between various viewpoints of QoS from [3] and [2]

Basic to this approach is the subdivision into network related QoS criteria and non network related criteria. The network related QoS criteria roughly correspond to the technical QoS parameters dealt with in this project. A mapping between the network related criteria and the network performance parameters is indicated, as shown in Figure B-2, but none is provided.

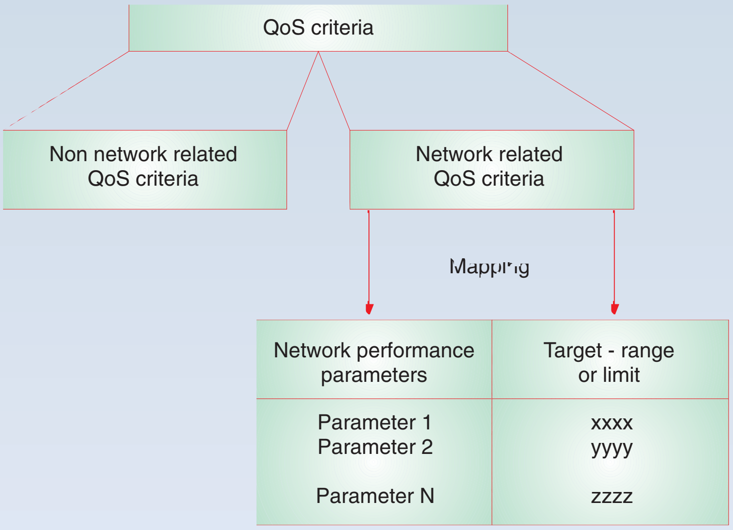

QoS criteria
Non network related QoS criteria
Network related QoS criteria
Mapping
Network performance parameters
Target - range or limit
Parameter 1
Parameter 2
Parameter N
xxxx
yyyy
zzzz

Figure B-2 Relationship between QoS and network performance (NP) from [3] and [2]

In their appendices these documents present a generic matrix capturing some customers' QoS requirements. This matrix is shown below. It is however not discussed how the various elements of the matrix relate to the end-user's perception of the service.

	Speed	Accuracy	Availability	Reliability	Security	Simplicity	Flexibility
Sales							
Service management - provision - alteration - service support - repair - cessation							
Call technical quality - connection/establishment - information transfer - connection release							
Billing							
Network service management by customer							

Within the ETSI framework the *Quality of Service indicators for Open Network Provision (ONP) of voice telephony and Integrated Digital Network (ISDN),* defined in ETSI technical report ETR 138 should also be taken into account [8]. These QoS indicators are listed below. For details, see ETR 138.

Voice telephony

- Fault reports per access line and year
- Unsuccessful call ratio
- Call set up time
- Speech transmission quality
- Supply time for initial network connection
- Response time for operator services
- Availability of card/coin phone
- Fault repair time

ISDN all bearer services

- Fault reports per access line and year
- Severely errored seconds

ISDN circuit switched

- Unsuccessful call ratio
- Call set up time

ISDN circuit mode permanent

ISDN all packet mode bearer services

- Throughput efficiency
- Round trip delay

ISDN packet mode switched

- Unsuccessful call ratio
- Call set up time

ISDN packet mode permanent

- Availability of …
- Number of service interruptions per year

The ETSI framework is used as a basis for the framework developed here. This will be briefly summarised in the main text of the paper. The simplicity of the more specific indicators of ETR 138 is sought (and enhanced with the structure of the services) rather the large number of general requirements of the ETR 003 matrix.

provided to the end-user. Outside the public telecommunications domain, however, the service as well as the QoS concept is used in a more general sense.

Note also that the generic service relationship is used recursively, as illustrated in Figure 2. To avoid confusion it is suggested to use the terms function provider and user when it refers to other relationships than that between end-user and network. An example of this recursive use is given in the outline of the ISO/IEC framework in the information box D.1.

1.4 Existing frameworks and approaches

There have been made several efforts towards making general QoS frameworks. These have been defined for different purposes. None of these have, to the authors knowledge, had as their

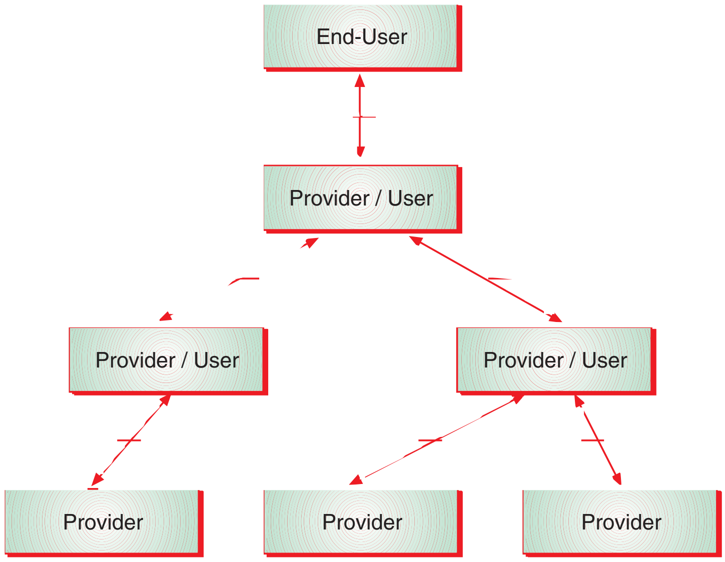


Figure 2 Recursive function (service) user provider relationship

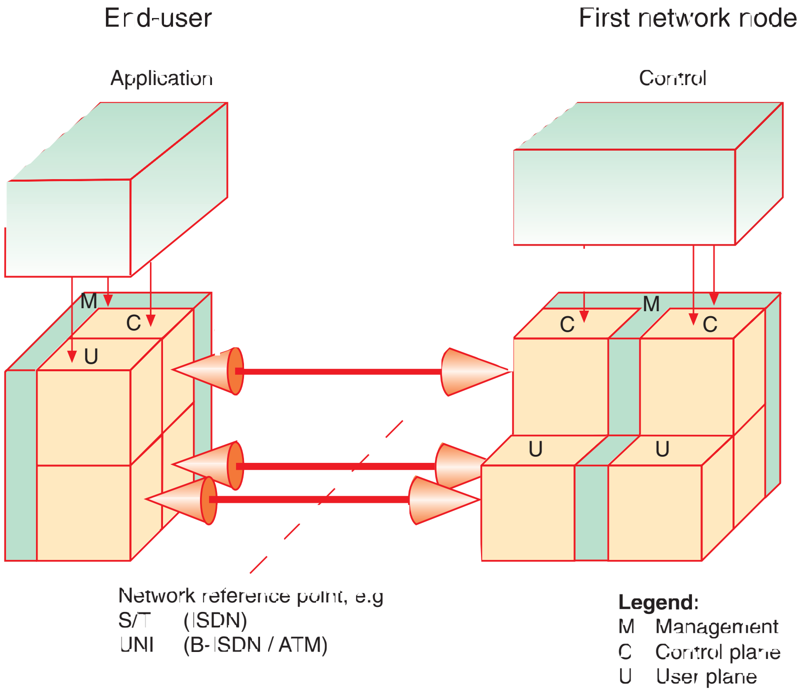


Figure 3 Principal sketch of interface between network and end-user. (Two channels are shown for the low level interface to account for the separate B and D channels in ISDN)

prime objective to form a basis for measurements. It is the objective of the work presented here to base a QoS measurement framework on these previous approaches and to make necessary adaptations and enhancements. For this reason some approaches forming the basis for the subsequent chapters is mentioned.

- *End-User opinion/satisfaction.* This aspect is briefly discussed as a separate item A.
- *ETSI framework,* [2]. This framework, which is based on the work of the FITCE Study Commission [3], is presented as item B.
- The *service failure concept* of EURESCOM P307. The attributes of a service failure is regarded independent of the service it affects and its cause in the network. See item C for details.
- *QoS in layered and distributed architectures* is presented as item D, including:
  - The ISO/OSI QoS framework
  - The Telecommunication Information Networking Architecture Consortium (TINA-C) QoS framework
  - Unification of Frameworks
    For a more thorough discussion on QoS in open distributed processing (ODP) systems, see [4].
- *ETNO Working Group 07/95 on QoS.* This group is working toward a consistently defined set of common European QoS parameters (QoS indicators). The aim is harmonised European QoS definitions and possibly performance targets for pan-European services, in order to facilitate comparison of the results of the measurements. The work is based on the approach of the FITCE Study Commission and ETSI summarized above. The work has hereto concentrated upon voice telephony.

2 Interface between user and network provider

It is a goal to be able to perform precise and unambiguous measurements. Results should be consistent over a number of measurements and measurement sites, and preferably across different bearer services providing the same basic functionality. To achieve this goal, the point where the measurement is taken must be well defined. A precise definition of where a measurement is taken is also necessary if the measurement shall serve within formal context, e.g. a legally binding business agreement between public network operator and end-user.

The requirement that the measurements carried out within the proposed framework shall be able to serve within formal contexts seems to be the most restrictive, and locates the point where the measurement must be taken to the "borderline" between customer and service provider / network operator. Hence, the natural choice of such a measurement point is the interconnection reference point between end-user and network as illustrated in Figure 3. Note that some of the elements in the figure will not be present for some networks/services, e.g. the distinction between a user and control plane for packet switching based on X.25.

QoS measurements should be related to single end-user sites. Measurements averaged over a large number of customers (thousands) and perhaps carried out over a long period have a poor ability to indicate the (dis)satisfaction of single users with the service. For instance, the unavailability of access to the network during a year may be acceptably low (e.g. one hour per year) if all end-users in the measured group (let us say 100,000) experience approximately the same unavailability. However, if only 100 of these have experienced a failure during the year, these receive a low availability – and the QoS may be unacceptable.

2.1 Connection reference point

Measurements should be taken on or unambiguously be connected to a reference point defining the interconnection between network and end-users. For example, UNI for ATM or the S/T interface for ISDN. The actual interface will depend on the end-user interconnection.

The end-user's terminal equipment and his eventual CPN (customer premises network) may influence his perception of the QoS. The performance of this equipment is the responsibility of the end-user or the A and B party end users, and cannot be taken into account in measurements of the QoS delivered by the public network.

C The service failure concept from EURESCOM P307

One of the objectives of EURESCOM project P307 was to determine the consequence of service failures for the individual end-users. The consequence will obviously depend on the end-user considered and the service repertoire he uses. A means toward this objective was to be able to describe how network failures affect the interface between the network and the end-user, irrespective of the services he uses/subscribes to. This approach is described in Chapter 3 and detailed in Appendix A of [7]. By obtaining results which are, as far as possible, independent of the service, it is feasible to apply the results to a range of existing services as well as new ones which will emerge in the years to come. The same is the case for measurement methods.

In P307, failures of the services delivered to an end-user are considered. The approach is generic for simplicity and to ensure usefulness in the future, as mentioned above. Five attributes describing a specific failure, i.e. how the failure is experienced by the end-user, are defined. A failure is a change, beyond given limits, in the value of at least one QoS parameter essential for the service and which will result in a (temporary) inability of the service to be provided to the end-user with a specified quality. The following requirements/objectives are set for the definition of these failure *attributes:*

i. There should be as few attributes as possible.

ii. The attributes should not describe partially overlapping characteristics of the failure.

iii. The attributes should be precise, descriptive and easily understood.

iv. The attributes should be service independent.

The attributes identified are listed below. It is seen that there is an underlying assumption of a connection oriented service. See Section 5.1 of the main text. For semi-permanent and connectionless services some of the attributes do not apply. The same is the case for non-bearer services handled entirely in the event dimension (cf. Item D.3)

Two important concepts from this approach will be used for the framework presented in the paper:

I. *Service independence* in the measurement techniques and measurement reference points.

II. The concept of a *service failures*, which causes end-user dissatisfaction with the QoS and may trigger complaints may be related to opinion poll types of measurements.

Service type	Quality measure
Delay sensitive digital (Interactive work towards remote comp.)	Delay
Digital volume (File transfer)	Throughput
Digital real time (Speech)	Bit error rate
Analogue (Speech)	Articulation Index [12]

- **Duration**, τ

  τ describes the duration of an end-user service failure.

- **Connection attempt rejection rate**, r

  describes the likelihood that a connection to or from the end-user cannot be established.

- **Transfer quality impairment**, θ

  θ is based on the primary transfer quality measure(s) of the actual type of service. For instance as in the table above.

- **Established connections lost**, c

  c is a binary variable. If c is equal to one (1) the established connections are lost at the beginning of or during an end-user service failure period, otherwise c is zero (0).

- **Impact**, ι

  The impact is a measure of how much of the end-user capabilities are affected by a failure, for instance in terms of relative number of connections and substitutable services.

What may be done to relate the measured QoS at the reference point to the subjectively perceived quality at the man-machine interface? The alternatives are to:

- Assume the CPN, User terminal and MMI to be ideal (no quality reduction), or

D QoS in layered and distributed architectures

This issue is presented in three subsections. The first two summarising the ISO and TINA-C approaches respectively, and a third considering a unification of the ITU-T framework, which is assumed to be well known to the reader, with these two.

D.1 The ISO/OSI QoS framework

The Basic Reference Model for Open System Interconnection (ISO's OSI model) defined in ITU recommendations X.200 and ISO/IEC 7498-1 provides a description of a model and the activities necessary for the interworking of systems by using a communication medium. The QoS framework provided in [5] is a supplement to the description of QoS of the Basic Reference Model. For detailed information on this framework, it is referred to the original document or the summary in any of [13], [14] or [15]. However, as a basis for the discussion in the main text, the following items should be pointed out:

- There is a set of ***generic QoS characteristics (parameters)***. A summary of these are given in the following. The name, definition and, in most cases, quantification are contained:
  - *Time delay* is the QoS characteristic that represents the difference in time between two related events. The Time delay characteristic is quantified in any units of time such as seconds, milliseconds, etc.
  - *Time variability* is the QoS characteristic representing the variability of a time or a time period. It relates to the dispersion or jitter that can be tolerated when time periods are involved. The Time variability characteristic is quantified as an amount of time or as a probability of variation from a time or a time period. It may be described by a number of mathematical descriptors, e.g. upper and lower bounds, variance or percentile bounds.
  - *Time window* is the characteristic representing a specific period of time. It is a bounded time interval which is defined by a starting time and a time delay, by a starting and an end time, or by a time delay and an end time. The Time window characteristic is quantified by providing either an absolute time (start or end time) plus a time interval or two absolute times. These are expressed in any units of time.
  - *The capacity* characteristic represents the amount of service that can be provided in a specified period of time. The Capacity characteristic can be applied to different types of OSI objects, and it is quantified using various units.
  - *Accuracy* is the QoS characteristic that represents the correctness of an event, a set of events or a condition. Accuracy is a QoS characteristic of concern to the user, for whom this characteristic refers to the user information only.

    (The implications of the integrity needs for headers and similar protocol control information can be subject to separate characteristics and measure.) This characteristic is statistical, and is evaluated (measured) over the defined length of user interaction units.
  - *Protection* is a QoS characteristic that represents the security afforded to resource or to information. Protection is quantified as a probability of failure of the protection.
  - *Cost* is the QoS characteristic that represents a means of assigning value to an object. Cost is measure in terms of a currency unit. The cost of a service is often a function of the QoS options selected and must be calculable by the service user from information supplied by the service provider.
  - *Priority* is the QoS characteristic that represents the importance of an object or the urgency assigned to an event. Priority can be quantified in various ways: as a rank of a set, as a measure relative to some reference or in comparison to some other object or event.
  - *Availability* is the QoS characteristic that represents the proportion of time when satisfactory service is available.
  - *Reliability* is the QoS characteristic that represents the probability that accuracy will remain above a defined requirement (i.e. that failures will not occur).
  - *Coherence* is the QoS characteristic that represents the degree of correlation between two or more objects (events, actions or information).
- The ISO/IEC framework has QoS functions associated with system as well as with each layer in the OSI model. Correspondingly, there are QoS requirements of each model layer toward the layer below and the peer entity. Hence, the QoS aspects are dealt with in a recursive manner as presented in Section 1.3 of the main text. The layers are using the layers below, and are referred to as (N)-service-user and (N)-service-provider, respectively.
- In the ISO/IEC framework there is a *QoS management function* which contains (a set of) subfunctions classified as *monitoring,* i.e. estimating by means of QoS measurements the values of a set of QoS characteristics (parameters) actually achieved during system activity.

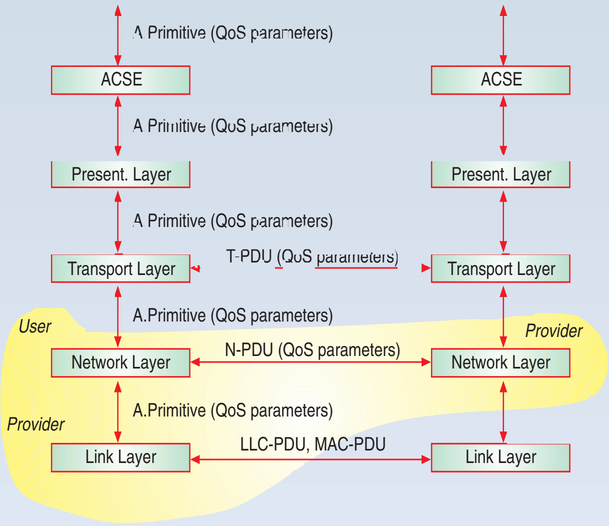


Figure D-1 Illustration of OSI QoS parameters

D.2 The Telecommunication Information Networking Architecture Consortium (TINA-C) QoS framework

The Telecommunication Information Networking Architecture Consortium (TINA-C) is an international collaboration which has as it goal to define and validate an open architecture for telecommunication services and management. It is referred to [16] for an introduction.

The TINA consortium is about to define its approach towards ensuring QoS in systems defined according to this architecture in its QoS framework [17]. See [4] for a presentation. Since this work is not commonly available and in its drafting phase, it will not be discussed in any detail. The following items should be taken into account when a framework for measurements are considered:

- The TINA framework is primarily specification and design oriented.
- It is based on the basic relation between entities as discussed in Section 1.3 of the main text.
- Several aspects of QoS are recognised, such as
  - timeliness (the semantics of the basic real time QoS dimension is defined by means of a Timing Behaviour Description Language – TBDL),
  - availability (currently detailed toward fault-tolerant design issues in distributed systems), and
  - high performance.

- - - oo0oo - - -

D.3 Unification of Frameworks

The QoS frameworks of ISO/IEC and that of ITU-T (ISDN and B-ISDN) are discussed and compared in [13], [14]. The TINA-C approach is included in this discussion in [15]. The objective of this work is to identify common features and differences between the various approaches and to reach a harmonised framework. One of the observations of this paper is: «There is no fundamental contradiction between the OSI/IEC, (B-)ISDN and ODP/TINA QoS frameworks. However, focus and also the use of concepts differ.»

For a measurement oriented framework, the following issues should be noted:

- Considering the *generic aspects of the QoS* frameworks we have:

  →*entities* (objects) performing traffic handling functions and offering QoS,

  →*QoS-parameters*,

  →*QoS-contracts*, i.e. the QoS agreed between user and provider.

- A *QoS dimension* is an aspect of a QoS framework. The following QoS dimensions are suggested in [15] (based on the TINA concept but with redefined semantics):
  - *Dependability dimension:* This dimension defines aspects concerning relationships between objects, as for instance availability.
  - *Event dimension:* This dimension defines aspects concerning the individual event that constitute the behaviour of an entity, e.g. the establishment of a connection.
  - Flow pattern dimension: This dimension focuses on the nature and structure of the information flowing in the system. In this context, the flow class concept of ATM based systems may be adopted. (It is easily seen that these classes are applicable for other transfer modes (switching principles) as well):
    - Constant bit-rate flow with strict real time constraints [CBR].
    - Variable bit-rate flow with strict real time constraints [VBR].
    - Asynchronous associate mode flow [AAM] which may be subdivided in transaction oriented sub-class (small data volume, medium real time constraints, e.g. interactive and distributed systems traffic) and a block transfer subclass (large data volume, weak real time constraints, e.g. file transfer).
    - Asynchronous message mode flow [AMM] (One way transfer with no immediate response, no real time constraint, e.g. e-mail).
- Layering is an important issue. The QoS requirements of an application are propagated downwards in the protocol stack defining both the flow and quality parameters associated with each level.

- Allocate a “fair share” of the quality reduction to the end-users network/ equipment.

2.2 Layer

In the introduction of the ISO/IEC framework [5], see special item D.1, it was outlined how quality of service can be associated with each level in a layered architecture. The quality of service the users perceive, is the one from the application level. Hence, it is the application level QoS that ideally should be measured. However, the public network operators are only responsible for the layers handled by their network. This is illustrated in Figure 3 where the responsibilities of the public network operator are indicated.

- In the user plane, low protocol level traffic is handled in the transfer phase of a connection. The actual levels will depend on the actual (bearer) service.

  Measurements should be related to the data transfer between end-users, i.e. end-to-end measurements.

  A single end-user may have multiple simultaneous connections, and ideally all of them should be measured.

- In the control plane the handling of connections and service features takes place.

  The network response to customer actions should be measured.

E ITU-T terms and definitions

The telecommunication standardization sector of ITU presents in its recommendation E.800 a comprehensive set of terms and definitions relating to quality of service, network performance and dependability standards pertaining to the planning, provisioning and operation of telecommunication networks [1]. Associated terminology covering statistical terms, recommended modifiers etc. is also included. Parts of this recommendation is adopted by IEC as terminology standard IEV 191. It is ITU-T's intention that these terms and definitions could be universally applied to all telecommunication services and networks.

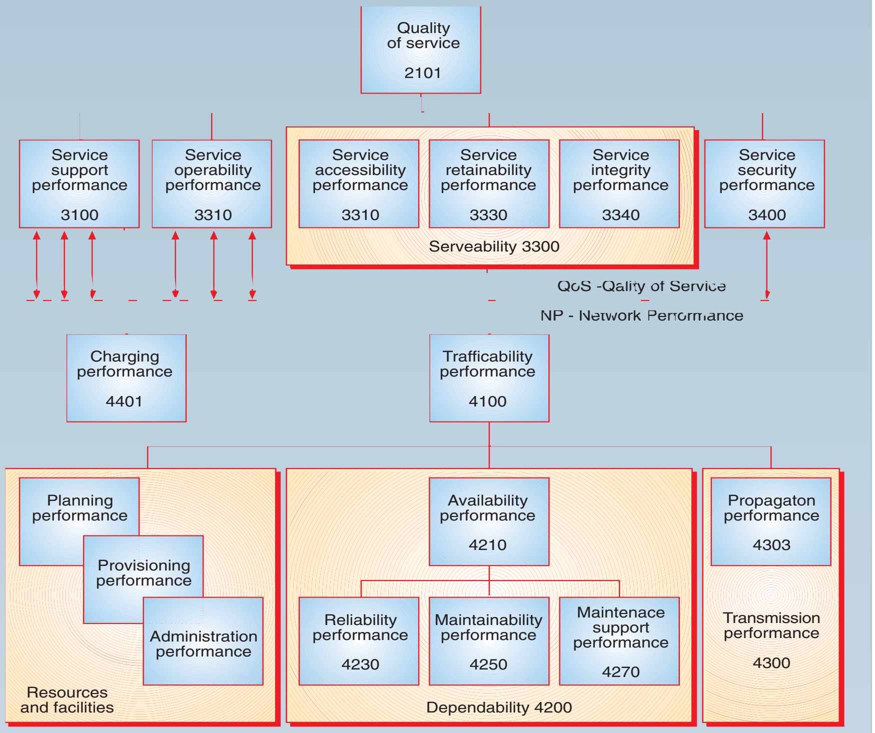

Quality of service 2101
Service support performance 3100
Service operability performance 3310
Service accessibility performance 3310
Service retainability performance 3330
Service integrity performance 3340
Serveability 3300
Service security performance 3400
QoS -Qality of Service
NP - Network Performance
Charging performance 4401
Trafficability performance 4100
Planning performance
Provisioning performance
Administration performance
Resources and facilities
Availability performance 4210
Reliability performance 4230
Maintainability performance 4250
Maintenace support performance 4270
Dependability 4200
Propagaton performance 4303
Transmission performance 4300

Figure E-1 Relation between QoS and network performance concepts as defined in ITU-T recommendation E.800

The relation between the various performance concepts of E.800 is shown in Figure E-1. It should be noted that E.800 makes a clear distinction between an ability and measures of this ability. For instance, the reliability performance may be measured by the reliability, the failure rate, the mean time to first failure, etc. The performance concepts are divided into two classes.

> **Quality of service (QoS),** which is the collective effect of service performance which determine the degree of satisfaction of a *user* of the *service.* QoS measures are only quantifiable at a service access point, and solely related to the users perception of the service.
>
> **Network performance (NP),** which is the ability of a network or network portion to provide the functions related to *communications* between *users.* From the provider's viewpoint, network performance is a concept by which network characteristics can be defined, measured and controlled to achieve a satisfactory level of service quality.

It is seen that by this subdivision between QoS and NP the ITU-T QoS concept lacks the recursivety discussed in Section 1.3 and inherent in other conceptual frameworks, see item D.

The following service performances contribute to the QoS:

- *Support:* The ability of an organization to provide a service and assist in its utilization.
- *Operability:* The ability of a service to be successfully and easily operated by a user.
- *Serveability:* The ability of a service to be obtained, within specified tolerances and other given conditions, when requested by the user and continue to be provided without excessive impairment for a requested duration.
- *Security:* The protection provided against unauthorized monitoring, fraudulent use, malicious impairment, misuse, human mistake and natural disaster. Details concerning security is for further study.

All the above service performances are dependent on the network characteristics. For instance the service support performance depends on certain aspects of the network performance like charging correctness. However, the serveability performance, which is the focus of this paper, is the most generally affected. It is subdivided into three service performances:

- *Accessibility:* The ability of a service to be obtained, within specified tolerances and other given conditions, when requested by the user.
- *Retainability:* The probability that a service, once obtained, will continue to be provided under given conditions for a given time duration.
- *Integrity:* The degree to which a service is provided without excessive impairments, once obtained.

Serveability performance depends on *trafficability performance* and its influencing factors of resources and facilities, dependability and *transmission performance*, as shown in Figure E-1. Note the similarity with Figure 5 of the main text. The trafficability performance is described in terms of losses and delay times. *Dependability* is the combined aspects of *availability, reliability, maintainability* and *maintenance support performance* and relates to the ability of an item to be in a state to perform a required function. The *resources and facilities* box, included the items within it, are for further study.

Hence, the activities across these interfaces are natural choices for measurement of the QoS provided by the public network operator.

The principle of measurements related to the interface between network and user and with activities in specific planes and layers has the advantage that:

- The measurements may serve in a business context (as pointed out in the beginning of the section)
- The measurements will be independent of the actual teleservice/application, as long as the protocols used are the same.

The disadvantage is obvious:

- It may be difficult to relate the observed values to "end-user-satisfaction with the service".

Hence, ideally we should with the proper knowledge of the end-user application be able to derive the performance of the end-user application(s) from the measurements performed.

Example:
A human end-user working interactively with a computer across the network is interested in the response time he gets. The network contribution to this response time is the end-to-end delay (both ways) of the last packet (or ATM cell) of the "chunk" of information transmitted and the end-to-end retransmission delays induced by packet/cell loss or corruption. Hence, for this application both these "low level" characteristics must be used to determine the effect on the application.

In this example, it is also worth to note that the processing, packetizing and depacketizing in both ends of the connection will contribute to the delay and may dominate overall response time.

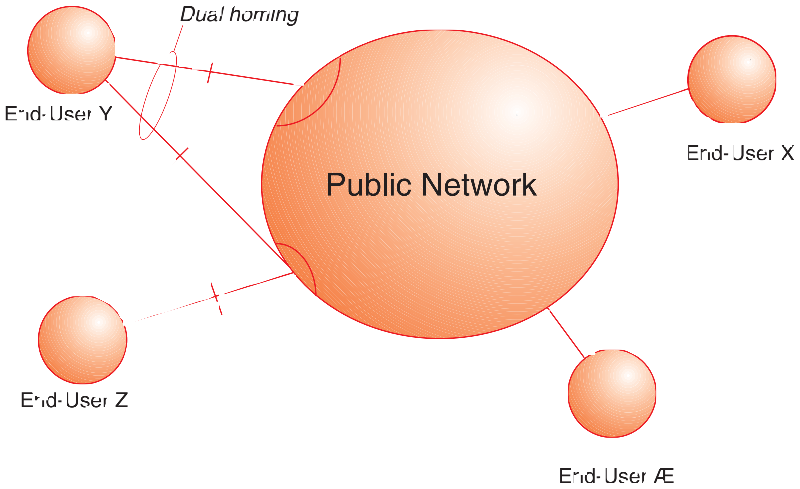


Figure 4 Sketch of end-user network interconnections

2.3 Number of connections

An end-user (or more precisely an end-user site) may have more than one physical interconnection point to the network, as illustrated for user Y in Figure 4. For these end-users, it is necessary to take all interconnections into account when QoS is measured. For instance, end-user Y will experience a better service accessibility performance than user Z, due to his dual homing. He may, however, experience a larger blocking during periods where one of his connections or access points are unavailable.

It is suggested that all interconnections from a single end-user toward the public network are regarded as a common reference point with respect to QoS measurements, when the same bearer services are provided over these. Measurements may be carried out for individual connections/interfaces, but these must be carried out in a way so that an overall QoS measure may be derived.

Note also that the end-user may have different teleservices provided over the same interface, for instance an ISDN interface. Similarly, an end-user may have partly substitutable services, like e-mail and fax provided (by different bearer services) over different interfaces. Simultaneous failures of these services will have a different effect on the end-user than independent failures, and thereby his perception of the quality of the service delivered. This effect will be larger the larger part of the end-user's service repertoire that is affected.

3 End-user experience of QoS

Quality of service, at least as defined by ITU-T [1], is basically subjective. How the end-user perceives the service quality, i.e. his satisfaction with the service as a function of the measured parameters, is highly dependent on a number of factors. Among them are:

- The different needs of individuals and/or organizations
- The service level (QoS – parameters) he is accustomed to
- The actual service and its usage
- Individuals and/or organisations using the service
- The cost of the service.

An investigation has been carried out on initiatives to measure QoS in UK, Australia, USA, Netherlands, France and Germany [6]. These initiatives have been facilitated by the country regulators under influence from country user bodies. This investigation, [6], confirms that users, and to some extent regulators, have a common view of QoS which goes beyond those QoS aspects related to network performance. Users' view of network performance also includes installation time, repair time, billing effectiveness, complaint handling and the many characteristics which influence the users' experience and perception of network performance, i.e. serviceability.

To go into any detail on the issue of the end-user's satisfaction of QoS aspects related to network performance/serviceability is outside the scope of this paper. However, it may be concluded, without any in-depth investigations, that:

A *The relation between degree of satisfaction and the measured parameter values are highly non-linear.*
For instance, a round trip delay of 100 ms is not noticed by an interactive computer user, while 1 second is pretty annoying and 10 sec is unacceptable. Similarly, network blocking of phone calls well below the B-subscriber busy probability (e.g. 0.5 %) is not considered a nuisance, while a blocking at the same level (≈ 10 %) is very annoying and twice this level is unacceptable.

B *The end-user perceived QoS is not governed by long term averages*[1].
As an extreme example, a lack of service accessibility measured as 1 % is perceived differently whether it is caused by a one percent blocking of calls during a year, or is constituted by complete outage lasting for four consecutive working days in a year.

C *In the current public network(s) the end-user is satisfied with the technical QoS*[2] *the dominant part of the time.*
By satisfied is in this context meant that the end-user accepts the service he is offered for the price paid. This statement is supported by the fact that the end-user rarely complains about the technical QoS. [Note that this statement does not imply that the end-user would not like to have an improved QoS if it were available at an acceptable (no) additional cost – or similarly would accept a reduced QoS if the cost of the service was substantially reduced.]

Hence, to capture the user's experience of the QoS in a simple way, the concept of service failures is introduced, where the performance (instantaneous QoS parameter values) deteriorates so much that the end-user notices it and it changes his subjective perception of the service. This is discussed in some more detail in the remaining subsection.

3.1 Experienced/monitored QoS as a stochastic process

The QoS parameters monitored or measured constitute a stochastic process, determined by:

- The network itself, which is deterministic.
- The failure and repairs of network functions and/or elements which are stochastic processes with activities on medium to long time-scales.
- The offered traffic, which is a stochastic process with activities on short to long time-scales. The offered traffic is to some degree influenced by the observed end-user.
- The environment, which is a stochastic process with activities on short to long time-scales.

A rough sketch of the relationships between these processes is shown in Figure 5.

The end-user's experience of the QoS will also depend on the end-user's utilization of the service. Cf. for instance the common argument that failures during the night are of less importance since it is rather unlikely that the individual end user will experience it. The end-user experiences the technical QoS by a sampling process defined by his utilization of the service. Note that this sampling

- is a stochastic process in itself, governed by the end-user's communication needs
- may influence the QoS observed significantly, if the actual usage of the service is a heavy use of network resources. (For instance a file transfer in a packet switched network may significantly increase the end-to-end delay across the network compared to what it was before the transfer started.)
- will depend on QoS received from the network. (For instance, if a telephone call is blocked due to network congestion, repeated call attempts will be made which are also likely to be blocked, and the end-user samples the network congestion more frequently than his ordinary call rate/process would have done.)

Hence, the resulting observed values of the QoS parameters form a very complex and composite stochastic process. It is stated above that the end-user perceived QoS is *not* governed by long term average. Neither is the end-user perception of QoS influenced by time constants in this stochastic QoS process that is substantially shorter than the time constants of human perception, as long as the QoS reductions do not cause consequences of a larger extent than the QoS reduction itself.

Figure 6 shows a sketch of how a QoS parameter is observed by an end-user, i.e. the QoS process, as a function of time. The QoS parameter observed may for instance be call set up delay or end-to-

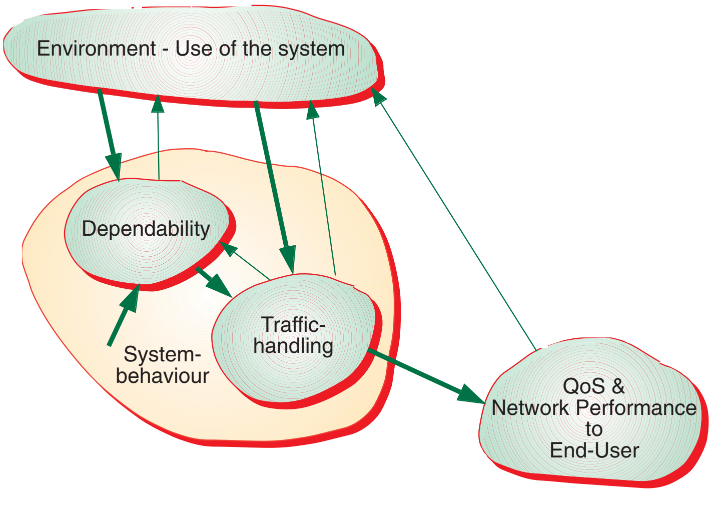


Figure 5 Sketch of elements and dependencies influencing the QoS

[1] *Formally, this statement is a consequence of the item pointed out above.*

[2] *The part of the complete service provided by the traffic machine.*

end delay through the network. The end-user samples the QoS parameter (which is a time variable stochastic quantity) by his utilization of the service. The utilization of the service is referred to as *interaction with the network* in the figure.

Figure 6 is drawn with definite interactions with the network in mind, i.e. a single transaction in the event dimension like the establishment of a connection. The picture will be a little different for the information transfer phase of a connection (i.e. in the flow pattern dimension) where it is more common to regard the observed parameter as a continuous variable. However, basically the same kind of process takes place, e.g. a bit is correctly transferred in a synchronous stream, or the delay of an ATM cell through an asynchronous network. The next subsection deals briefly with the measurement of this kind of processes.

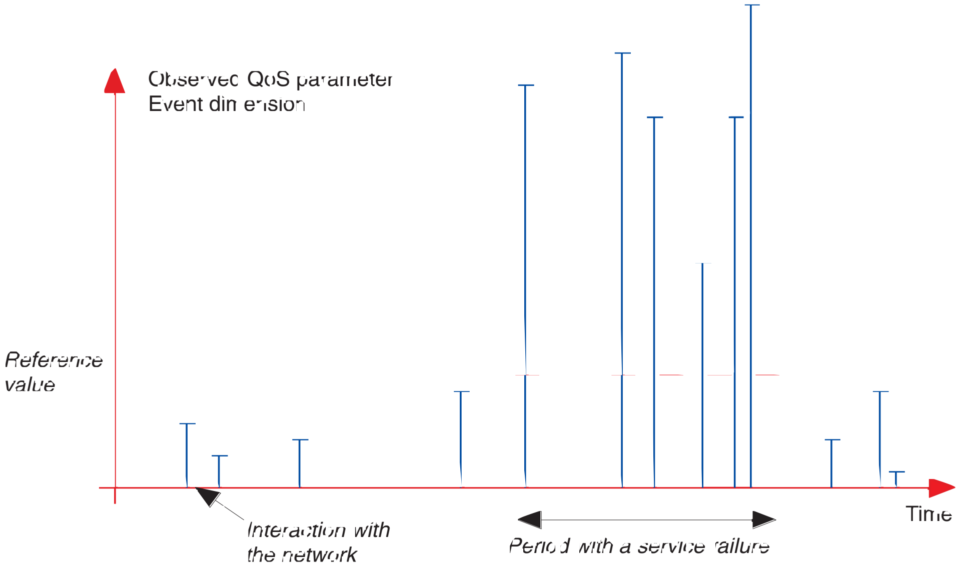


Figure 6 General sketch of how a QoS parameter is observed by the end-user

3.1.1 Observation process

It should be decided whether the observation of the QoS parameters should be performed

- *uniformly.* In this case the statistics will be obtained irrespective of typical or specific usage. An example of this observation method is to make regular test calls during the day (and night) during all days of the week.
- with a sampling *corresponding to the typical or specific usage* of the service considered. Examples of this kind of observation method are, a) to make test calls with a frequency corresponding to the daily and weekly call frequency profiles, or b) to observe the value of the QoS obtained for every or a fraction of the interactions with the network, e.g. the set-up delay of a connection.

3.1.2 Averaging time intervals

This issue has relevance for the transfer of information between end users, i.e. the flow pattern dimension. With respect to the statistics of some QoS parameters, short term averages must be obtained, either to make the measurements meaningful, e.g. bit failure rate, or to make measurements feasible due to a very large number of observations, e.g. end-to-end ATM cell delays, or both. How long should these time intervals be to capture the effect of the variation of the QoS parameter on the end-user? Too short intervals would create an excessive volume of data and too long intervals would hide the effect. For instance, for speech quality in analogue telephony, the duration of a connection is a suitable interval to form averages of observed values, or should the interval be substantially shorter, in the order of the duration of one syllable?

3.1.3 Regular statistics

Measurements on a real network will not be carried out under stationary conditions, cf. the discussion at the beginning of Section 3.1. Hence, the dynamics of the QoS may have a large influence on how the QoS is perceived. It is suggested that at least the following statistics is obtained for the QoS parameter (process):

- Mean, eventually means over short periods as discussed above
- Variance
- The instantaneous distribution of the parameter (occurrence frequency)
- Autocorrelation. The autocorrelation indicates how rapid the QoS level changes and may give valuable information on how the QoS is perceived.

3.1.4 Extreme process statistics

With respect to QoS experienced by the end user, special emphasis should be put on measurements of the extreme behaviour of the QoS process. More specifically, values of the QoS parameters in a range that represents service failures should be brought into focus, see Figure 6, next section and Section 5.2. Suggested minimum of statistics obtained for each QoS parameter relevant for the service:

- Rate of service failures. (It may be discussed if it is worthwhile to obtain the inter-service failure statistics in more detail.)
- Duration of service failures
- Mean and variance of the QoS parameter *during* service failures
- Autocorrelation in the service failure process. (An exact definition is yet to be developed, also depending on the observation process. It is, however, suggested that the autocorrelation is based on indicator functions of whether the service is satisfactory in an attempt or not, relative to the reference value defined for the actual QoS parameter, cf. Section 5.2.)

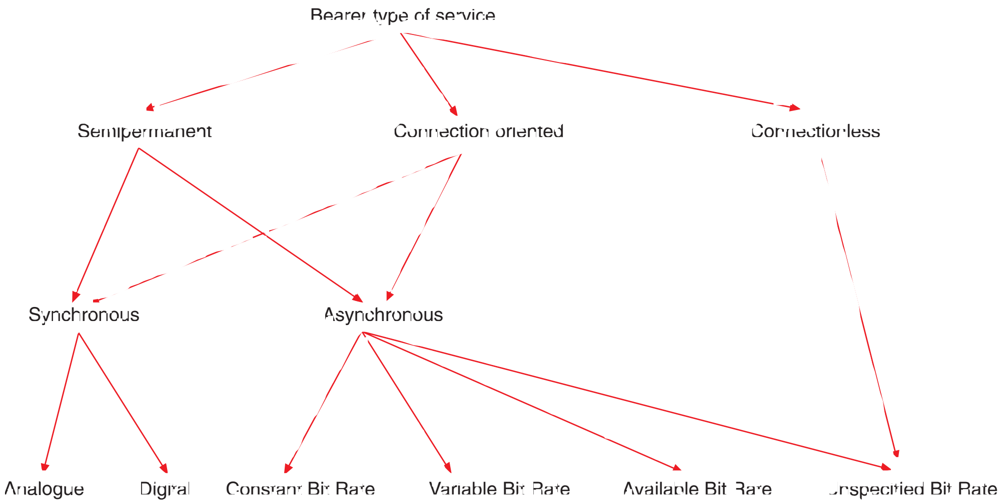


Figure 7 The structure of bearer services

3.2 Service failures

The concept of service failures, or more precisely, end-user service failures is adopted from [7], see information item C. The end-user considered is defined by the following two criteria:

a The entity paying the public network operator for the services received. [The person(s) using the services may be the end-user himself, his employees, his customers, etc., and they may use the services directly (e.g. telephone) or indirectly (e.g. automatic teller machine).] This is similar to the definition of a customer in [2].

b The part of the entity's activity located at a single site. For instance in Figure 4, X, Y, Z and Æ are all considered separate end-users even if the same entity (e.g. a bank) is paying for, say, both X and Y[3].

The end-user service failure is related to QoS, the latter being defined in ITU-T recommendation E.800, [1], as: *The collective effect of service performances which determine the degree of satisfaction of a user of the service.* However, here only the aspects directly related to network performance are relevant, namely the serviceability performance See item E for a brief introduction.

Definition of end-user service failure:

> *Temporary inability of a service to be provided to the considered end-user, characterised by a change, beyond given limits, of at least one parameter essential for the service.*

With respect to the definition of service parameters and the limits (or reference values) for the parameters, it is referred to Section 5.2. The limits may be agreed between network provider and end-user on a (semi)permanent basis, e.g. as a service level agreement (SLA), or negotiated between network and user on a per connection basis, e.g. as discussed in [4].

It should also be kept in mind that the limits/ reference values are what is agreed for the service/connection. For instance for "best effort" connections[4] there are no given limits and no service failures will occur.

Informally, an end-user service failure is present when one of the service's QoS parameters is poorer than its nominal/ specified/negotiated value. Note that the values of these parameters also describe the failure, as long as it is present, together with the failure duration.

The above definition is closely related to the ITU-T definition of interruption; break (of service)[5]. The differences are that

- No lower limit is set on the time duration of a failure/interruption
- The end-user is not necessarily a single person, but a composite user as defined above.

The end-user service failure is defined irrespective of the cause. Hence, causes like network overload and human misoperation may also result in an end-user service failure.

The most common indication of QoS to the end user – or rather lack thereof – is customer complaints per number of lines and time period. It is hypothesized that this complaint frequency is closely related to the frequency of service failures, and eventually other statistics of the end-user service failure process (see Section 3.1.4.) Hence, service failures may form a basis for using the customer feed-back (through complaints) as a means for improving the "correct" network performance issues.

[3] *The single site criteria in the definition of an end-user are introduced for conceptual simplicity, i.e. to make measurements of the QoS provided to a specific end-user easier to understand. For bearer services, this does not seem to introduce any restrictions. If the framework is extended to other services this requirement should be reconsidered. For instance in an 800 service, it is the receiving party that is the end-user according to criterion* **a.** *The calling party may be directed to one out of several sites according to various criteria. Hence, in this case the single site criteria becomes restrictive.*

[4] *Examples are unspecified bitrate (UBR) ATM connections and connections with the compulsory OSI QoS requirements.*

[5] ***Interruption; break (of service)*** *(definition 4101): Temporary inability of a* service *to be provided persisting for more than a given* time duration, *characterised by a change beyond given limits in at least one parameter essential for the* service. *(*Note 1 – *An* interruption *of a* service *may be caused by* disabled states *of the* items *used for the* service *or by external reasons such as high service demand.) (*Note 2 – *An* interruption *of a* service *is generally an* interruption *of the transmission, which may be characterised by an abnormal value of power level, noise level, signal distortion,* error *rate, etc.)*

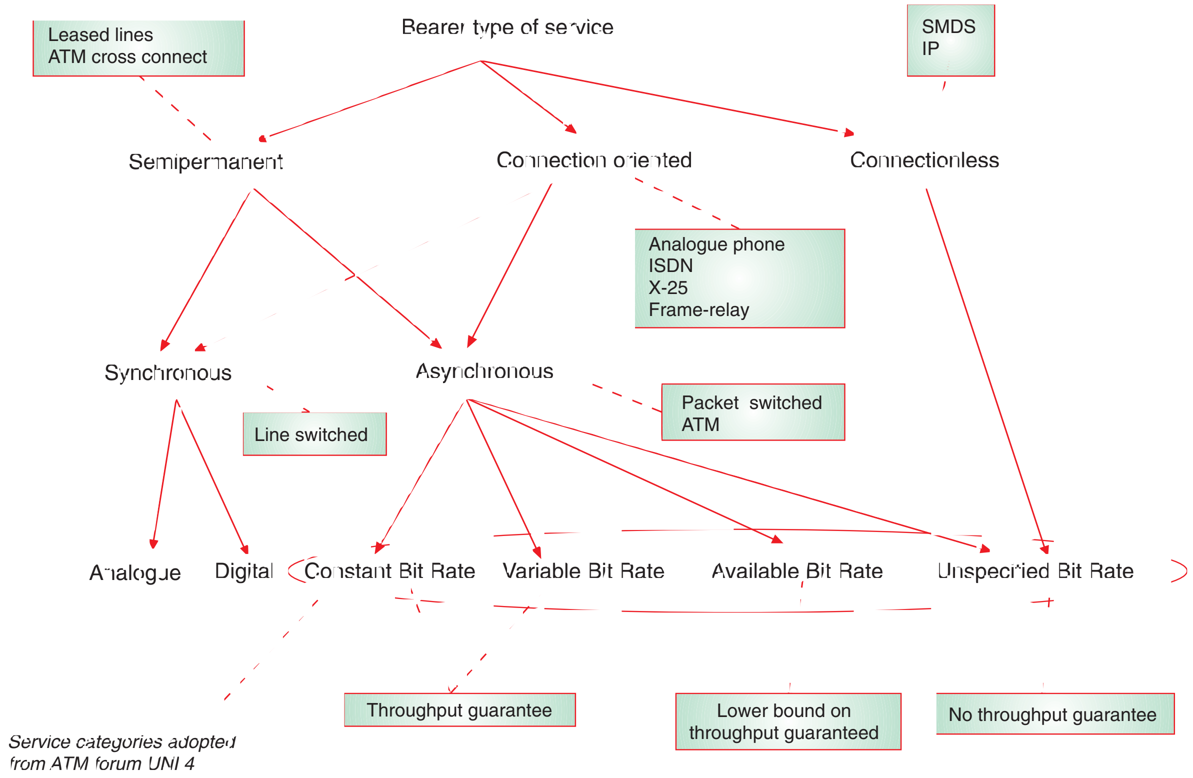


Figure 8 Annotation of the bearer service structure

4 The structure of bearer services

There are a number of functions associated with the provision of telecommunication services. For instance, in [2] the following service functions have been defined:

- Sales
- Service management
  - provision
  - alteration
  - service support
  - repair
  - cessation
- Call technical quality
  - connection/establishment
  - information transfer
  - connection release
- Billing
- Network service management by customer.

These service functions include both the organizational and technical aspects. With respect to the E.800 QoS definition, the quality in provisioning all of these contributes to the QoS provided to the end-user. See [2] for a more detailed discussion.

As pointed out in the introduction, this paper concentrates on the quality aspects related to the technical performance and characteristics of services. They are those referred to as "Call technical quality" in the above list and the ability of the end-user to access the service once it is provided. The latter is determined by the technical performance of the network as well as the maintenance support performance of the operator(s). Hence, the repair function is also considered.

Before starting a discussion of the quality aspects related to the technical performance and characteristics of services, we consider the basic structure of services. The objective is to identify common characteristics among services which will guide the choice of types of measurements. These should be independent of specific services (or put in other words, common to many services). Such a structure is shown in Figure 7. The first level shows the connection type. The next shows the transfer mode, and the bottom level shows the stream characteristics.

To exemplify this structure, it is annotated in Figure 8. Note that the annotation is not intended to be exhaustive.

The asynchronous transfer types are adopted from ATM Forum's UNI 4.0 implementation guidelines, since it is assumed that these will be generally adopted. As will be seen below, the generic QoS service parameters suggested are the same for constant and variable bitrate transfer, and for available and unspecified bitrate transfer. Hence, it may be discussed whether the granularity used here, i.e. four types, is necessary or whether it is sufficient to consider only two types,

- Real time transfer (constant and variable bitrate), and
- Best effort transfer (available and unspecified bitrate).

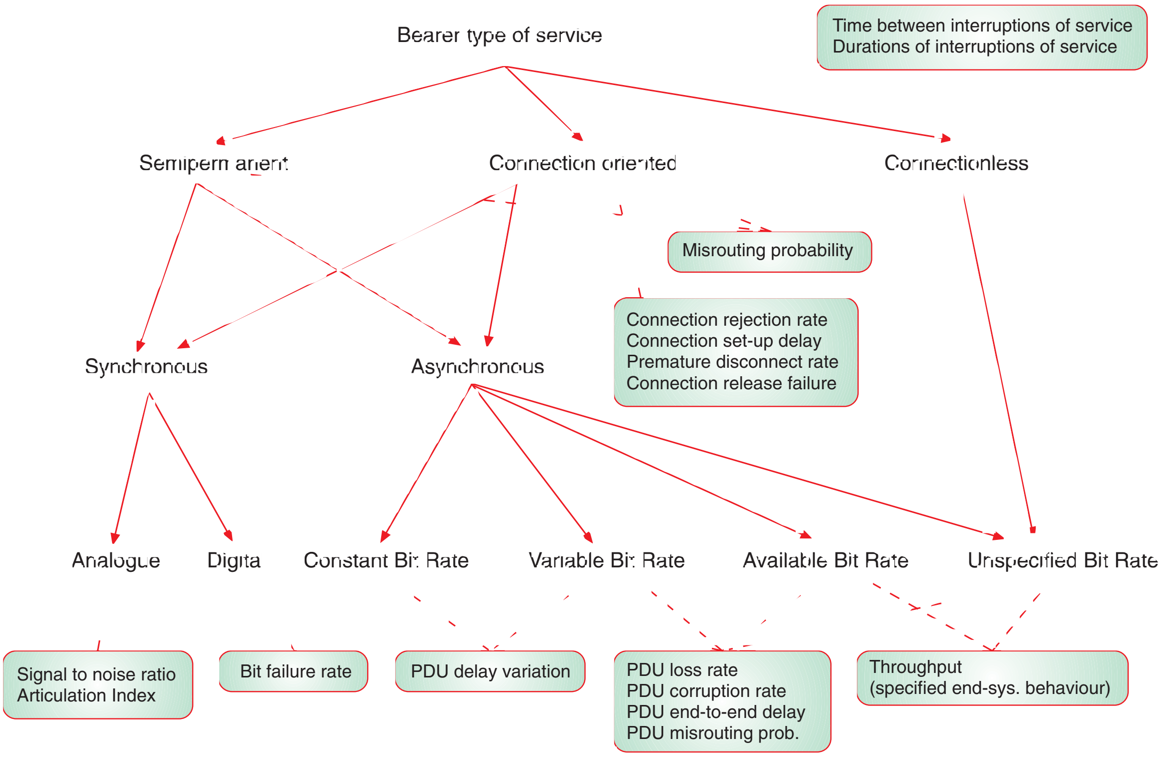


Figure 9 Generic QoS parameters related to the service structure

5 QoS parameters

This chapter introduces QoS parameters related to the service structure above. It is intended that the QoS indicators of ETSI Technical Report ETR 138 for voice telephony and Integrated Digital Network (ISDN) [8], with some modifications, should fit into the framework suggested here. The proposed framework is intended to be a generalisation of the one in ETR 138 to all services. Furthermore, it is the intention to simplify ETR 138 by introducing QoS parameters common to a number of services, e.g. connection set-up delay. The measurements principle and the way a sufficient QoS is specified, is also intended to be common across a range of services. However, how the detailed measurements are carried out, e.g. which signals are registered on the interface between end-user and network as discussed in Chapter 2, is service specific. To exemplify: in ETR 138 call set-up time is defined for a) voice telephony, b) circuit mode switched bearer services, and c) packed mode switched bearer services. The measurement of these are based on different signals between end-user and network, but the QoS parameters obtained are essentially the same.

5.1 Generic QoS parameters associated with service structure

The QoS parameters *suggested* related to the service structure of Figure 7 are introduced in Figure 9. It is stressed that the parameters are suggestions and that additional parameters may be introduced and some of the suggested ones may be regarded as superfluous. Furthermore, the parameters have not yet been precisely defined.

The QoS parameter is associated with aspects of the services, irrespective of the specific service. For instance, a basic function in all asynchronous services is to transport packetized data units (PDUs) across the network. A QoS parameter associated with this function is the rate (or probability) of lost PDUs. This parameter is the same, irrespective of whether the PDU is an SMDS packet, an ATM cell in a semi-permanent cross connect (“leased virtual path”) or an X.25 packet.

The number of service parameters are sought kept as small as possible.

Some comments related to the suggested QoS parameters:

- Service interruptions, as defined in [1], are used as an indicator of the serviceability performance of the system. The time between interruptions and the duration of interruptions are used as parameters. A more precise definition of what constitutes an interruption is given in Section 5.2.
  - The time between interruptions and the duration of interruptions are the QoS parameters which describe service failures as outlined in Section 3.2.

- These parameters cover the same properties as the “failures per thousand access lines” and the “severely errored minutes” defined in [8]. One of the reasons for using service interruptions instead of these is that to the end-user it is of minor interest why the service is rendered useless. Another reason is that there are more causes for service interruptions than excessive noise (bit-failure rate) and loss of access.
- The discussion of Section 3 should be kept in mind. Hence, the actual duration of the time between the interruptions and the duration of the interruptions should be measured. In fact, to be able to properly derive relations between the service failures (service interruption process) and the end-user perception of the service, higher order statistics like the correlation and auto-correlation should also be collected.

- The availability of a semi-permanent connection (leased line) is not considered explicitly, but is captured through the service interruption QoS service parameters. (A specific leased lined is in this context regarded as a service which is ordered from a public network operator.)
- The only “connection characteristic” considered for semi-permanent connections is the *misrouting probability*. This characteristic is obviously relevant for connection oriented services (wrong B-party connected). For semi-permanent connections it is interpreted as a) the connection is established to a wrong destination, b) an established connection is redirected to a wrong destination.
- It is seen that for asynchronous (packet oriented) services, the probability of a PDU reaching a wrong destination is also a property associated with the stream characteristics, e.g. a single PDU of a virtual connection reaches a wrong destination. This is denoted *PDU misrouting probability*. This QoS parameter also accounts for connectionless transfer when a PDU is directed to a wrong destination.
- For the stream characteristics (delays, jitter, PDU losses, bit failure rates, etc.), the properties of the two uni-directional end-to-end connections are regarded separately.
- It is left for further study how the various QoS parameters should be measured. For instance, the *PDU loss rate* may be defined and measured as any of the alternatives below:
  - the relative number of PDUs lost during a period of …,
  - the absolute number of PDUs lost during a period of …, (becomes

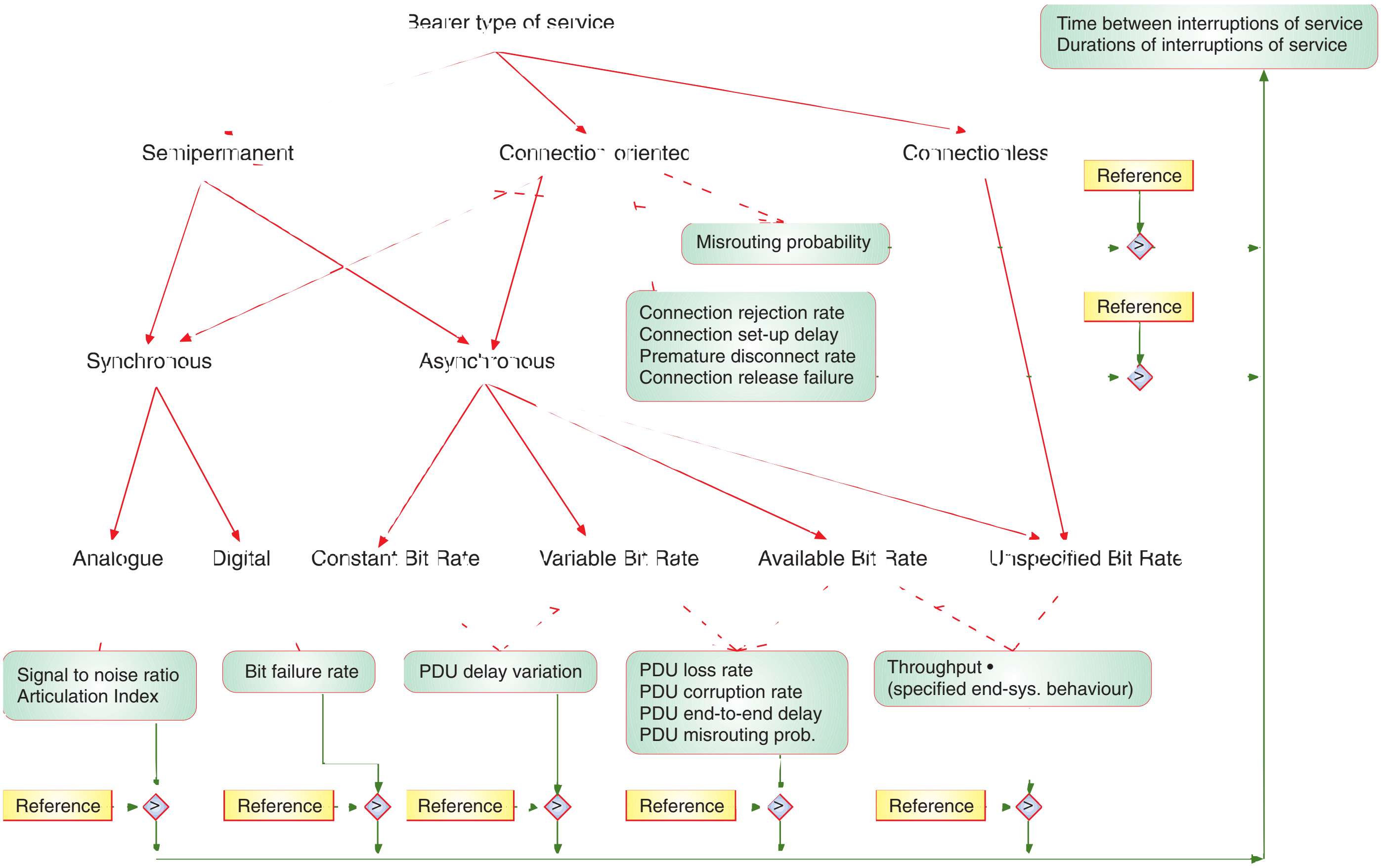


Figure 10 Illustration of the service failure definition

Table 1 Comparison between the ETSI QoS indicators and the service independent QoS parameters

QoS indicators	Service independent QoS parameter; comment
Voice telephony	
• Fault reports per access line and year	Covered by time between interruptions of service
• Unsuccessful call ratio	Included
• Call set up time	Included
• Speech transmission quality	Included, but articulation index is suggested as a non subjective metric.
• Supply time for initial network connection	Not relevant
• Response time for operator services	Not relevant
• Availability of card/coin phone	Not included
• Fault repair time	Covered by duration of service interruptions
ISDN all bearer services	
• Fault reports per access line and year	Covered by time between interruptions of service
• Severely errored seconds	Covered by time between interruptions of service and the duration of interruptions triggered by a large bit failure rate
ISDN circuit switched	
• Unsuccessful call ratio	Included
• Call set up time	Included
ISDN circuit mode permanent	
ISDN all packed mode bearer services	
• Throughput efficiency	Included
• Round trip delay	Covered by two end-to-end delays
ISDN packed mode switched	
• Unsuccessful call ratio	Included
• Call set up time	Included
ISDN packed mode permanent	
• Availability of …	Covered by time between interruptions of service and the duration of interruptions
• Number of service interruptions per year	Covered by time between interruptions of service
NOT INDICATED	
•	Premature disconnect rate
•	Misrouting probability
•	Connection release failure
•	PDU delay variation
•	PDU loss rate
•	PDU misrouting probability

meaningful only if the traffic is known and the time constants of the traffic process are less than the measurement period)

- the (the distribution of the) distance between lost PDUs in number of PDUs and/or in absolute time (seconds),
- the same statistics as above plus the autocorrelation of the loss process (which may be necessary to determine the effect on the higher level traffic handling as discussed in Section 2.2),
- etc.

5.2 Definition of service failures

A reference value is associated with each QoS parameter relevant for the service considered. As long as all relevant QoS parameters are below their reference values (which may be stated in a service level agreement between end-user and public network operator), the service is considered satisfactory. When one of them is exceeded, an interruption of service occurs and we have a service failure. This is illustrated in Figure 10. In the figure, a single reference is shown for several QoS parameters. This is solely done to not overload the figure and make it unreadable. It is intended that there should be a reference value for each QoS parameter defined.

The box "reference" indicates well defined criteria for when the service is good enough with respect to the corresponding parameter. For instance, the bit failure rate should be lower than a certain reference value. It still remains open how the reference criteria should be defined. There may be simple threshold values, but it is foreseen that more composite criteria will be used. In any case, it must be clearly stated how the actual QoS parameter should be measured and compared to the reference values.

An example of a simple threshold value is that the bit failure rate shall not exceed 10^{-6} measured over a predefined period of a second (jumping time frame). An example of a composite threshold value is that the bit failure rate shall not exceed 10^{-6} measured over a predefined period of a second (jumping time frame) and not exceed 10^{-4} measured over a predefined period of 50 ms.

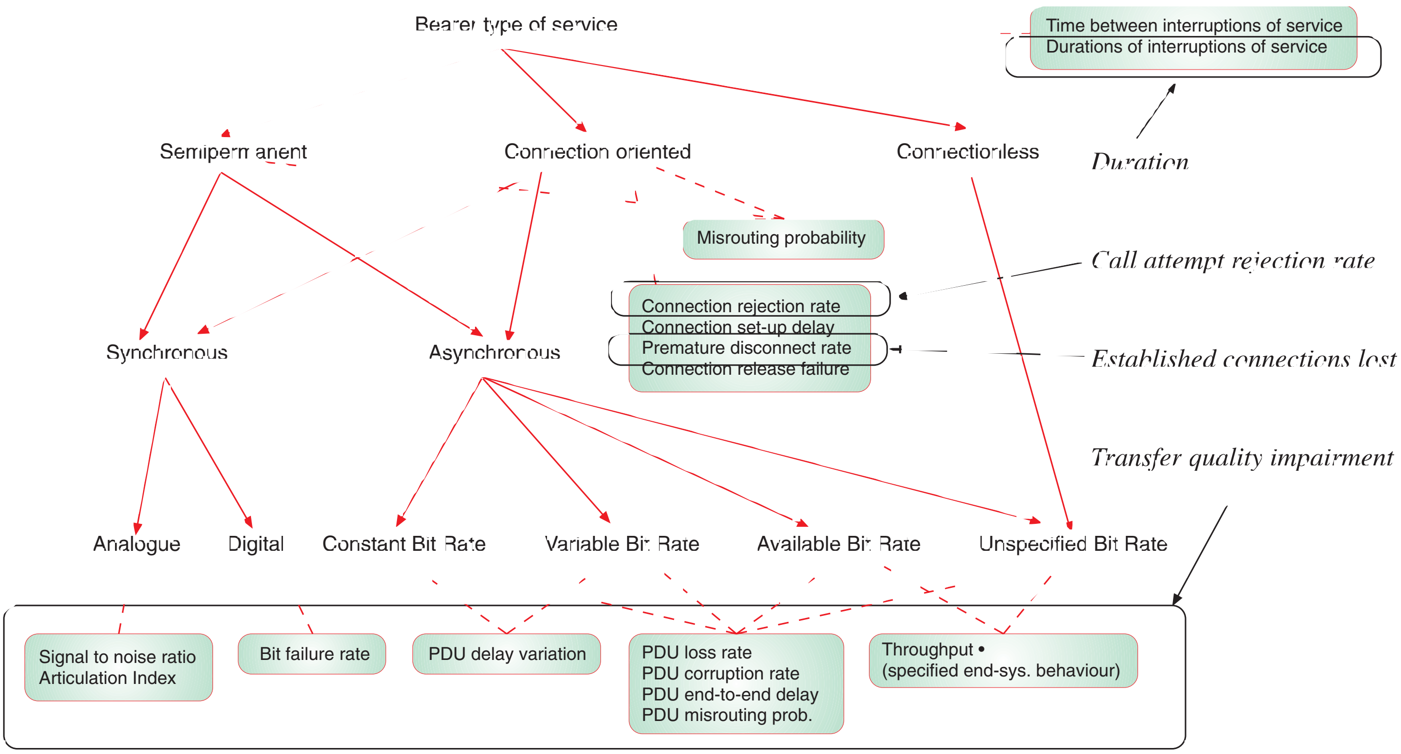


Figure 11 Relation between the proposed QoS parameters and the service independent failure attributes introduced in [7]

The actual reference criteria must depend on what is agreed between network operator and end-user/customer for the service at hand. The measurement method chosen for the corresponding QoS parameter must at least be capable of determining whether the reference criteria is met or not.

How the actual measurements are carried out in detail will depend on the service considered. For instance, which signals in the interface between network and end user constitute the start of a set up sequence and which signals (or lack of signals within a certain time) define a call attempt rejected by the network. If standard metrics for the QoS parameters exist, and corresponding reference criteria, these should of course be used.

In [8] the availability of semi-permanent connections[7] is considered. None are explicitly defined in Figure 9. However, this QoS parameter is covered by the time between service interruptions and the duration of service interruptions. The average of these determines the availability of the service.

5.3 Revisit of previously presented frameworks

Before proceeding to the concluding remarks, it is interesting to see how the proposed structure ties in with some of the previously suggested frameworks summarised in the separate boxes B, C and D.

5.3.1 QoS indicators for ONP

These indicators are defined in ETR 138, [8]. See the separate information box B for a summary. The QoS indicators for ONP are to some extent already discussed above. To get a more comprehensive overview, Table 1 shows the QoS indicators with a comment on their service independent QoS parameter counterpart. It should be pointed out that the latter is also intended to cover services not included in [8].

5.3.2 Service failure attributes

Four of the service independent failure attributes introduced in [7], cf. the previous discussion and information box C, are recognised in the current framework. These are indicated in Figure 11. The fifth attribute, impact, is not considered in this framework. See however the discussion at the end of Section 2.3. Comparing the two approaches, it should be kept in mind that [7] considers the attributes of a failure in the service provisioning to the end-user, given that a failure has occurred. Hence, the time between service interruptions is for instance not considered in [7].

5.3.3 QoS dimensions

In the unification of frameworks presented in information box D.3, the concept of QoS dimensions is introduced. These are recognised in the proposed framework as shown in Figure 12.

Compared to the layered and distributed architecture approach, it is seen that analogue and digital synchronous transfer is introduced in the flow pattern

[7] *E.g. Availability for the bearer service for packet mode permanent services.*

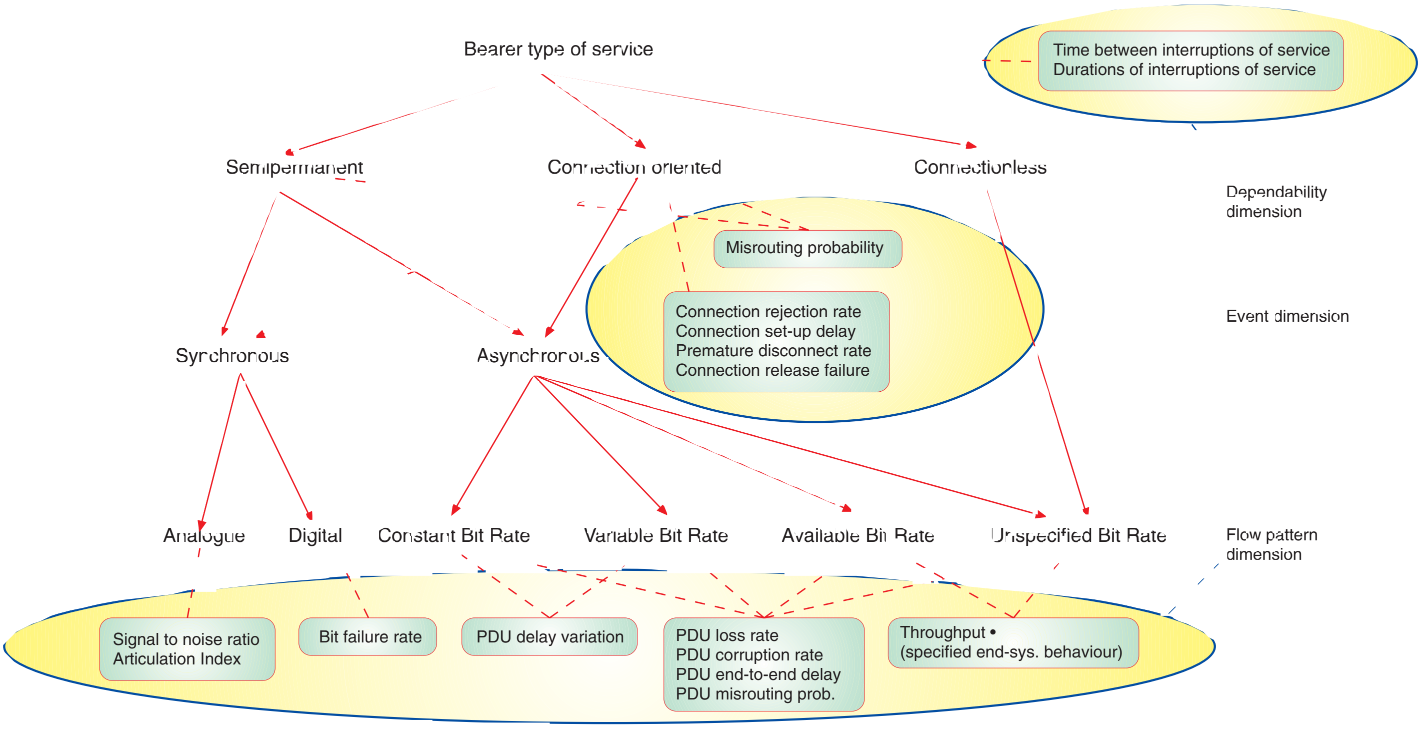


Figure 12 The proposed QoS parameters classified according to QoS dimension

dimension. Furthermore, two of the asynchronous transfer modes AAM and AMM are replaced by the newly introduced transfer services ABR and UBR (ATM Forum UNI 4.0). However, the semantics between these are roughly the same.

6 Concluding remarks

The paper summarises some actual QoS frameworks. These are partially overlapping, but essentially supplementary since they focus on different aspects. Even though concepts may be different, and there are considerable differences in terminology, a unification seems feasible. None of these are however primarily intended as a basis for measurements, nor are they intended for a quantitative comparison of QoS aspects across different types of bearer networks.

Aspects with respect to actual measurements of QoS parameters are briefly discussed with respect to

- Which signals should be observed at the user network interface, and
- Which statistical quantities should be obtained.

The notion of service structure is introduced. QoS parameters may be associated with the elements of this service structure. Hence, specific measurement methods may be developed for service elements and reused for a range of services. The actual signals which are observed and the QoS reference values will be specific for the actual services. This approach is an enhancement of the QoS indicators of ETR 138 [8].

An essential problem is the mapping between end-user perceived QoS and the measured (technical) QoS parameters. Customer complaints and customer satisfaction queries are a common indicator of end-users' (dis-)satisfaction with a service. Furthermore, there are indications that the end-users consider the technical QoS as satisfactory most of the time and react (negatively) only when one or more of the QoS parameters of the actual service deteriorate beyond its normal/agreed values. To get a simple measure of the quality of the service provided, the concept of service failures are introduced. It is hypothesised that the characteristics (in terms of frequency, duration, etc.) of the service failures will correlate well with the customer satisfaction and customer complaints, and may provide a good indicator of end-user perceived QoS.

Acknowledgements

The author would like to thank Ola Espvik, Albert Lee, Mechthild Stör and Finn Arve Aagesen for comments on an earlier version of the manuscript.

References

1 ITU-T. *Terms and definitions related to Quality of Service and Network Performance including Dependability.* ITU-T Recommendation E.800, August 1994.

2 ETSI TC-NA. *Network Aspects : general aspects of Quality of Service and Network Performance.* ETSI Technical Report ETR 003 (ref. RTR/NA-042102), October 1994.

3 FITCE Study Commission. *The study of Network Performance considering Customer Requirements, Final Report,* 1993.

4 Aagesen, F A. QoS frameworks for open distributed processing systems. *Telektronikk,* 93, (1), 1997. (This issue.)

5 ISO/IEC JTC1-SC21. *Quality of Service Framework, Working Draft No. 2,* SC21-WG1 meeting, Yokohama, 1993.

6 Lee, A. *An overview of world Quality of Service activities at April 1996 prepared for EURESCOM P603,* June 14, 1996.

7 EURESCOM. Project P307 Reliability engineering; Deliverable D3. *An approach to dependability modelling of telecommunication networks and services,* November 1995.

8 ETSI TC-NA. *Network Aspects : Quality of Service indicators for Open Network Provision (ONP) of voice telephony and Integrated Digital Network (ISDN),* ETSI Technical Report ETR 138 (ref. DTR/NA-049001), July 1994.

9 Lee, A. *Publication of comparable performance indicators for operators in the UK telecommunications industry,* The Telecommunication Managers Association (TMA), January 1996.

10 Lee, A. Publication of comparable performance indicators for telecommunications Companies. *News and Views,* March 1996

11 Delamere, G. Comparing telecoms companies' performance. *Consumer Policy Review,* 6, (1), 7–10, 1996.

12 *American National Standard Method for the Calculation of the Articulation Index.* ANSI standard S3.5-1969, 16 January 1969.

13 Aagesen, F A. *Quality of Service in layered architecture.* Trondheim, SINTEF, 1994. (SINTEF report no. STF40 F94088.)

14 Aagesen, F A. Architectures for the modelling of QoS functionality, *Telektronikk,* 91, (2/3), 56–68, 1995

15 Aagesen , F A, Hellan, J K. On constructive models for QoS functionality. *International Teletraffic Seminar '95,* Bangkok, November 1995, p. 50-1 to -12.

16 Nilsson, G, Dupuy, F, Chapman, M. An overview of the telecommunication information networking architecture. *TINA'95 Conference,* Melbourne, 1995.

17 TINA-C (Telecommunication Information Networking Architecture Consortium). *Quality of Service framework,* Doc. no. TR_MRK.-001_1.0_94, TINA consortium, 1994.

QoS frameworks for open distributed processing systems

BY FINN ARVE AAGESEN

A *service* is the behaviour of some functional capability. *Quality of service* (QoS) is a measure of the relative frequency of specific events or duration of times between specific events within a service, used as quality criteria for proper service functioning. A QoS framework contains the constructive aspects related to QoS parameters, contracts and functionality, comprising specification, implementation, execution and management.

This paper discusses state-of-the-art for QoS frameworks for distributed systems. After a short review of the ISO OSI and ITU ISDN QoS frameworks, the OSI ODP (Open Distributed Processing) and TINA (Telecommunication Information Networking Architecture) QoS frameworks are presented and discussed. A short summary of relevant research related to distributed systems and constructive aspects of QoS are given.

1 Introduction

High capacity networking and processing and the emergence of new types of information teleservices have impacted system architecture standards. The OSI Reference Model has played an important role as a framework for system interconnection and has made a basis for interoperability. The motivations behind the development of OSI were limited in scope. Object-based architectures are now emerging and new dimensions are considered, like e.g. genericity, portability, reuse, manageability, ease of development, ease of integration, dynamic binding and safe interactions. Object-based distributed computing is being established as a basis for the support of teleservices. Several international organizations, such as ITU, ISO, OMG (Object Management Group), OSF (Open Service Foundations) and TINA-C (Telecommunication Information Networking Architecture Consortium) are currently defining similar object-based frameworks as a basis for open distributed computing.

QoS has been an aspect of both OSI and the ISDN specifications [1]. For OSI, a QoS Framework is under establishment. For ISDN, QoS is handled both in the generic documents I.350 [17] and E.800 [18], in specific B-ISDN documents and in the ATM Forum UNI specification [3]. ITU does not specifically address QoS in the Q.1200 recommendations for Intelligent Networks.

For distributed systems, QoS is undoubtedly an important aspect. However, flexibility and openness has more focus than has efficiency. This may be compensated by appropriate QoS handling functionality. QoS and QoS constraints are expressed to be important aspects of these architectures. But as QoS is not the top priority aspect, it remains to see what the outcome will be.

The objectives of this paper is both to give an insight into the problem domain of QoS related to distributed systems and to describe status for QoS frameworks. Section 2 defines basic concepts related to QoS. Section 3 gives a short summary of the OSI and ISDN QoS frameworks, previously handled in [1]. Section 4 presents the ISOs reference model for ODP (Open Distributed Processing) and the QoS aspects of this model. Section 5 presents TINA (Telecommunication Information Networking Architecture) and the QoS aspects of this architecture. Section 6 gives a summary of relevant research related to QoS in distributed systems. An overall summary is given in Section 7.

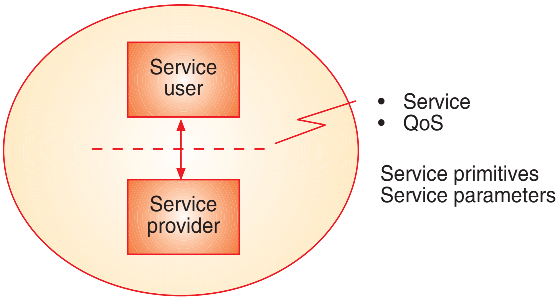


Figure 2.1 A generic service relationship

2 Basic concepts

2.1 Service

Within the generic framework of distributed systems, a service is defined as *the behaviour of some functional capability provided by a service provider to a service user* (Figure 2.1). The generic service can be defined by service primitives, and these service primitives can carry service parameters.

Within the context of OSI, a *service* is related to a specific layer, and an (N)-service is the service offered to the (N+1)-layer. ITU defines *teleservices* and *bearer services*. A teleservice is a service that the user gets from the user terminal, while a bearer service is a service offered at some interface between the user and the network. Within ISOs Reference Model for Open Distributed Processing (RM-ODP) [16], and also within TINA, it is focused on application services provided by *application objects*.

2.2 QoS and QoS parameters

Quality of Service (QoS) is a *measure* of the relative frequency of specific events or duration of times between specific events within a service, used as a quality criteria for proper service functioning. ITU defines QoS as “the collective effect of service performance which determine the degree of satisfaction of the user of the service” ([17],[18]). In the “Quality of Service Framework"” for OSI [15], QoS is defined as “a set of qualities related to the provision of an (N)-service, as perceived by an (N)-service-user”. RM-ODP and TINA define QoS as “a set of quality requirements on the collective behaviour of one or more objects”. There is no conflict between these definitions. These seemingly different definitions reflect the different system architecture models considered within the context of ISDN, OSI and TINA, respectively. QoS will be an aspect of models on various abstraction levels, related to:

- Stakeholders (users, subscribers and service providers)
- Administrative domains
- Hierarchical layers of software and hardware functionality
- Programmatic as well as hardware reference points
- Traffic sources and resources.

In the OSI QoS framework, concepts such as QoS *user categories*, QoS *characteristics* and QoS *parameters* are defined (Figure 2.2). A *QoS user category* is “a policy objective that leads to the identification of a set of QoS characteristics”. The basic idea is to identify various classes of users and to define the QoS requirements based on these classes. The QoS user categories defined are: the *secure,* the *safe,* the *time critical,* the *highly reliable,* the *easy to use*, the *extensible/flexible*, the *low cost* and the *QoS monitorable/testable/auditable*. A *QoS characteristic* is "a quantifiable aspect of QoS, which is defined independently of the means by which it is represented or controlled". QoS characteristics are intended to be used to describe the actual behaviour of systems. A *QoS parameter* is "a variable relating to one or more QoS characteristics, values of which are conveyed between objects as a part of a QoS mechanism". The set of generic QoS characteristics defined are: *time delay, time variability, time window, capacity, accuracy, protection, cost, priority, availability, reliability* and *coherence*. The definitions are given in the paper by B.E. Helvik [11] in this issue of *Telektronikk.*

QoS parameters are classified as *QoS requirement parameters* and *QoS data parameters*. *QoS requirements* are a statement of requirements for one or more QoS characteristics. Vogel [15] classifies QoS parameters as:

- User-oriented
- Cost-oriented
- Synchronization-oriented
- Format and coding-oriented
- Traffic-oriented.

User-oriented parameters are the parameters describing the user's subjective apprehension of the image, sound and speech quality. *Cost-oriented* parameters are related to charges. *Synchronization-oriented* parameters are parameters related to the beginning of image and sound in multimedia and simple image/ sound services. *Format-oriented* parameters are video resolution, frame rate, storage format and coding and compression schemes parameters. *Traffic-oriented* parameters are parameters varying with traffic load. QoS is directly related to the use of common traffic resources. Examples of *traffic resources* are: nodes, transmission capacity, transmission links, routes, logical channels, buffers, windows, and also system internal and processing resources such as CPUs, buses and interface-circuits within nodes and end-systems. The quantitative measure of QoS is directly related to the utilization of the resources involved in providing the service. So, *traffic performance* and *QoS* are two strongly related concepts.

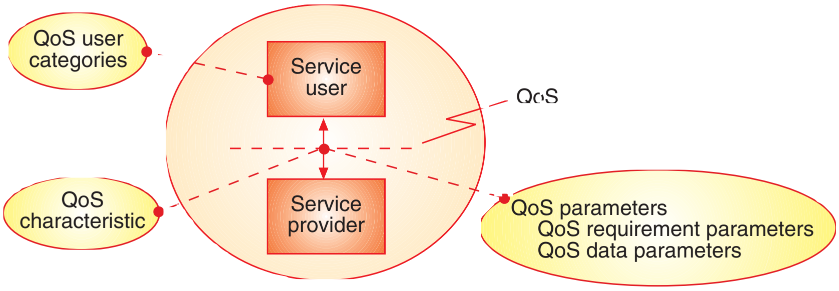


Figure 2.2 QoS user categories, characteristics and parameters

2.3 QoS – functionality, life-cycle and system components

Various aspects of *QoS handling functionality* are illustrated in Figure 2.3. QoS handling functionality will comprise: *assessment* of QoS *requirements* in terms of user's subjective wishes or satisfaction with the quality of the application: performance, synchronization, cost, and so forth. The assessment result must further be *mapped* into QoS parameters for various system components/ layers. There must be some *negotiation* between system components/layers to ensure that the QoS can be met before necessary resources are allocated. During the session there must be some *control* of user traffic parameters as well as QoS, and QoS must be renegotiated if needed. Finally, resources are *deallocated.*

Generic functionality has *life-cycle aspects* such as: user perception, requirement specification, functional specifi-

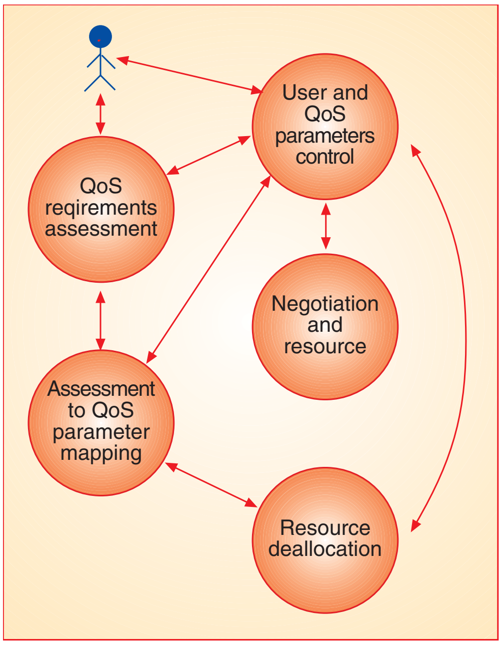

User and QoS parameters control
QoS reqirements assessment
Negotiation and resource
Assessment to QoS parameter mapping
Resource deallocation

Figure 2.3 QoS handling functionality

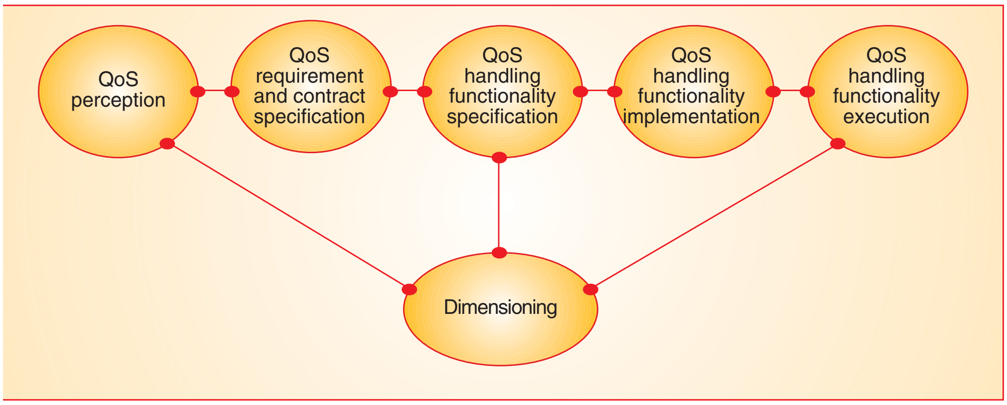

QoS perception
QoS requirement and contract specification
QoS handling functionality specification
QoS handling functionality implementation
QoS handling functionality execution
Dimensioning

Figure 2.4 Life-cycle related aspects of QoS handling functionality

cation, construction, implementation and execution. For QoS handling functionality we here consider the aspects: *QoS perception, QoS requirement and contract specification, QoS handling functionality specification, QoS handling functionality implementation* and *QoS handling functionality execution*. However, QoS is also strongly related to the very important *dimensioning* aspect. By varying the number or size of traffic resources, QoS is influenced. The *QoS handling functionality* and the *dimensioning* must therefore be considered as *a whole*. These aspects are illustrated in Figure 2.4.

The "distributed processing view" of QoS is more comprehensive than the "networking views" of QoS, as defined for OSI and ISDN. The specification language aspects introduced by ODP and TINA are not considered for OSI and ISDN. However, the principal elements of *the QoS handling functionality* are quite analogous. For B-ISDN, *QoS handling functionality* is classified as either *traffic control* or *congestion control* functionality [1]. Traffic control is the set of actions taken by the network to *avoid* congested conditions while congestion control is the set of actions taken by the network to *minimize* the intensity, spread and duration of congestion. In networks "up to LAN", the traffic control functionality was: *switching, multiplexing, routing, access, priority control* and *ack-based flow-control*. With high-capacity networks came transport protocols with *rate-based flow-control*. With B-ISDN came concepts such as: *connection admission control* (CAC), *usage/network parameter control* (UPC/NPC) and *traffic shaping*. CAC is the set of actions taking place during call set up in order to establish a *traffic contract* and a connection, UPC/NPC is the set of actions taken by the network to monitor and control traffic, and *traffic shaping* comprises modification of the traffic characteristics.

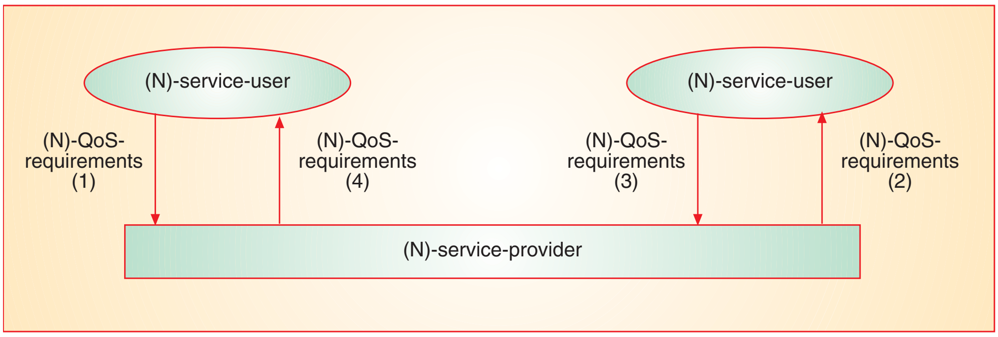


Figure 3.1 Flow of QoS requirements in a confirmed (N)-service facility

Table 3.1 OSI QoS requirement parameter semantics

- **Target:** the value is a desired level of QoS characteristic.
- **Lowest quality acceptable:** the value is a minimum level of quality that is acceptable to the user.
- **Guaranteed:** the desired QoS must be guaranteed by reservation of resources (or other means) so that the requested level will be met.
- **Compulsory:** the achieved QoS must be monitored by the provider and the communication aborted if it degrades below the compulsory level. However, the desired QoS is not guaranteed, and it may be deliberately degraded.
- **Threshold:** The achieved QoS must be monitored by the provider. Either or both users are warned if it degrades below the threshold level.
- **Controlled highest quality:** The provider and users must constrain their operation so that defined QoS level is not exceeded.

There is also a *system component view* of QoS handling functionality. The QoS parameters and QoS handling functionality can be related to various *system components*. Vogel [26] has a discussion of the various *technical aspects* of the distributed processing view of QoS. Aspects discussed are:

- User interface
- End-system (hardware, operating system, communication protocols)
- Encoding
- Communication protocols
- File-servers
- Databases
- Multimedia document modelling.

Such a system-component related consideration is not part of whether the OSI, ISDN, ODP or the TINA QoS frameworks.

3 A summary of the OSI and ISDN QoS frameworks

The OSI and ISDN QoS *architectures* are constituted by entities performing *QoS handling functions*, which behaviour is related to *QoS parameters*. The OSI entities are layer entities, while the ISDN entities are abstractions of signalling capabilities within the user end-system or within the exchanges in the network. B-ISDN defines a *traffic contract* between the user and the network, and the B-ISDN framework defines a *transmission-oriented* service based on the *traffic contract*. This is different to OSI, which defines an *application-oriented* service based on QoS requirement parameters [1].

3.1 The OSI QoS framework

The OSI QoS framework [15] is a supplement to the description of QoS contained in the Basic Reference Model. In addition to the definition of QoS-related *concepts,* it contains the definition of a *model* of QoS and the definition of QoS parameter *semantics*. The OSI concepts: *QoS user category, QoS characteristic* and *QoS parameters* were defined in Section 2.2.

The external flow of QoS requirements in a confirmed (N)-service-facility is illustrated in Figure 3.1. The model of QoS defines two types of entities performing functions related to the operation and management of QoS, namely *layer* and *system* QoS entities. *Layer* QoS entities are associated with the operation of a particular (N)-subsystem. *System* QoS entities have a system-wide role.

QoS parameter *semantics* is an important issue. The QoS parameters are classified as *requirement* and *data* parameters. Some requirements are classified as *weak* (best effort) requirements and some as *strong* requirements. The classification scheme is illustrated in Table 3.1. The "target" and "lowest quality acceptable" types are considered as "best effort" semantics. For one session, both "compulsory", "threshold" and "controlled highest quality" values can be defined for the same parameter.

3.2 QoS in present OSI recommendations

QoS is related to various layers within the OSI reference model. QoS parameters related to service primitives and PDUs are illustrated in Figure 3.2. The application layer consists of "Specific Application Service Elements" (SASEs) and "Common Application Service Elements" (CASEs). The *only* application service element that has explicitly defined QoS parameters is the CASE element ACSE (Association Control Service Element). A QoS parameter field is defined on the A.ASSOCIATE service primitives. These parameters are just sent further down to the transport layer.

The transport service provides transparent transfer of data between transport service users. There is both a connection-oriented and a connection-less service, but only a connection-oriented protocol. The transport service provides for a QoS selection. QoS is specified through the selection of values for QoS parameters representing characteristics such as *throughput, transit delay, residual error rate* and *failure probability*.

The QoS parameters for the *connection-oriented* case is shown in Table 3.2. *Throughput* and *transit delay* are defined for each direction of transfer on a transport connection. In each direction both *maximum* and *average* values are specified. The QoS parameters for the *connection-less service* are a subset of the connection-oriented parameters. They comprise *transit delay, protection, residual error probability* and *priority*.

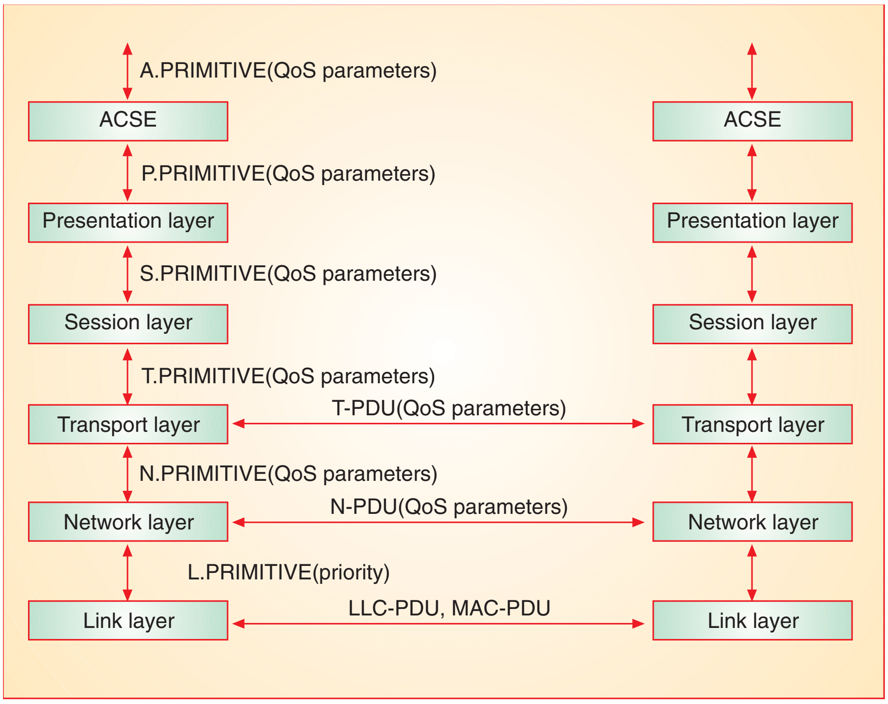


Figure 3.2 OSI QoS parameters in LAN-environments

The OSI service can be characterized as *application-oriented*. This is because the (N+1)-layer defines its QoS requirement parameters based on (N+1)-layer concepts, and there is an intention of a downwards mapping of the application-defined QoS requirements. There is a consistent set of session, transport and network service QoS requirement parameters. There is a *concept of a totality*, but some important elements are missing: 1) The QoS parameters of the application layer service are not well-defined. 2) The transport layer is the vital point for handling the end-to-end QoS, but there is not defined any transport layer functionality to handle this responsibility except the delegation of this responsibility to the underlying network layer. 3) The network layer does not have the defined functionality to take this responsibility. 4) In addition, the need for parameters describing other aspects of the traffic pattern "offered" than the aspects reflected by the defined "requirements", is not explicitly expressed.

Table 3.2 Connection-oriented transport service QoS parameters

Phase	Performance criterion:				
	Speed (time delay)	Accuracy/reliability	Others		
TC establishment	TC establishment delay	TC establishment failure probability (misconnection / TC refusal)	Protection	Priority	Cost
Data transfer	Throughput	Residual error rate (corruption, duplication/loss)			
	Transit delay	Resilience of the TC			
		Transfer failure probability			
TC release	TC release delay	TC release failure probability			

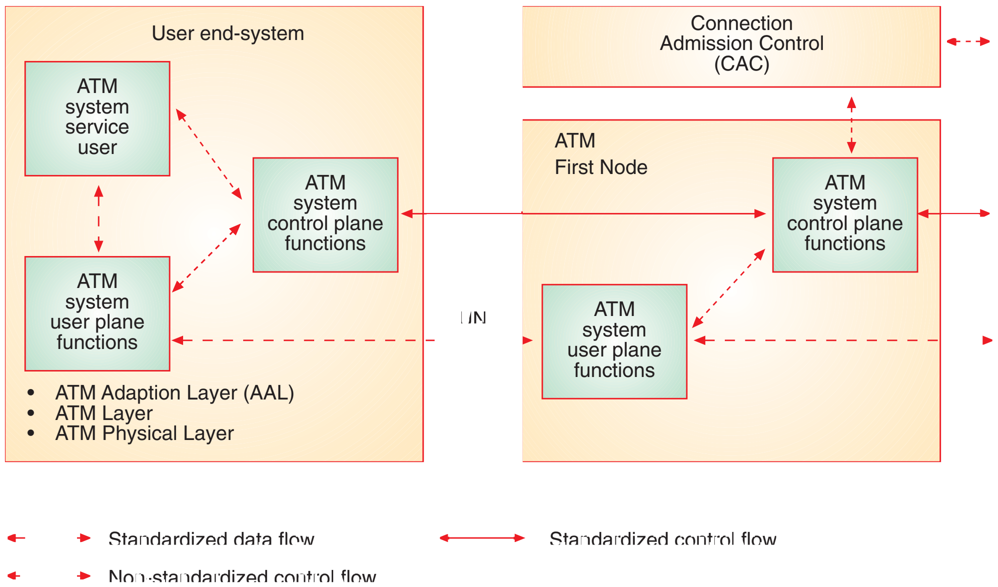

User end-system
ATM system service user
ATM system control plane functions
ATM system user plane functions
ATM Adaption Layer (AAL)
ATM Layer
ATM Physical Layer
UNI
Connection Admission Control (CAC)
ATM First Node
ATM system control plane functions
ATM system user plane functions
Standardized data flow
Standardized control flow
Non-standardized control flow

Figure 3.3 B-ISDN generic architecture

3.3 The ISDN QoS framework

ISDN is here used as a common concept for narrowband (N-ISDN) and broadband ISDN (B-ISDN). The ISDN functional architecture defines a *user plane* for data-transfer, a *control plane* for call and connection control and a *management plane* for management. Important QoS-related aspects are defined in several documents (see [1]). The ITU document I.350: "General Aspects of QoS and Network Performance in Digital Networks, Including ISDN" [17] was intended for N-ISDN. The generic aspects of I.350 are also applicable to B-ISDN. I.350 defines a *method* for identifying QoS parameters. QoS is not very much visible in the *specific* N-ISDN recommendations. The document I.356 "B-ISDN ATM Layer Cell Transfer Performance" is a supplement to I.350. A model for the B-ISDN functionality is shown in Figure 3.3. The ATM system has *five* different service and protocol concepts: 1) the ATM system service classes, 2) the AAL protocol types, 3) the ATM layer protocol class, 4) the ATM system signalling protocol type, and 5) the ATM layer QoS classes.

The User Network Interface (UNI) plays an important role within the B-ISDN QoS framework. A *traffic contract* specifies the negotiated characteristics of an ATM layer connection at a private or public UNI, consisting of:

- Source traffic parameters
- Cell Delay Variation (CDV) tolerance
- Conformance definition
- Requested *QoS class* for each direction of the ATM connection
- The definition of a *compliant* connection.

Peak Cell Rate, Sustainable Cell Rate and *Burst Tolerance* are potential source traffic parameters. The contract items related to these source traffic parameters as well as Cell Delay Variation tolerance and Conformance are all based on:

- GCRA (Generic Cell Rate Algorithm)
- Reference Models.

GCRA consists of two parallel working algorithms: a *Virtual Scheduling* algorithm and a *Leaky Bucket* algorithm. The Virtual Scheduling algorithm controls the time between individual cells, and the Leaky Bucket algorithm controls the accumulation of cells. GCRA has two parameters: I and L. I is the increment parameter (specified time between cells) and L the limit parameter (specified time period for accepting burst of cells). The Leaky Bucket leak-rate is 1 per cell time, the limit is L and the increment per event is I.

Peak Cell Rate is defined according to the reference model illustrated in Figure 3.4. The Peak Cell Rate of an ATM connection is defined as the inverse of the minimum interarrival time T (peak-emission interval) between the appearance of two ATM-PDUs. The *Cell Delay Variation* tolerance is the upper bound on the "cell clumping" measure. The tolerance allocated to a particular connection at the

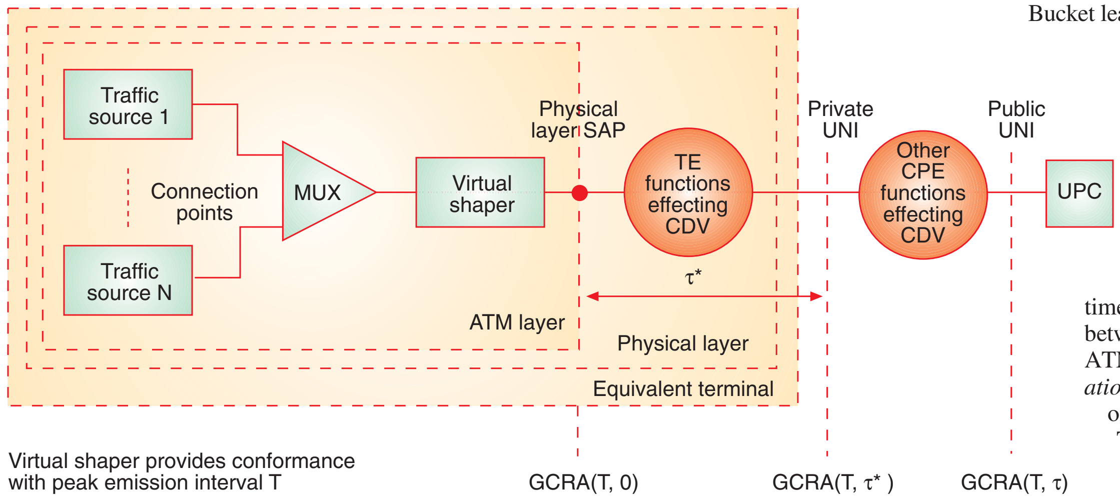


Figure 3.4 Peak cell rate reference model

private UNI is denoted by the symbol τ^* and at the public UNI by τ.

Sustainable Cell Rate and *Burst Tolerance* defined according to reference models *similar* to the model illustrated in Figure 3.4. *Conformance* applies to cells as they pass the UNI and are in principle tested according to some combination of GCRA algorithms. *Conformance Definition* is different from the *Usage Parameter Control* algorithm. The network provider may use any Usage Parameter Control as long as the operation does not violate the QoS objectives of compliant connections.

The B-ISDN *service* can be characterized as *transmission-oriented*. The ATM system service class is the consequence of a selected ATM layer *QoS class* and an AAL *protocol class*. The ATM system user needs to know the concepts of cell-time, cell delay variation, burst tolerance and GCRA.

4 The ISO reference model for open distributed processing systems (RM-ODP)

Among the various object-based frameworks for teleservices mentioned in Section 1, the ODP Reference Model for Open Distributed Processing (RM-ODP) [16] claims to be the most generic architecture. ODP is suitable for a wide range of applications. It defines a framework for service and application development in a heterogeneous environment based on:

- *Viewpoints,* in order to simplify the description of complex systems
- *Concepts* for the viewpoint specifications
- A model for an *infrastructure.* An important task for the infrastructure is the support of distribution transparencies which are important features of the ODP framework.
- Principles for assessing *conformance* to ODP systems.

QoS and QoS constraints are important aspects of the standard, but there is currently *no* specific QoS framework. RM-ODP consists of four parts: Part 1) *Overview,* Part 2) *Foundations,* Part 3) *Architecture* and Part 4) *Architectural Semantics*. Part 1 provides a tutorial introduction and gives a guidance to the understanding of the model. Part 2 defines the basic concepts and the analytical framework. It introduces the notion of viewpoint and defines the concept and vocabulary that are common to all ODP viewpoints. The architecture defined in Part 3 is a prescriptive model. It defines the five ODP viewpoints and the ODP functions. Part 4 contains a formalization of the ODP concepts and viewpoints.

This section has four subsections. In 4.1, the two ODP structuring approaches *viewpoint* and *transparency* are defined. In Section 4.2, viewpoint independent modelling concepts are presented. One of these viewpoint independent concepts is the *contract.* A specialisation of the contract denoted as the *environment contract* is presented in Section 4.3. Section 4.4 gives a more detailed presentation of the viewpoints and with special attention to QoS aspects.

4.1 Viewpoints and transparencies

The complete specification of any non-trivial distributed system involves a large amount of information. Attempting to capture all aspects of the design in a single description is seldom workable. Most design methods aid to establish a coordinated set of models, each aimed to capture one facet of the design. As a **first** structuring approach, ODP defines five **viewpoints**, each associated with a viewpoint language. These five viewpoints are:

- The *enterprise* viewpoint, which is concerned with the business activity of the system
- The *information* viewpoint, which is concerned with the information that needs to be stored and processed in the system
- The *computational* viewpoint, which is concerned with the description of the system as a set of *objects* that *interact* at system *interfaces*
- The *engineering* viewpoint, which is concerned about the mechanisms supporting system distribution
- The *technology* viewpoint, which is concerned with the technical realization.

Viewpoint specifications represent various dependent specifications of the *same* system. Some aspects of one viewpoint model can be mapped into some aspect of another viewpoint model. This is illustrated in Figure 4.1.

ODP-RM does not discuss methods related to various life-cycle phases of open distributed system development. ODP has concepts that can be used within various methods. A viewpoint can follow a system through various phases. And also, it must be allowed to iterate viewpoint models through the life-cycle phases. The viewpoints must be considered as flexible tools, and *not* static restrictions.

The **second** structuring approach is taken by identifying a number of **transparencies.** When considering a distributed system, a number of concerns becomes

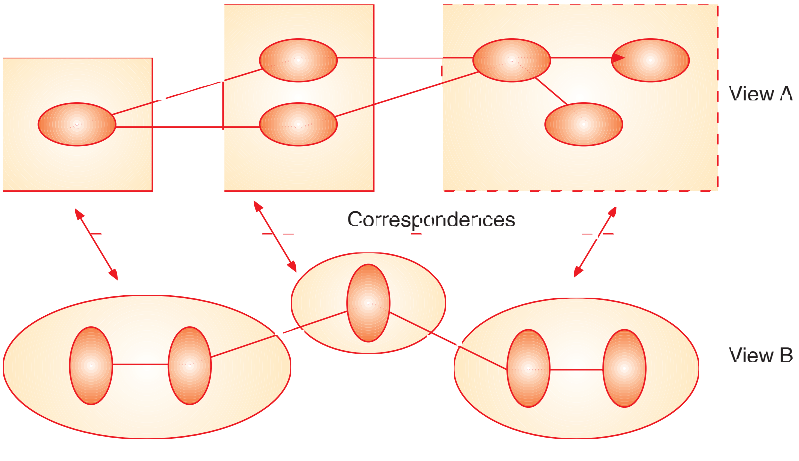


Figure 4.1 Correspondence between different views of a system

apparent which are a direct result of distribution: the system components are heterogeneous, they can fail independently, they are different, possibly varying locations, and so on. These concerns can be either solved directly as a part of the application design or standard solutions can be selected. If *standard* solutions are chosen, the application designer works in a world which is transparent to that specific concern. The standard mechanism is said to provide transparency. Application designers simply select which transparencies they wish to assume. The transparencies defined are:

- *Access transparency*, which mask differences in data representation and invocation mechanisms to enable interworking between objects
- *Failure transparency*, which mask from an object the failure and possibly recovery of other objects (or itself) to enable fault tolerance
- *Location transparency*, which mask from the use of information about location in space when identifying and binding to interfaces
- *Migration transparency*, which mask from an object the ability of a system to change the location of that object
- *Relocation transparency*, which mask relocation of an interface from other interfaces bound to it
- *Replication transparency*, which mask the use of a group of mutually behaviourally compatible objects to support an interface
- *Persistence transparency*, which mask from an object the deactivation and reactivation of other objects (or itself)
- *Transaction transparency*, which mask coordination of activities amongst a configuration of objects to achieve transparency.

4.2 Basic modelling concepts

There is a set of *general* concepts that give a common ground to the ODP viewpoint specifications, such as: *object, action, encapsulation, abstraction, behaviour, state, interface, service, composition/decomposition, template, role, type* and *class, class* and *subclass hierarchy, base class* and *derived class hierarchy, naming, group, domain, contract, liaison* and *binding*. This makes it possible to relate the viewpoint specifications to each other.

Objects and *actions* are the most basic modelling concepts. All things of interest are modelled as objects. Anything of interest that can happen is an *action.* The objects are involved in *interactions* resulting from their defined internal *behaviour.* The *behaviour* of an object is defined as the set of all potential actions an object may take place in. State and behaviour are related concepts. *State* characterises the condition of an object at a given instant.

An object offers *services* to its *environment.* Objects can only interact at *interfaces,* where an interface represents a part of the object's behaviour related to a particular subset of its possible interactions. An object can have several interfaces. *Contract* is a general concept for characterising and regulating the co-operation of objects. The contract is considered in the following section.

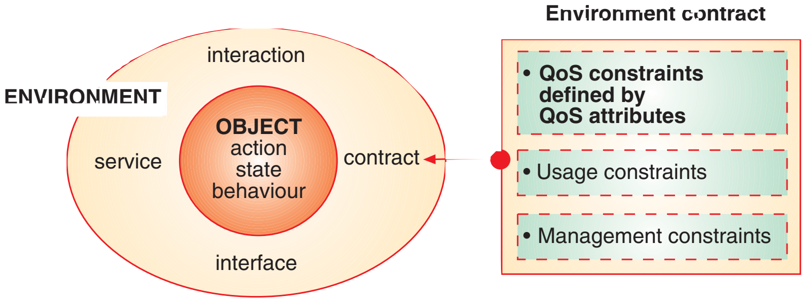


Figure 4.2 An object and its environment

4.3 The environment contract

A contract is an agreement that governs the cooperation among a number of objects. The contract is based on the idea of defining the obligation and expectations associated with the co-operating objects. As an example, the specification of a contract may include: specification of the different roles that objects involved in the contract may assume and the interfaces associated with the roles, QoS aspects of object cooperation, here denoted *QoS attributes*, indications of duration of periods of validity, the kind of behaviour that invalidates the contract and liveness and safety conditions.

The contract may be agreed, redefined and terminated. Whenever objects co-operate there is some contract between them. However, such contracts are often derivable from the viewpoint language rule and do not need to be stated explicitly. A particular kind of contract is the *environment contract* (Figure 4.2). This contract describes the requirements placed between the object and its environment and vice versa. An environment contract includes three classes of constraints: 1) QoS constraints, 2) usage constraints, and 3) management constraints. The *QoS constraints* include:

- *Temporal* constraints (e.g. deadlines)
- *Volume* constraints (e.g. throughput)
- *Dependability* constraints covering aspects of availability, reliability, maintainability, security and safety (e.g. time between failures).

4.4 QoS in the various viewpoints

The **enterprise** language introduces basic concepts necessary to represent an ODP system in the context of the enterprise in which it operates. It defines the stakeholders and other participating institutions and establishes the roles, describes the legal environment and the set of rules and regulations. Contracts are important parts of the enterprise viewpoint. *Environment contracts* with *QoS constraints* for the enterprise viewpoint objects are essential.

In the **information** viewpoint, an ODP system is represented in terms of information objects and their relationship. ODP is *not* yet prescriptive with respect to information modelling. The model and notation are not prescribed. ODP does not either indicate what kind of information that is relevant to this viewpoint. The information viewpoint can be very complex and large enough to encompass all other viewpoints. The common understanding for a *minimal information modelling* is that it suffices to take into account the information entities *directly supporting* the system being considered.

Some *QoS information* will appear in the information viewpoint. This can be information related to *performance management*, various *media* like voice, data and video, *sessions, connections* and *relationships* between users and user agents. The information viewpoint language will *not* have the concept of interface. The *QoS information* defined in this viewpoint will in the computational viewpoint appear as related to *interfaces* and *computational objects*.

The **computational** viewpoint is directly concerned with distribution. The system is composed into objects which are abstractions of distributed applications performing individual functions and interacting at well-defined interfaces. The computational specification thus provide the basis for decisions on how to distribute the jobs to be done, assuming communications mechanisms can be defined in the engineering specification to support the behaviour at those interfaces. It is focused on: *computational objects, interfaces* between objects and *interaction* between objects.

There are three kinds of *interactions:* 1) signals, 2) operations, and 3) flows (or streams). *Signals* represents the elementary building units for interactions. There are two kinds of *operations:* announcements and interrogations. Interrogations are analogous to procedure invocations. A *flow* is an abstraction of a sequence of interactions, resulting in the conveyance of information from a producer object to a consumer object. A flow represents a logically continuous information transfer. The semantics is application dependent and is not defined within RM-ODP.

A computational *interface* is characterized by 1) a *signature,* 2) a *behaviour,* and 3) an *environmental contract*. The signature depends on the interface type, which can be either operational, stream or signal. An *operation interface* has a signature that defines the set of operations supported at the interface and whether the interface has the role of the client or server for that set of operations.

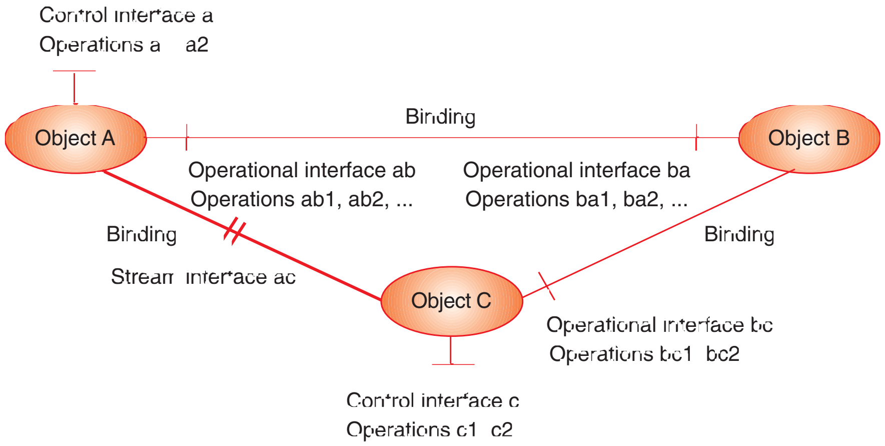


Figure 4.3 Computational objects and interfaces

A *stream interface* has a signature that defines the set of flows supported at the interface and for each flow, whether the interface has the role of producer or consumer. A *signal interface* has a signature that defines the set of signals supported at the interface and for each signal, whether the interface has the role of initiating or responding. A *computational viewpoint environment contract* specifies a set of QoS constraints placed on a computational object and its environment. A particular notation for specifying QoS is not prescribed by the computational language. Figure 4.3 illustrates computational objects bound by various interfaces.

There are *two* kinds of computational *objects,* namely *basic* objects and *binding* objects. Whenever computational objects are able to interact, a *binding*

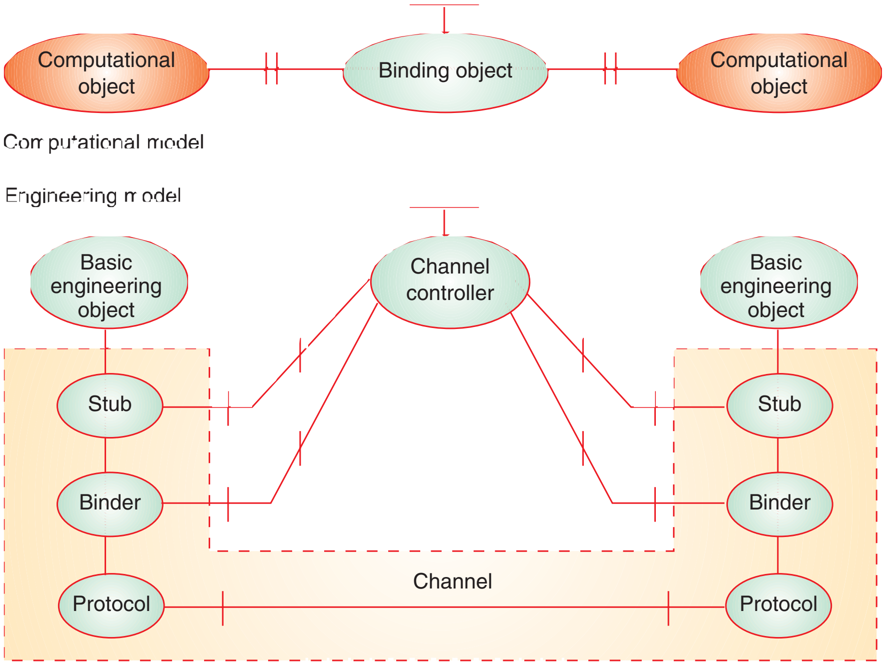


Figure 4.4 Explicit binding in the computational and engineering viewpoints

exists between them. Environment contracts provide a declarative way of specifying the QoS of a binding. It is also possible to actively manipulate QoS during the lifetime of a binding. Where this is desired, the binding must be modelled as a *binding object,* which may provide interfaces for controlling QoS. When a binding object exists, the binding is said to be *explicit.* The paper by Février et al. in this issue of *Telektronikk* gives a QoS specification of ODP binding objects. Figure 4.4. illustrates the relationship between the computational and engineering model of generic explicit binding.

The **engineering** viewpoint gives an implementation language and operating system *independent* description of how the computational objects and their interfaces are implemented. The engineering viewpoint focuses on mechanisms for supporting distributed interaction between objects. This includes mechanisms for ensuring that an object meets its QoS obligations. Supporting the communication between computational interfaces, with specific QoS constraints, is an important function provided in the engineering viewpoint. QoS monitoring and QoS budget administration are also important engineering viewpoint QoS functionality. In Figure 4.4, a Computational Object maps into a Basic Engineering Object and a Binding Object maps into the Channel Controller. *Channel, Stub, Binder and Protocol* represent engineering model functionality. The *Stub* is responsible for application service to application level protocol mapping, the *Binder* is responsible for the application level entity association control, and the *Protocol* is responsible for the protocol adjustment to transport network.

The **technical** viewpoint describes the implementation. The generic engineering components are matched with existing pieces of hardware and software, thus providing real integrated products and implementations which comply with the specifications provided with the other viewpoints.

5 Telecommunication Information Networking Architecture (TINA)

The TINA Consortium (TINA-C) is an international collaboration aiming at defining and validating an open architecture for telecommunications *services* and *management* in the coming era of B-ISDN, multimedia and rapid service innovation. The architecture is based on distributed computing, object orientation, and other concepts and standards from the telecommunication and computing industries. The intention is to make use of recent advances to improve interoperability, re-use of software and specifications, and flexible placement of software on computing platforms/nodes. In addition, the *consistent* use of software methods and principles to both *service* and *management* functionality will be possible.

The TINA architecture is both broad in its scope and deep in its details. In the present versions, QoS has a certain role. In addition to being an aspect of various documents, a specific TINA QoS framework is defined. Section 5.1 gives a short presentation of the overall concepts and principles of TINA. Section 5.2 discusses various aspects of QoS as treated in various documents. Section 5.3 presents the TINA QoS framework. The presentation is based on [4], [6], [19], [22], [25].

5.1 TINA concepts and principles

The TINA functionality is decomposed according to the dimensions illustrated in Figure 5.1. These dimensions are: 1) computing layering, and 2) management layering. *Computing layering* considers telecommunication software as structured into three software layers above a *hardware resources* layer. These layers are:

- Telecommunications application layer
- Distributed Processing Environment (DPE) layer
- Native Computing and Communication Environment (NCCE) layer.

The TINA applications are the software actually implementing the capabilities provided by the system. They are located in the Telecommunications application layer. DPE is software supporting the distributed execution of telecommunication application. The DPE provides a technology independent view of computing resources and it also shields from applications the distributed nature of the system. Both these layers are designed in an object-oriented way. However, the implementation of TINA applications and DPE does not need to be based on object-oriented languages. NCCE contains the operating system, communications and other support found in computer systems.

In the *management layering* dimension there are: *services, resources* and *elements.* This layering is an adaptation of the TMN layering. The management layering is related to the partition of the software in the *Telecommunications application layer* as defined in the *com-*

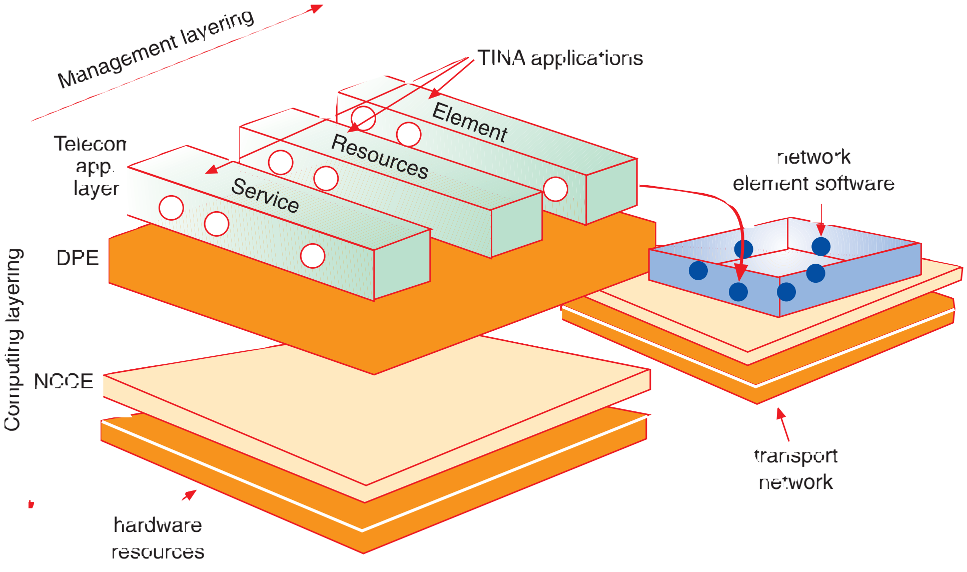


Figure 5.1 Layers and separation in TINA

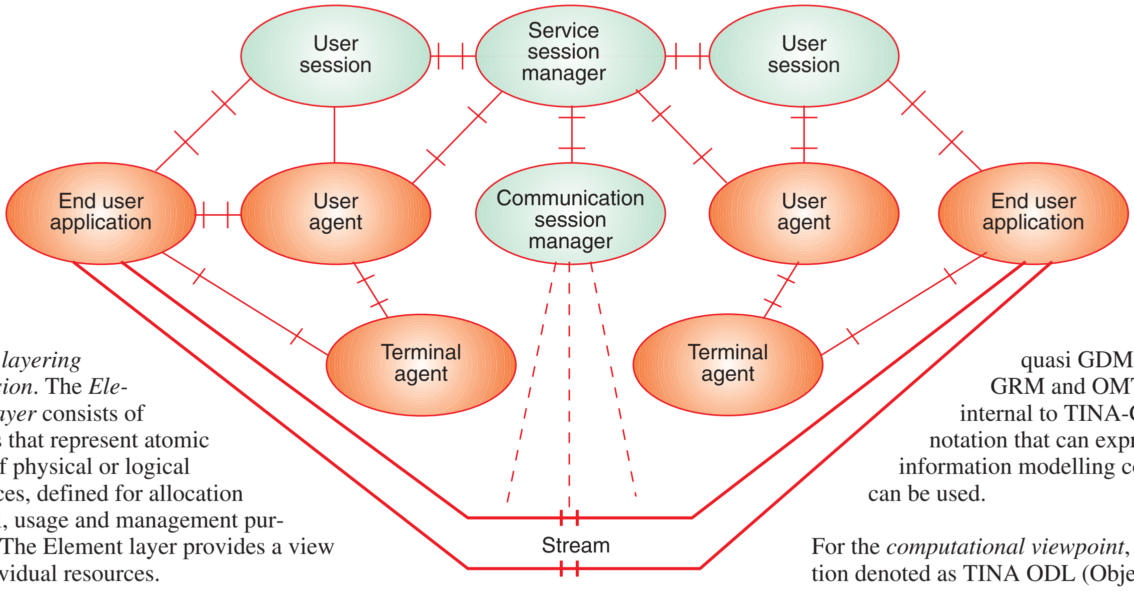


Figure 5.2 Access and session components

puting layering dimension. The *Element layer* consists of objects that represent atomic units of physical or logical resources, defined for allocation control, usage and management purposes. The Element layer provides a view of individual resources.

The *Resources layer* contains objects that maintain views and manipulate collection of elements and their relationships. It also provides the service layer with an abstract representation of elements. The Resources layer is responsible for providing technology independent views of elements, so as to be suitable for services. Anything that is represented in the Element layer should be handled by the Resources layer. *Service* layer consists of objects involved in the provision of services to stakeholders. Objects in this layer are either specific to a given service or are service-independent.

TINA also defines an **overall architecture**. This overall architecture governs *four* main subsets:

- Management architecture
- Service architecture
- Network architecture
- Computing architecture.

The concepts *computing* and *management* previous related to layering are *not quite consistent* with the corresponding concepts related to this *overall architecture*. The *Management architecture* defines a set of concepts and principles for the design, specification and implementation of software that are used to manage services, resources, software and underlying technology. The *Service architecture* defines a set of concepts and principles for the design, specification, implementation and management of telecommunication *services.* It further aims to define a set of reusable components from which to build telecommunication services. The *Network architecture* defines a set of concepts and principles for the design, specification, implementation and management of transport networks, where a transport network encompasses both transmission and switching technology. And finally, the *Computing architecture* defines a set of concepts and principles for designing and building distributed software and the software support environment.

5.1.1 Computing architecture

The Computing architecture defines the *modelling concepts* for the specification of *object-oriented* software. It also defines DPE. The Computing architecture refines and adapts the RM-ODP standard, so that it is suitable for the design of telecommunication systems. The *RM-ODP viewpoints* are important aspects of TINA Computing architecture.

For the *information viewpoint* a notation chosen for information specifications is a form of GDMO (Guidelines for the Definition of Managed Objects) with GRM (General Relationship Model). TINA uses only aspects of GDMO and GRM that are suitable for TINA information modelling. The notation is called quasi GDMO-GRM. The OMT (Object Modelling Technique) [24] graphical notation has been adopted for the diagrammatic representation. The use of quasi GDMO-GRM and OMT are internal to TINA-C. Any notation that can express the information modelling concepts can be used.

For the *computational viewpoint*, a notation denoted as TINA ODL (Object Definition Language) is chosen. ODL enhances OMG IDL (Object Management Group Interface Definition Language). Use of ODL notation is an internal TINA language. Any notation that can express the computational modelling concepts can be used. At present, no notation is used in TINA for *engineering viewpoint* specifications.

5.1.2 Network architecture

The purpose of the Network architecture is to provide a set of generic concepts that describe transport networks in a technology independent way, and also to provide mechanisms of the establishment, modification, and release of network connections. The Network architecture defines a set of *abstractions* that the Resource layer can work with.

One aspect of this Network architecture is *network layering and partitioning*. A network may be decomposed into subnetworks and links between them. Each subnetwork may be further decomposed into smaller subnetworks interconnected by links until the desired level of detail is reached. Transport networks can also be viewed as composed of layer networks.

The computational model that provides for the establishment, modification, and release of connections is called *connection management*. The connection management functionality reside in the *Resources layer*. One of these components is the *Communication session manager* that provides an interface to service software (Figure 5.2). This interface consists of operations to build and modify the *abstract* representations of *connections.*

5.1.3 Service architecture

Services and their environments are modelled according to the modelling concepts in the Computing architecture. There are *three* main sets of concepts in the service architecture:

- *Session* concepts
- *Access* concepts
- *Management* concepts.

The session concepts address service activities and temporal relationships. Access concepts address user and terminal associations with network and services. Management concepts address service management issues. These concepts are part of the *Management architecture*.

Even if services by their nature are different from each other, they all have a fundamental property in that they provide a context for relating activities. Such a context is termed a *session.* Four types of sessions have been identified: 1) *service* session, 2) *user* session, 3) *communications* session, and 4) *access* session. Service session represents a single activation of a service. A user session represents a single user's interaction with the service session. Communications session represents the connections associated with a service session, and access sessions represents a user's attachment to a system and the attachments involvement in services.

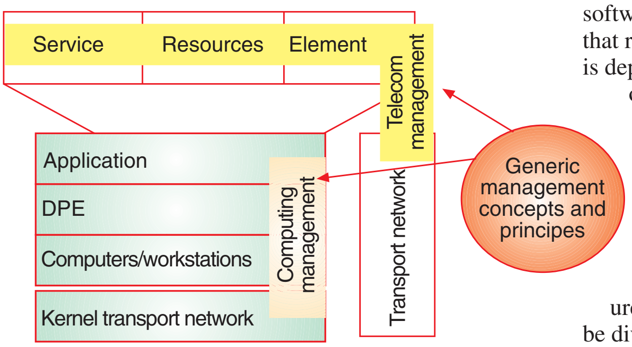


Figure 5.3 Two types of management

Users need to have flexible *access* to services, in terms of locations from which they access the service and the types of terminals they use. *User access* is therefore distinguished from *terminal access*. An agent concept is used in the access model. An agent is a computational object, or collection of objects, that acts on the behalf of another entity. Two types of agents have been identified: 1) *user agent,* and 2) *terminal agent*. A *user agent* is a computational object that represents and acts on behalf of a user. It receives requests from users to establish service sessions, or to join existing service sessions. A user agent also receives requests to join service sessions themselves. *Terminal agent* is a computational object responsible for a terminal. It is responsible for obtaining the precise location of a terminal. In order to access a service, a user must associate its user agent with a terminal agent. A user may simultaneously be associated with many terminals. So the creation of relationships between terminal agents and the services related to a user is a dynamic and challenging function. Figure 5.2 gives an illustration.

5.1.4 Management architecture

TINA management architecture provides the concepts and principles for functionality to manage the entities in TINA systems. As with the service architecture, the computing architecture is the methodology basis. A TINA system consists of a computing environment upon which *service, resource* and *element* applications run. In accordance with this separation, two classes of management are defined:

- Computing management
- Telecommunication management.

Computing management involves the management of NCCE, DPE, and the software (independent of functionality) that runs on the DPE. The main concern is deployment, installation and operation of software and computing nodes. Telecommunication management involves the management of the transport network, the management of the applications that use and control this network and the management of services. These relationships are illustrated in Figure 5.3. Management functionality can be divided into five aspects: fault, configuration, accounting, performance and security. The five areas of management are very broad in scope and have the same separation as found in OSI management.

5.2 TINA and QoS

Several generic QoS requirements are put upon the TINA architecture:

- It should support an ability to meet a wide range of QoS as they can be presented by new services.
- It should meet QoS targets for existing services as defined by the ITU-T standards pertinent to the service.
- Customer selection of QoS should be possible.
- Congestion and unavailability should be managed, i.e. the architecture should provide means to deal with unavailability of software resources and congestion.

TINA-C has started its definition of an approach to QoS in a QoS framework [25]. In addition, QoS aspects are part of "the other" functional TINA documents. Many aspects of QoS are immature. The QoS framework gives an indication of what may come. TINA addresses QoS mainly at the *computational* and *engineering* viewpoints. And *most* concern is related to *QoS requirement* and *contract specifications* (see Figure 2.4) and accordingly *language aspects*. In Section 5.2.1, QoS as handled in the present functional documents are presented. In Section 5.2.2, the QoS framework is presented.

5.2.1 QoS in the present functional specifications

Considering the **computational viewpoint**, *QoS attributes* play an important role. QoS attributes are associated with *operations* and *stream* flows. The attributes are *typed.* Examples of QoS attributes for *operations* are: maximum response time, maximum invocation time and maximum invocation interval. Examples of QoS attributes for *stream flows* are: throughput, maximum jitter and sampling rate.

TINA provides the concepts necessary to associate QoS service statements with the operations and flows. It does *not* provide the *semantics* of these QoS specifications as done in the OSI QoS framework described in Section 3.1. It is the intention that the semantics of QoS attributes shall be contained in the TINA QoS framework.

Computational viewpoint specification are based on TINA ODL (Object Definition Language). In the present version of ODL, QoS parameter semantics is left to the programmer. ODL allows for a *QoS service type* and variable identifier with any of *operation* and *flow*. Each operation and flow is followed by the keyword “with”, which in turn is followed by a datatype, and variable identifier, specifying parameters describing the QoS service associated with the operation or flow.

In the current version of ODL, *values* to the QoS variables cannot be given at specification time. The QoS value associated with the QoS type is expected to be assigned, or negotiated, when the interface is instantiated. Values can be renegotiated. Semantics are programmer dependent, but the semantics of each QoS type is expected to be documented as *text comments*.

Figure 5.4 gives an example specification. Two interfaces are specified, one operational and one stream interface. For the stream interface I2, QoS is offered using an instance of VideoQoS. An instance of VideoQoS is represented as a float, but depending upon the value of Guarantee (either Statistical or Deterministic) the float is to be interpreted as a “mean” or “peak” frame rate.

The treatment of QoS in the **engineering viewpoint** is related to the TDPE (TINA DPE) kernel support for QoS. Seen from the kernel, QoS can be either 1) end-to-end in nature, or 2) between an application and the TDPE kernel itself. In cases where the TDPE kernel is used to support QoS, it is the responsibility of the kernel to manage QoS negotiation and control between the applications and the underlying NCCE (Native Computing and Communication Environment). Applications negotiate for and are allocated a guaranteed level of service from the TDPE kernel.

A *QoS contract* is a formal agreement between an application and the TDPE kernel. In cases where the NCCE is not capable of handling the QoS requirements, TDPE may be needed to implement its own QoS support. A *QoS Support Model* is introduced. This model needs: 1) A QoS specification language, used to represent QoS offers and QoS requirements, 2) A mapping of the QoS requirements to the QoS offered by the TDPE kernel and NCCE, and 3) Mechanisms for reporting failures to meet QoS guarantees. Language for QoS specification is not decided. The mapping of application requirements to the QoS offered by the underlying NCCE enables the TDPE kernel to support a number of roles. These include: the monitoring of the existing QoS contracts, and reacting to changes in the offered QoS from underlying NCCE, deciding whether new contracts can be issued without affecting the state of the existing QoS contracts and notifying applications of failure to meet required QoS guarantees.

The QoS concepts introduced in the engineering viewpoint is application of the generic QoS concepts and functionality at this lower functionality level. The computational viewpoint QoS functionality is strongly related to the existing QoS support by the kernel and NCCE. So far, this relationship is not considered.

5.2.2 The QoS framework

The TINA Quality of Service Framework is a framework for specifying QoS aspects of distributed telecommunications applications in the *computational specifications* of these applications. It further outlines an *engineering framework* for the realization of QoS aspects of the computational specification. Several *dimensions* of QoS are defined, such as: *timeliness, high performance* and *dependability.* Each dimension of QoS has its own associated QoS parameters. It should be possible for application designers to use previously defined QoS parameters in specifications and also to define new QoS parameters as necessary. The concept of *QoS domain* is therefore introduced. A *QoS domain is defined as the computational model of a QoS dimension.*

A QoS domain contains a set of QoS attributes, where each attribute denotes a QoS guarantee that either is provided by the server of an interface or is required by the clients of the interface or vice versa. Based on these definitions, the QoS framework contains: extensions of the QoS aspects of the *Computing* architecture for the computational and engineering viewpoints, a draft definition of a *timeliness* dimension (denoted as the real time architecture) and a draft definition of an *availability* dimension (denoted as the availability architecture). Availability is one aspect of the more generic *dependability* dimension (Figure 5.5).

Note that a *dimension* comprises a *total view* of a system, while a *domain* only has a *computational view* of a system. The draft definitions of the timeliness and availability dimensions so far *only* have the *computational view* of these dimensions. This means that it only comprises the timeliness and availability *domains.* TINA QoS framework is well elaborated for the timeliness domain. The semantics is defined by means of a *Timing Behaviour Description Language – TBDL.* The language has *events* and *durations* as the two central concepts. It has been used to define the semantics of a number of QoS attributes for operational interfaces. Extending the framework to include flows is for further study. Some work has been done on the *availability domain*. In the following Section 5.2.2.1, extensions to the computing architecture are presented. The timeliness domain is presented in Section 5.2.2.2.

5.2.2.1 Computing architecture extensions

In the **computational** specification three kinds of entities are subjects for QoS-related specification: *interfaces, objects*

```
interface I1{
// QoS type
// max time server allowed for operation completion
typedef float Boundedresponsetime;
...
void operation (in datatype1 var1)
with BoundedresponseTime op1QoS;
..
};// end of I1

interface I2{
// QoS type
// mean or peak throughput
struct VideoQoS{
union Throughput Switch (Guarantee) /* in frames per sec. */
case Statistical : float mean;
case Deterministic: float peak;
};// end of VideoQoS
...
sink VideoFlowtype display with VideoQoS requiredQoS;
...
};// end of I2
```

Figure 5.4 QoS attributes example

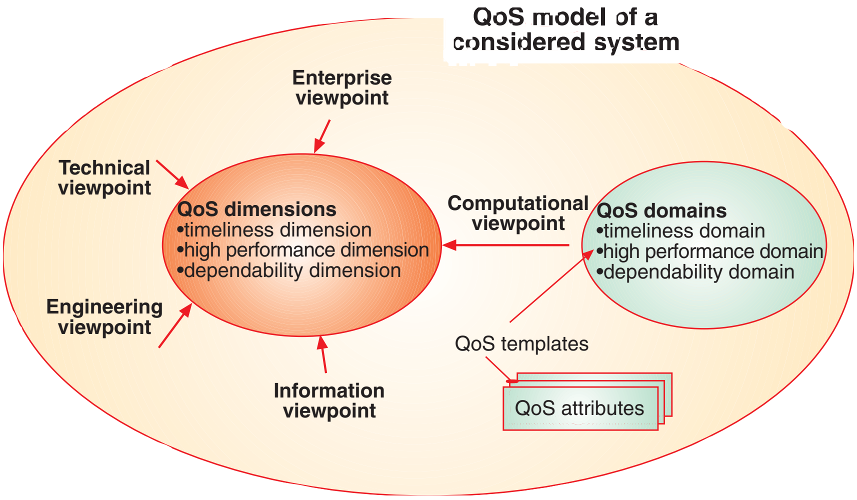


Figure 5.5 QoS domains and dimensions

and *packages.* A package is a set of objects. In the specification of an *individual interface*, the objective of including QoS specifications is to clearly specify the QoS parameters involved in interactions via the interface. In the specification of an *object,* the objective of including QoS specifications is to describe the QoS relationships that exist between the object and its environment. The objective of QoS specifications for *packages* is similar to the objective for object QoS specifications, except that the QoS relationships of interest here encompass a collection of objects and not just an individual object. QoS attributes are essential elements. A QoS attribute applies either to an entire object, an interface or an individual operation.

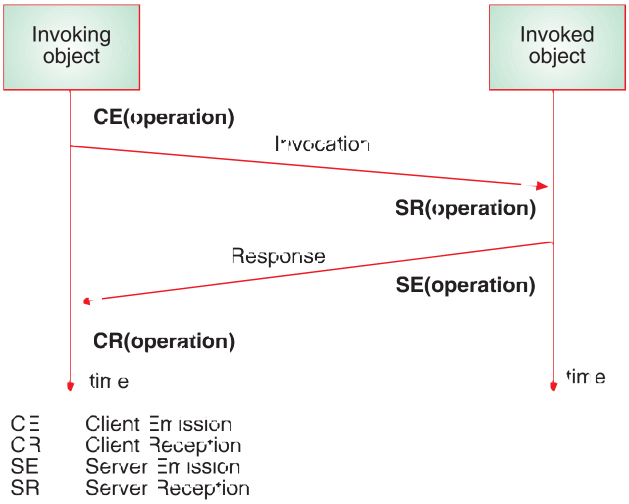


Figure 5.6 The four events associated with an invocation

It is stated that TINA-ODL should be refined to include specification structures for *QoS domain* specifications. A *QoS template* is a specification structure for specifying a QoS domain. A QoS template has the following structure:

<QoS template name>
<Name and type of each QoS attribute>
<Semantics of QoS attribute>

The semantics is described using concepts defined in the model on which the QoS domain is based. An object template can import one or more QoS template definitions. Each attribute defined in an imported QoS template is a service attribute in the importing object template. *Inheritance* is needed in the definition of QoS templates.

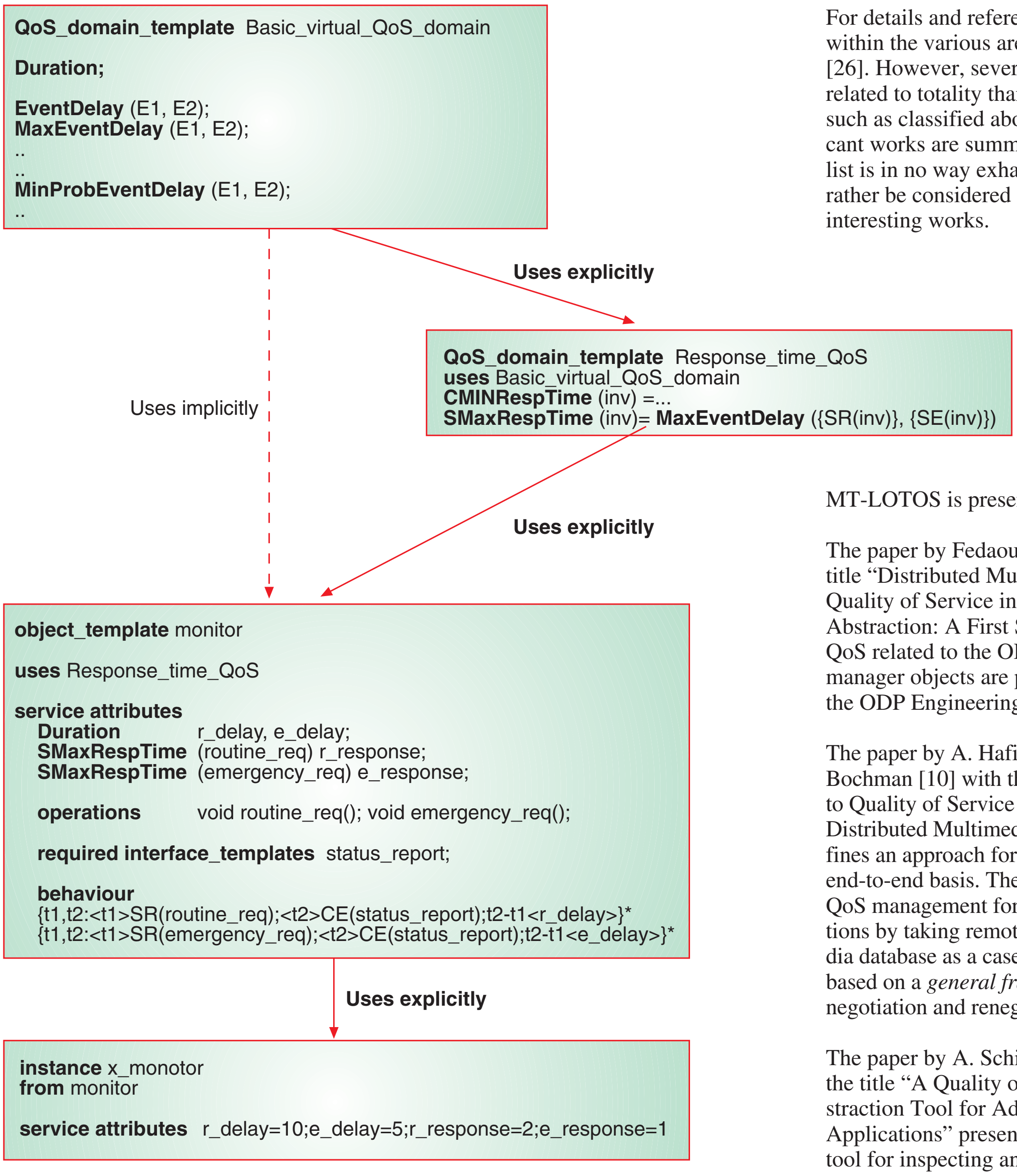
QoS aspects in the **engineering viewpoint** is concerned with which kind of support is required from the environments for realizing QoS guarantees. This is related to *resource management.* Hence, in the engineering viewpoint, one is interested in identifying the different resource managers, involved in provision of the QoS guarantees and how the various resource managers interact and cooperate in the provision of QoS guarantees. In the development of the TINA-C QoS framework, one is concerned with concepts needed for describing these aspects. TINA has defined an engineering QoS binding model similar to the ODP model in Section 4.4, but with the concepts QoS Binding object, QoS Binder Object, QoS Channel, QoS Channel Control Function, QoS Stub, QoS Binder and QoS Protocol Adaptor. These QoS objects are special cases of the similar RM-ODP objects.

5.2.2.2 Timeliness domain

The discussion of the timeliness domain is to some extent related to the TINA-TBDL language. TINA-TBDL formalises the content of the timeliness domain and intends to describe behaviour and timing relationships between invocations/announcements emitted and received by objects. Figure 5.6 shows the four events associated with an invocation. The similar two events associated with an emission is CE(operation) and SR(operation), where "operation" is the name (signature) of the operation.

The language will be extended to include stream interfaces at a later point in time by defining the necessary events relevant for stream interfaces. Behaviour description with TBDL can be given in: *Interface Templates, Object Templates* and *Package Templates*. TBDL enables the specification of timing requirements for activities spanning one or several interfaces in one or more objects. TBDL has similarities with constraint-oriented languages for temporal ordering and interaction such as MT-LOTOS, described in the paper by Février et al. in this issue of *Telektronikk* [21].

TBDL is based on a partial ordering of event occurrences. Events form the basic elements of the language. The language also has time and duration concepts. Each event occurrence can be associated with time references which are used to state timing constraints for a given behaviour. Several *real-time QoS attributes* are defined. These are *duration, delay bounds* for specific event pairs, *response time* for invocations, *message delay* for invocations, *emission delay, receipt delay, delay bounds* for event sets, *response interval, jitter* and *jitter bounds*. Figure 5.7 gives an example of interrelation of QoS domains templates and object templates with QoS attribute definitions based on TINA-TBDL.



```
QoS_domain_template  Basic_virtual_QoS_domain

Duration;

EventDelay (E1, E2);
MaxEventDelay (E1, E2);
..
..
MinProbEventDelay (E1, E2);
..
```

```
QoS_domain_template  Response_time_QoS
uses Basic_virtual_QoS_domain
CMINRespTime (inv) =...
SMaxRespTime (inv)= MaxEventDelay ({SR(inv)}, {SE(inv)})
```

```
object_template monitor

uses Response_time_QoS

service attributes
   Duration          r_delay, e_delay;
   SMaxRespTime  (routine_req) r_response;
   SMaxRespTime  (emergency_req) e_response;

   operations        void routine_req(); void emergency_req();

   required interface_templates  status_report;

   behaviour
   {t1,t2:<t1>SR(routine_req);<t2>CE(status_report);t2-t1<r_delay>}*
   {t1,t2:<t1>SR(emergency_req);<t2>CE(status_report);t2-t1<e_delay>}*
```

```
instance x_monotor
from monitor

service attributes  r_delay=10;e_delay=5;r_response=2;e_response=1
```

Figure 5.7 Example of QoS domain templates

6 QoS-related research activities

There is a significant research activity related to the topic "QoS in distributed systems". The topic is handled within a broad spectrum of conferences. The IFIP Fifth International Workshop on Quality of Service (IWQoS'97) is here mentioned as an example.

The paper by Vogel et al. [26] with the title "Distributed Multimedia and QoS: A Survey" gives an overview of state-of-the-art by 1995. An overview of research areas and works are given. The research areas were classified as:

- User interface and user issues
- End-system and operating system
- Encoding
- Communication protocols
- File-servers
- Multimedia document modelling and multimedia queries.

For details and references to papers within the various areas, it is referred to [26]. However, several papers are rather related to totality than to specific topics such as classified above. Some significant works are summarised below. This list is in no way exhaustive and should rather be considered as examples of interesting works.

First, the paper by Février et al. in this issue of *Telektronikk* [21], where a complete QoS specification of ODP binding objects using MT-LOTOS is presented.

The paper by Fedaoui et al. [7] with the title "Distributed Multimedia Systems Quality of Service in ODP Framework of Abstraction: A First Study" discusses QoS related to the ODP viewpoints. QoS manager objects are proposed as part of the ODP Engineering Viewpoint.

The paper by A. Hafid and G. v. Bochman [10] with the title "An approach to Quality of Service Management for Distributed Multimedia Applications" defines an approach for QoS support on an end-to-end basis. The paper deals with QoS management for multimedia applications by taking remote access to multimedia database as a case study. This case is based on a *general framework* for QoS negotiation and renegotiation.

The paper by A. Schill et. al. [24] with the title "A Quality of Service Abstraction Tool for Advanced Distributed Applications" presents a management tool for inspecting and controlling QoS characteristics on top of XTPX (eXpress Transfer Protocol Extended) and TCP/IP. The design and implementation of a higher-level QoS manager tool is described.

The paper by R. Friedrich et. al. [9] with the title "Integration of Performance Measurement and Modelling for Open Distributed Processing" discusses an architecture and prototype for an efficient measurement infrastructure for heterogeneous distributed environments. The benefits of integrating modelling and measurements for application design, deployment and management in ODP is highlighted. However, a stronger support of performance metrics in ODP is stated as needed.

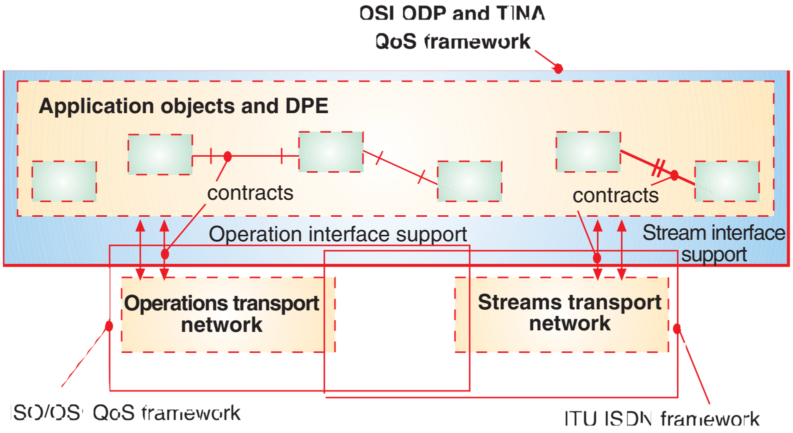


Figure 7.1 Distributed systems architecture-related QoS focus

The paper by G. Coulson et. al. [5] with the title “The Design of a QoS-Controlled ATM-Based Communication System in Chorus” describes the design of an application platform able to run distributed real-time and multimedia applications alongside conventional UNIX programs. It is focused on resource management aspects of the design and deal with CPU scheduling, network resource management and memory management issues. An architecture is presented that guarantees QoS levels of both communication and processing with varying degrees of commitment as specified by the QoS parameters.

The paper by J.Y. Hyi et al. [12] with the title “Quality-of-Service Control in GRAMS for ATM Local Area Network” focuses on Gopher-style real-time ATM multimedia services (GRAMS). A server that can determine the service rates and successfully multiplex media transmission according to QoS requirement is a key component of this system. Various *counters* are used to measure resource utilization and individual QoS. These counters are used for admission control and also user parameter control. Experimental results with video and image transfer using low-cost workstations are given.

Finally, the paper by K. Nahrstedt and J.M. Smith [20] with the title “The QoS Broker” is mentioned. In human affairs, *brokers are intermediaries, with specialized knowledge, who work toward a mutually desirable outcome through negotiation*. The QoS broker uses this principle to arrange for the delivery of an end-to-end QoS in distributed multimedia systems. Semantically, the QoS broker is not new. It is covered by the functionality which in Figure 2.3 is denoted “Negotiation and resource allocation”. This paper is interesting, however, not because of the broker concept, but because of a total design of the QoS broker in relation to underlying network and local operating system. An experimental prototype using a telerobotics application is also presented.

7 Summary and conclusions

There is no fundamental contradiction between the OSI, ISDN, ODP an TINA QoS frameworks. As a common feature there are: i) objects offering services and performing QoS handling functions, ii) QoS parameters/attributes, and iii) contracts between objects. However, the focus and also the use of concepts differ. The difference of focus is roughly as indicated in Figure 7.1 and Figure 7.2.

Considering the system architecture (Figure 7.1), ODP and TINA objects are abstractions of applications. OSI objects are abstractions of layer entities, while the ITU ISDN objects are abstractions of signalling capabilities within the user end-system or within the exchanges in the network. ODP and TINA are mostly

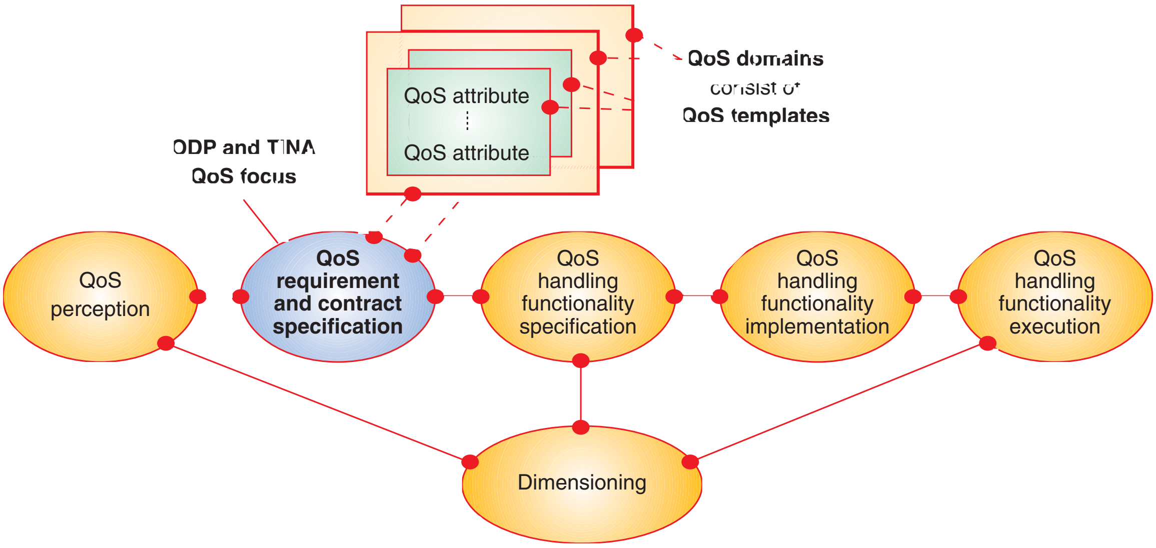


Figure 7.2 Distributed systems life-cycle-related QoS focus

concerned about the applications and also the needed DPE functionality for providing the ODP transparencies. OSI is concerned about packet switched networks while ITU is primarily concerned about circuit and cell-switched networks.

Considering the life-cycle view in Figure 7.2, ODP and TINA are mostly related to requirement and contract specification and accordingly languages for non-ambiguous specifications. By applying object-oriented specification of QoS attributes, QoS specifications can be extended and refined according to object-oriented techniques.

A QoS framework with an appropriate QoS handling functionality is the basis for meeting the design objectives of a telecommunication service providing system. Traffic and reliability researchers and engineers are mostly concerned about traffic modelling and performance. System researchers and engineers are mostly concerned about architectures and languages that are the basis for constructive design.

The QoS framework for distributed systems does not so far solve the problems related to QoS handling functionality as illustrated in Figure 2.3. QoS handling functionality must be included as part of the DPE functionality. A common QoS framework covering the *architecture* aspects as well as the *life-cycle* aspects, needs *a meta concept paradigm* with a common concept structure. In addition, adjustment on conflicting parts and effort to fill the holes in the framework are needed.

References

1 Aagesen, F A. Architectures for the modelling of QoS functionality *Telektronikk,* 91, (2/3), 1995, 56–68. ISSN 0085-7130.

2 Aagesen, F A, Hellan, J K. On constructive models for QoS functionality. *International Teletraffic Seminar, Bangkok,* November 1995.

3 The ATM Forum. *ATM User-Network Interface (UNI) Signalling Specification, version 4.0,* July, 1996.

4 Berndt, H et al. Service and management architecture in TINA-C. *TINA'95 Conference,* Melbourne 1995.

5 Coulson, G et al. The design of a QoS-controlled ATM-based communication system in Chorus. *Journal on selected areas in communications,* 13, (4), 1995.

6 Demounem, L. Intelligent networks and TINA service models : present and future. *INDC'96,* Trondheim, June 1996.

7 Fedaoui, L et al. Distributed multimedia systems quality of service in ODP framework of abstraction : a first study. *IFIP conference on open distributed processing, 1994.* North Holland, Elsevier, 1994.

8 Fevrier, A et al. Compositional specification of ODP binding objects. *INDC'96,* Trondheim, June 1996.

9 Friedrich, R et al. Integration of performance measurement and modeling for open distributed processing. *Proc. 3rd IFIP TC 6/WG 6.1 International Conference on Open Distributed Processing.* Chapman and Hall, 1994. (ISBN 0412-71150-8.)

10 Hafid, A, Bochmann, G V. An approach to quality of service management for distributed multimedia applications. *Proc. 3rd IFIP TC 6/WG 6.1 International Conference on Open Distributed Processing.* Chapman and Hall, 1994. (ISBN 0412-71150-8.)

11 Helvik, B E. Elements of a QoS measurement framework for bearer services. *Telektronikk,* 93, (1), 1997. (This issue.)

12 Hui, J Y et al. Quality-of-service control in Grams for ATM local area networks. *Journal on selected areas in communications,* 13, (4), 1995.

13 ISO. *Revised Text for ISO/IEC DIS 8073 - information technology - telecommunication and information exchange between systems - connection oriented transport protocol specification,* Geneva, 1991. (ISO/IEC JTC1/SC6 DIS 8072.)

14 ISO. *Information technology : transport service definition for open systems interconnection,* Geneva 1993. (ISO/IEC DIS 8072.)

15 ISO. Quality of service framework, Working Draft No. 2. *SC21/WG1 meeting, Yokohoma,* 1993. (ISO/IEC JTC1/SC21.)

16 ISO. *Reference model of open distributed processing.* Output from meeting in Helsinki, May 1995. (ISO/IEC JTC1/SC21 DIS 10746.)

17 ITU-T. *Recommendations, overall network aspects and functions.* Geneva, 1988. (CCITT Blue Book, Volume III - Fascicle III.8, Part III, I.300-Series.)

18 ITU-T. *E.800, terms and definitions related to the quality of telecommunication services, quality of service and dependability vocabulary,* Geneva, 1993. (ITU-T Blue Book, Fascicle II.3.)

19 Minerva, R. TINA service architecture : some issues in Service control. *TINA'95 Conference,* Melbourne, 1995.

20 Nahrstedt, K, Smith, M. The QoS broker. *IEEE Multimedia,* 1070-986X/95, 1995.

21 Février, A et al. QoS specification of ODP binding objects. *Telektronikk,* 93, (1), 1997. (This issue.)

22 Nilsson, G et al. An overview of the telecommunication information networking architecture. *TINA'95 Conference,* Melbourne 1995.

23 Rumbaug, J et al. *Object-oriented modelling and design.* Prentice Hall, 1991. (ISBN 0-13-630054-5.)

24 Schill, A et al. A quality of service abstraction tool for advanced distributed applications. *Proc. 3rd IFIP TC 6/WG 6.1 International Conference on Open Distributed Processing.* Chapman and Hall, 1994. (ISBN 0412-71150-8.)

25 TINA-C. *TINA Quality of Service Framework,* 1994.

26 Vogel, A et al. Distributed multimedia and QoS : a survey. *IEEE Multimedia,* 1070-986X/95, 1995.

QoS specification of ODP binding objects

BY ARNAUD FÉVRIER, ELIE NAJM, GUY LEDUC AND LUC LÉONARD

We present a QoS oriented notation suitable for the ODP framework. In particular, we focus on a computational view of objects: we consider systems described as configurations of interacting objects and we deal with two types of communications: message passing and flows. In message passing, signals (from one object to another) are conveyed through the (implicit) underlying infrastructure. This form of interaction is suitable for client/server applications where no strong real time or ordering constraints are needed from the communication infrastructure.

In contrast, in a flow type of communication, signals are conveyed through third party (binding) objects that may be explicitly called for in order to ensure specific QoS requirements needed by specific applications.

A building blocks approach for the formal specification of binding objects in the ODP computational model is presented. The formal notation that is used is based on LOTOS extended with two features – real time and gate passing. These features are among the extensions that are currently studied in the ISO standardisation Formal Description Techniques group. We apply our building blocks approach to the specification of a multicast, multimedia binding object.[1]

1 Introduction

The ODP reference model [1, 2] provides for a multiple-viewpoint specification of distributed applications and systems. Five viewpoints have been defined within ODP and are considered to encompass the different areas of concerns that need to be covered when one develops a system or application. These five viewpoints are: enterprise. information, computation, engineering and technology. For a given system or application, the enterprise viewpoint defines its requirements at a strategic level; the information viewpoint describes the information needed to represent it; the computational viewpoint provides an abstract implementation of it; the engineering viewpoint describes how the computational description is supported in terms of generic system components and communication protocols; and the technological viewpoint maps the generic engineering components onto existing pieces of hardware and software.

A new work item has now been launched within ISO and ITU aimed at developing a QoS architecture which is consistent with the ODP architecture. The first working draft is now available [3]. In particular, the document presents a preliminary identification of those aspects of the QoS that are appropriate to the different ODP viewpoints. For instance, it is stated that an Entreprise QoS description should include the requirements placed on a system by its users, and the corresponding guarantees and claims made by the system elements in meeting these requirements. This document is intended to be used as a basis for future discussion and development. Concerning the Information viewpoint, document [3] states that an Information QoS description should define the information needed to support the instantiation and dynamic behaviour of objects for the purpose of meeting their QoS requirements. As for the computational viewpoint, QoS descriptions should relate the requirements as stated in the Entreprise descriptions to the guarantees and claims made by the computational elements to meet the requirements. Moreover, the Computational description should exhibit the behaviour of objects that meet the requirements.

In the present paper, we will concentrate on the computational viewpoint which is of particular relevance to application programmers. The computational viewpoint is also interesting for systems designers as they are concerned with mapping computational descriptions onto generic execution components in the engineering model. In particular, we focus on object behaviour and exhibit, using a specification language, on a concrete example, a modular specification that complies with QoS requirements.

The Computational Model is the (abstract) language used to describe applications in the computational viewpoint: in the Computational Model, an application is represented as a dynamic configuration of interacting objects. Objects are the key concept in the Computational Model. A Computational object has a state that may be accessed externally only through interactions at its interfaces (we will hereafter, when there is no confusion, refer to computational objects simply as objects). An Object may possess (possibly many) interfaces which may be dynamically created and deleted. Interfaces are names of locations of interactions between objects. Objects may change dynamically their communicating partners by exchanging interface names. Objects may also create and delete other objects.

There are two kinds of objects in the Computational Model, namely, basic objects and binding objects. Binding objects are used to convey interactions between interfaces: (i) either explicitly through a binding object, or (ii) implicitly without exhibiting a binding object. When specifying an explicit binding object, the programmer may incorporate the QoS requirements (order, timeliness, throughput, ...) on the transport of the interactions supported by that binding object. In contrast, in an implicit binding between two interfaces, no specific requirements are made on the transport of interactions: interfaces interact by message passing with no explicit ordering or delay required on the transport of these messages.

There are three kinds of interfaces in computational objects: signal, operational and stream. Signal interfaces are the most primitive: operational and stream interfaces can be modeled as special types of signal interfaces. A signal is an operation name and a vector of values (references to interfaces). A signal interface is an interface that emits and receives signals. An operational interface is an interface that can receive invocations and possibly react with result messages. Invocations and result messages are signals. An operational interface has a type which is, roughly, defined to be the type of the operations it can handle (where the type of an operation includes the types of its return messages). A subtyping system allows for the safe substitution of an interface of a given type by another interface having a subtype of this type. A stream interface is an abstraction of a signal interface: the type of a stream interface is simply a name and a role (sender or receiver).

The development of ODP is a new challenge for formal techniques [2, 4]. Since July 94, the Formal Description Techniques (FDTs) group within ISO has become part of the Open Distributed Pro-

[1] *Parts of this paper was presented at IFIP/ICCC's INDC'96 in Trondheim, June 1996.*

cessing (ODP) standardization committee (SC21/WG7). Thus, supporting the formal design of open distributed systems is a new objective for this group. Indeed, the ISO-FDT team of experts is now working on the standardization of an extension of LOTOS – temporarily called E-LOTOS [5], which is targeted, among other things, at providing support to the design of ODP systems. This group has established a list of desirable features together with a list of requirements that E-LOTOS should aim to fulfil. Aspects related to real-time, constructive data representations and modularity are being actively studied. Dynamic reconfiguration of communication structures is also being tackled.

The present paper is an exercise in the specification of a binding object using a formal description technique. Binding objects are important for both application and system designers and developers and they can be used in many different ways. For instance, application designers may specify their transport requirements and let system designers develop new networks and protocols that match these requirements. On the other hand, application programmers may use the abstraction provided by existing binding objects to develop and analyze their applications. A specification of a binding object should cover the functional and QoS requirements. Functional requirements include: connection establishment, dynamic reconfiguration, orderly transport of information, etc. QoS requirements involve: connection establishment delay, jitter, throughput, error rate, inter and intra flow synchronization, etc. Thus, the specification language should be expressive and able to address real-time constraints and to capture dynamic reconfiguration of communicating components.

We use for our specification exercise the LOTOS language [6, 7] extended with two features that are currently under study in the ISO/FDT group: real time and gate (i.e. reference) passing à la p-calculus [8]. We call this language MT-LOTOS. We show how the MT-LOTOS indeed allows for a modular construction of our binding object. We introduce first a collection of generic building blocks specified in MT-LOTOS. Each of these blocks has a self contained meaning and can be composed with other building blocks. The genericity and reusability of these blocks is illustrated in the construction of a multimedia, multicast binding object.

The remainder of the document is structured as follows. Section two is a short presentation of MT-LOTOS. In section three we introduce the collection of building blocks specified in MT-LOTOS. Section four is devoted to the specification of the binding object example, which is first presented informally. In section five we conclude.

2 A brief presentation of MT-LOTOS

As said earlier, MT-LOTOS is a combination of two extensions to LOTOS. These extensions are formally specified and discussed in [9, 10] and [11, 12]. In this paper we give a short informal presentation, leaving aside the data typing aspects. In order to make our presentation clear and self contained, we present MT-LOTOS as a language on its own, without discussing its differences with LOTOS. It is however important to note that MT-LOTOS is an upward compatible extension of LOTOS.

The primitive concepts of MT-LOTOS are: actions, processes and agents. These can be composed as follows: (i) actions can be composed to form processes, (ii) processes can be composed to form processes and/or agents; (iii) agents can be composed to form agents. We briefly introduce each of these concepts.

2.1 Actions

Actions can be internal, represented by the symbol i, or external. An external action is a gate and an ordered list of offers: g off1...offn; where an offer is one of four possible forms: (i) presenting a value: !E (resulting from the evaluation of expression E), (ii) accepting a value of some type ?x:t (and storing it in variable x), (iii) presenting a gate name: !g, (iv) accepting a gate name: ?h:gid (which is stored in h. Note the special type for gates, *gid*). A typical feature of MT-LOTOS is that actions may be conditioned by a predicate and/or a time constraint, and may have a side effect of creating new agents. The syntax of actions is captured by the following definitions (using a BNF grammar rule).

```
a ::= i{time-const} new C |
g off1...offn {time-const}[pred] new C
off ::= !E | ?x:t | ?h:gid
time-const ::= t in t1..t2 | t | t1..t2
```

The general form of time-const is t in t1..t2 where t is a declared variable that records the time elapsed between action offering and action occurrence. Two other forms are allowed, respectively when there is just a recording variable t without actual constraint, or a constraint without time recording. Pred is a Boolean expression, possibly referring to variables used in the offers offi of the action or declared in time-const. Finally, C is an agent that is created as a side-effect of the execution of the action (we will see how agents are constructed in the sequel). An unconstrained action a, is equivalent to: a {0...inf} [true]. Note that no predicate Pred can be associated with an internal action i.

2.2 Processes

Processes are obtained by composition of actions and/or other processes. Let us first introduce three simple process constructs, then we turn to more sophisticated ones: (i) stop is the simplest process, representing the do-nothing-while-letting-time-pass behaviour, (ii) if B is a process then: c; B is the process representing the behaviour: "perform action c and then enable (the actions) of process B", (iii) wait t; B represents the behaviour "let t time units pass and then enable (the actions) of B". Assuming B, B1 and B2 represent generic processes, the general form of MT-LOTOS processes is given by the following definitions.

```
B :: =  stop
     |  c ; B
     |  wait t; B
     |  B1[]B2
     |  B1 |[g1, ..., gn]| B2
     |  exit
     |  B1 >> B2
     |  B1 [> B2
     |  hide g1, ..., gn in B
     |  P [h1, ..., hk](E1, ..., Em)
```

A few words on each of the newly introduced constructs.

B1 [] B2 is a disjunctive choice between the actions of B1 and the actions of B2. The choice is resolved by the execution of the first action. For instance, the behaviour of (c1; B1) [] (c2; B2) is: "perform c1 then enable B1 (thus disabling c2; B2) or perform c2 then enable B2 (thus disabling c1; B1).

B1 [> B2 is the disabling of the behaviour of B1 by the first action of B2. For instance, the behaviour of (c1; B1) [>(c2; B2) is: "perform c1 then enable B1 [>(c2;B2) or perform c2 then enable B2 (thus disabling c1; B1)".

exit is the process which performs one special action, the successful termination action, and then stops. This termination action is used to enable a new process as explained in the following construct.

B1 >> B2 is the enabling of the behaviour of B2 by the last action of B1. For instance, the behaviour of exit >> B2 is: "perform a (hidden) termination action and then enable B2"; and the behaviour of (c; B1) >> B2 is: "perform action c then enable B1 >> B2".

B1 | [g1, ..., gn] | B2 is: "run B1 and B2 in parallel enforcing the synchronisation on actions occurring at gates g1, ..., gn while letting the other actions free". In order for two actions to be synchronisable, they must have the same gate, two matchable lists of offers (i.e. where the ith offer of one list matches the ith offer of the other list), and the predicates, if any, must evaluate to true. Offers are matchable as follows: (i) two values can be matched if they are of the same type and they are equal, (ii) a value and a variable can be matched if they have the same type; the matching, in this case, results in a transfer of the value to the variable, (iii) two variables of the same type are always matchable. Note that if B1 and B2 synchronise on two actions, then they are said to perform jointly a synchronised action, and this action can be further synchronised with yet a third process, like e.g. in the expression: (B1 | [g] | B2) | [g] | B3. Note that i is the action that can never synchronise with any action.

hide g1, ..., gn in B hides (transforms into the internal action, i) the actions occurring on gates g1, ..., gn. Thus, in an expression (hide g1, ..., gn in B) | [g1] | B', the actions of B occurring on gate g1 are internal and thus cannot synchronize with actions from B'. hide has also another function: it creates new gate names. For instance in the expression, hide g in h! g; B, a new gate, g, is created and sent on gate h.

The specifier may define named processes by a set of equations of the form: Process P[g1, ..., gk] (x1:t1, ..., xn:tn):= B Endproc where P is the name of the process, g1, ..., gk is a list of gate name parameters and x1:t1, ..., xn:tn a list of value parameters (typed variables). Hence, the behaviour of P[h1, ..., hk] (E1, ..., Em) is defined to be the same as B where each gi has been substituted with hi and each xi has been substituted with Ei.

2.3 Agents

The communicating architecture of processes is static and imposed by the parallel operators. Agents have dynamic communicating structures: agents may discover new agents and interact with them, agents may also forget about previously known agents. Agents are made from processes using the embedding operator <_>: if B is a process, then <B> is an agent. <B> is the simplest form of an agent. The function of the embedding operator, <_>, is to put a boundary around a process, thus allowing it to interact with other agents. Agents can be put in parallel with other agents using the operator |. For instance <B1> | <B2> is the parallel composition of agents <B1> and <B2>. In contrast with processes, interaction between agents is binary and the synchronisation gates are not given explicitly in the parallel operator. In <B1> | <B2>, <B1> and <B2> can perform actions freely (without synchronisation) and can also synchronise on matching actions. In this case, the resulting joint action is hidden. The behaviour of an agent can be restricted by disallowing actions occurring at a specified list of gates. For instance, if C is an agent, then restrict g1, ..., gn to C is an agent which has a behaviour similar to C except that it does not perform any action occurring on gates g1, ..., gn.

One last remark concerning the creation of agents. We have seen that new agents can be spawn as a side effect of some actions. The following is an example of this feature. Take the agent <g!h new <B1>; B2>. This agent performs action g!h (offering gate h on gate g) which enables then the agent: <B1> | <B2>, i.e. the parallel composition of <B2> with the spawn agent <B1>.

Finally, one can define named agents in a way similar to that of named processes. The syntax of agents is given by the grammar rules:

```
C :: =  <B>
     |  restrict g1, ..., gn to C
     |  C1 | C2
     |  A [h1, ..., hk](E1, ..., Em)
```

3 A collection of building blocks

One of the most important specification styles of (MT-)LOTOS is the constraint oriented style. Thanks to this style, one can obtain specifications composed from generic modules where each module represents a constraint that acts upon a designated part of the system. The constraints can be of different forms, such as the order of actions on a given gate, the timeliness of actions, the structure of the data conveyed in the actions, etc.

We have identified a collection of generic components that are suitable for the specification of functional and QoS requirements of multimedia and multicast binding object. We present them below, together with their MT-LOTOS specification. In these specifications, dt represents a packet of a certain media (audio or video).

Medium: This component describes a point-to-point transmission medium between two points. Packets are received on gate ist (input stream), and are delivered on gate ost (output stream). Medium is very general. The only constraint it expresses is that no packet is lost. On the other hand, the transmission delay of each packet is totally unconstrained and the ordering of the packets is not preserved. After the reception of a packet on ist, i{0..inf} introduces a nondeterministic delay before the delivery on ost. In parallel a new occurrence of Medium handles the following packets.

```
PROCESS Medium [ist, ost]: NOEXIT :=
   ist ? dt:data; ( i{0..inf}; ost ! dt; STOP
   |||
   Medium [ist, ost])
ENDPROC (* Medium *)
```

FIFO_Const: This component also considers gates ist where packets are received, and ost where these packets are delivered. It enforces that the packets be delivered in the same order as they are received. In process FIFO_Const, the ordering is handled with an appropriate data structure: q, that describes a FIFO queue. We will not enter here into the details of the datatypes definition. At any time, FIFO_Const can accept (ist ? dt:data) a new packet that is added to q, or deliver (ost !first(q)) the first packet in q.

```
PROCESS FIFO_Const
[ist, ost](q:fifo):NOEXIT :=
      ist ?dt:data; FIFO_Const
      [ist,ost](append(dt,q))
   [] [not(IsEmpty(q))] -> ost !first(q);
   FIFO_Const [ist,ost](rest(q))
ENDPROC (* FIFO_Const *)
```

Delay_Const: Here again, two gates are considered. Delay_Const enforces that at least a minimal delay delmin elapses between the receiving of a packet on ist and its delivery on ost.

```
PROCESS Delay_Const [ist, ost]
(delmin):NOEXIT :=
   ist ? dt:data ; (Wait delmin ; ost!dt; STOP
   |||
   Delay_Const [ist, ost](delmin))
ENDPROC (* Delay_Const *)
```

Delay_Obs: Delay_Obs expresses a requirement on the service provided by a transmission medium. It verifies that the delay between the reception and the delivery never exceeds a maximal value. If the packet is not delivered before this maximal delay, an error message is sent on the management gate m. After the reception of a packet, Delay_Const proposes ost!dt during a time delmax. On the other hand, m!error_delay!ost is delayed by delmax+epsilon. In other words, m!error_delay!ost is enabled when the delivery cannot occur anymore.

```
PROCESS Delay_Obs[ist, ost, m]
(delmax):NOEXIT :=
   ist ? dt:data ; ( ( ost ! dt {0..delmax};
   STOP
      []
      wait(delmax+epsilon);
      m!error_delay!ost ; STOP ) |||
      Delay_Const [ist, ost, m](delmax))
ENDPROC (* Delay_Obs *)
```

Jitter_Const: Jitter_Const has an effect similar to Delay_Const, but on just one gate. It enforces that at least a minimal delay jmin elapses between any two successive deliveries of packets at gate ost.

```
PROCESS Jitter_Const [ost](jmin):
NOEXIT :=
   ost ? dt:data; Jitter_Const2 [ost](jmin)
where
   PROCESS Jitter_Const2 [ost](jmin):
   NOEXIT :=
      wait jmin ; ost ? dt:data;
      Jitter_Const2 [ost](jmin)
   ENDPROC (* Jitter_Const2 *)
ENDPROC (* Jitter_Const *)
```

Jitter_Obs: Jitter_Obs has an effect similar to Delay_Obs, but on just one gate. It verifies that the delays between successive deliveries of packets on gate ost do not exceed jmax. Like Delay_Const, it signals an error if this happens.

```
PROCESS Jitter_Obs [ost, m](jmax):
NOEXIT :=
   ost ? dt:data;
   Jitter_Obs2 [ost, m](jmax)
where
   PROCESS Jitter_Obs2 [ost, m](jmax):
   NOEXIT :=
      ost ? dt:data {0..jmax}; Jitter_Obs
      [ost, m](jmax)
   [] wait (jmax+epsilon);
   m!error_jitter!ost ; STOP
   ENDPROC (* Jitter_Obs2 *)
ENDPROC (* Jitter_Obs *)
```

One_Ind_Flow: One_Ind_Flow gives a first example of the modularity allowed by MT-LOTOS. It describes a flow that combines the effects of the previous components. So, this flow loses no packet and preserves their order; the transmission delay of each packet is undetermined, but it is at least of delmin, and it cannot exceed delmax, otherwise an error message is sent and the transmission is stopped; the delay between successive deliveries of packets (the jitter) is at least of jmin and at most of jmax, if this maximal value is exceeded, an error message is also sent and the transmission is stopped.

One_Ind_Flow is simply obtained by putting in parallel the various constraints (or processes) and by enforcing their synchronisation on the gates ist and ost. In this case, One_Ind_Flow integrates all the constraints, but any other combination of them would have been possible too (for example with no lower bound on the transmission delay or with no preservation of the order) resulting in a less constraining flow. Furthermore, One_Ind_Flow allows the handling of a disconnection through the management gate m: the occurrence of m !Dreq!ost interrupts the flow and the whole process turns into stop.

```
PROCESS One_Ind_Flow [ist, ost, m]
   (q:fifo, delmin, delmax, jmin, jmax):
   NOEXIT :=
      ( ( Medium [ist, ost]
      | [ist, ost] |
      FIFO_Const
      [ist, ost](q)
      | [ist, ost] |
      Delay_Const [ist, ost] (delmin)
      | [ist, ost] |
      Delay_Obs [ist, ost, m] (delmax))
   | [ost] | (Jitter_Const [ost](jmin)
   | [ost] | Jitter_Obs [ost, m](jmax))
   ) [> m !Dreq!ost ; STOP
ENDPROC (* One_Ind_Flow *)
```

Inter_Sync_Const: Until now, we have only presented constraints handling one flow. Inter_Sync_Const controls the synchronisation between the packets delivered by two flows. The complete synchronisation mechanism requires two brother instances of process Inter_Sync_Const: one per flow. Inter_Sync_Const controls the packets delivered on ost by its local flow and exchanges on gate s synchronisation information with the Inter_Sync_Const responsible for its brother flow. The way Inter_Sync_Const is combined with a flow is illustrated by the next component: One_Sync_Flow. The effect of Inter_Sync_Const is to ensure that the packets on its local flow are not delivered too late or too early with respect to the packets on the brother flow. The local flow (resp. the brother flow) may be ahead of the brother flow (resp. the local flow) of at most my (resp. ym) time units. If these constraints cannot be met, an error message is sent on gate m and the flow is interrupted. We will not enter here into more details about the synchronisation mechanism. The actual values for these parameters are given in the following section. The meaning of the parameters used in this process is the following:

- ml is the ideal time for my last packet
- yl is the ideal time for your last packet
- me is the time elapsed since my last packet
- ye is the time elapsed since your last packet
- my is the accepted advance of my stream over your stream
- ym is the accepted advance of your stream over my stream
- mm is the interval between two successive packets of my stream

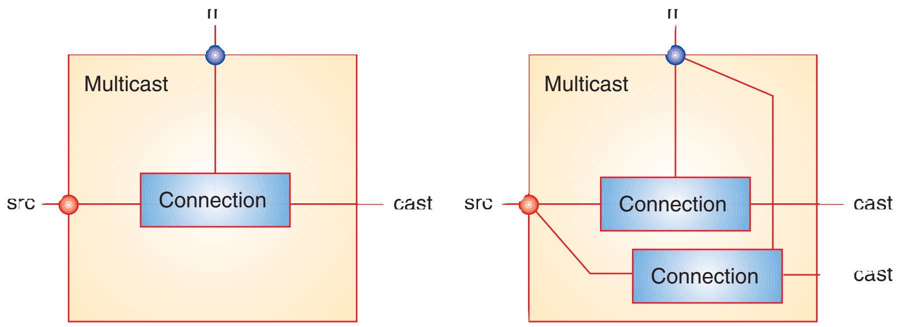


Figure 1 The multicast component, with one connection (left) and two connections (right)

- yy is the interval between two successive packets of your stream
- ms is the name of my stream : takes one of two values a or v
- ys is the name of your stream : takes one of two values v or a.

```
PROCESS Inter_Sync_Const [ost, s, m]
  (ml, yl, me, ye, my, ym, mm,
  yy:time, ms, ys:streams):NOEXIT :=
  ost ? dt:data {t}
  [ my > ml+mm-(yl+ye+t) > 0-ym ] ;
    Notify_Other_Stream[ost, s, m]
      (ml+mm, yl, 0, ye+t, my, ym,
      mm, yy, ms, ys)
  []
  s!ys; Inter_Sync_Const [ost, s, m]
  (ml, yl+yy, me+t, 0, my, ym, mm,
  yy, ms, ys)
  []
  wait (ml+mm - (yl+ye)+ym+epsilon);
  m!inter_sync_error!ost ;
  STOP
WHERE
  PROCESS Notify_Other_Stream [ost, s, m]
    (ml, yl, me, ye, my, ym, mm,
    yy:time, ms, ys:streams):
    NOEXIT :=
    s!ms {0..0}; Inter_Sync_Const[ost, s, m]
      (ml, yl, me, ye, my, ym, mm, yy,
      ms, ys)
    []
    s!ys; Notify_Other_Stream [ost, s, m]
      (ml, yl+yy, me+t, 0, my, ym, mm,
      yy, ms, ys)
  ENDPROC (* Notify_Other_Stream *)
ENDPROC (* Inter_Sync_Const *)
```

One_Sync_Flow: This component gives a new example of the modularity allowed by MT-LOTOS. One_Ind_Flow is already the composition of several features. One_Sync_Flow enhances it with Inter_Sync_Const, a synchronisation mechanism with another flow. Again, the addition of a constraint is simply obtained by putting both processes in parallel. Furthermore, this last interflow constraint may be removed if a request is made on gate m. The constraint is then replaced by the neutral process Sink.

```
PROCESS One_Sync_Flow [ist, ost, m, s]
  (q:fifo,delmin, delmax, jmin,
  jmax:time,
  ml, yl, me, ye, my, ym, mm, yy:time,
  ms, ys:streams); NOEXIT :=
  ( One_Ind_Flow [ist, ost, m]
  (q, delmin, delmax,jmin,jmax)
  | [ost] |
  ( Inter_Sync_Const [ost, s, m]
  (ml, yl, me, ye, my, ym, mm, yy, ms, ys )
    [> m!dis_other_stream!ost;
    Sink[ost])
  )[> m!Dreq!ost ; STOP
ENDPROC (* One_Sync_Flow *)
```

Sink: This component enforces no constraint on the actions occurring on a gate. It is specified by a process with a single gate: st. In process sink no predicate or time constraint restricts the acceptance of packets on st.

```
PROCESS Sink [st] : NOEXIT :=
st ? dt:data ; Sink [st]
ENDPROC (* Sink *)
```

Multicast: This component is independent from the previous one. It describes a multicasting mechanism. Considering one "input" or source gate (src) and a collection of "output" gates (described by the generic gate variable cast), Multicast conveys messages, without delay, between the "input" gate and every "output" gates. The command to create a new output gate cast and an associated connection from src to the new cast is received on gate m. Remark that Multicast receives the new gate name in variable cast of type gid (which is the special type used for gate names). Figure 1 is a graphical representation of the multicast process, in two situations: one active connection (left) and two active connections (right). Note that processes are represented by rectangular boxes whereas agents will be represented by rounded boxes.

```
PROCESS Multicast [src, m]: NOEXIT :=
  src? dt: data ; Multicast [src, m]
  []
  m ! Creq ? cast : gid;
    (  ( Connection [src, cast] [>
      m !Dreq ! cast ; Sink [src] )
      | [src] |
      Multicast [src, m]
    )
WHERE
  PROCESS Connection [src, cast]:
  NOEXIT :=
    src ? dt : data ; cast !dt {0};
    Connection [src, cast]
  ENDPROC (* Connection *)
ENDPROC (* Multicast *)
```

4 A multicast multimedia binding object

We want to specify an ODP Binding Object that supports a video broadcast application. The binding object we aim at fulfils the following functions:

- It listens to a source emitting two synchronised flows, an audio and a video, and multicasts the two flows to a dynamically changing set of clients
- At any time a client can request to join the audio or the video or both the audio and video streams by providing the reference of one (or two) receiving interface(s)
- At any time a client may request to leave the audio, or the video or both audio and video flows
- It tries to enforce the intra and inter synchronisation of flows and notifies failures to do so.

The source flows have these characteristics, inspired from [13]:

- There are 25 images per second, i.e. the video stream is constituted by packets delivered every 40 ms
- The sound is sampled every 30 ms, i.e. a sound packet is delivered every 30 ms.

We suppose that the two source flows do not deviate from the above figures and that both flows are fully synchronized. The Binding object accepts these flows and delivers them to any requiring customer. Since the binding object will encapsulate the behaviour of a concrete network, it will have to deal with usual networks problems (jitter, packet loss, end-to-end delay, ...). Nevertheless, the customers expect a minimal QoS. The QoS is twofold:

- Each flow must respect a QoS,
  - the sound may suffer no jitter
  - the video allows a jitter of 5 ms, i.e. consecutive images may be separated by 35 to 45 ms
- Both must be *reasonably* synchronous. This is known as *lip synchronization*.

The lip synchronization is considered correct if the sound is not too far (back or ahead) from the corresponding lip movement. The actual figures are:

- The sounds must not come more than 15 ms before the lip movement
- The sounds must not come later than 150 ms after the lip movement.

We give the MT-LOTOS specification below. The above figures will be used in the specification as follows:

- abv = 15 ms (allowed advance of the audio stream on the video stream)
- vba = 150 ms (allowed advance of the video stream on the audio stream)
- ar = 30 ms (audio packets rate)
- vr = 40 ms (video packets rate).

Figure 2 represents the initial configuration of agents, i.e. when no clients have joined the casts. Note that MT-LOTOS agents are represented by rounded rectangles. Figure 3, at the end of the specification, represents one client connected to the (synchronised) audio and video flows.

```
SPECIFICATION binding-object
[srca, srcv, c]
(* gate c is the controlling gate of the
binding object *)
(* gates srca and srcv are the audio and
video source gates respectively *)
TYPE SORTS streams
OPNS a, v -> streams
ENDTYPE
BEHAVIOUR
RESTRICT mma, mmv TO
```

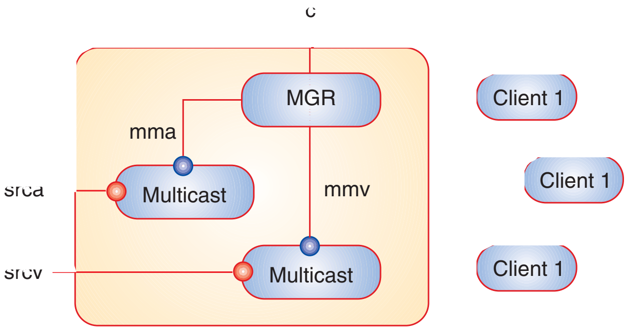


Figure 2 The binding object with no connected clients

```
< MGR[c, mma, mmv] > |
< Multicast[srca, mma] > |
< Multicast[srcv, mmv] >
```

At the initial state, the binding object configuration contains two Multicast objects, one for audio and one for video and a manager object, MGR. At this initial state, the Multicast objects only listen, each to its specific media source, and no destinations are active, i.e. no client is being serviced.

The manager of the binding object, MGR, is accessed by the clients at gate c, and manages the audio and video Multicast objects through gates mma and mmv respectively.

```
WHERE
PROCESS MGR [c, mma, mmv]:
NOEXIT :=
```

MGR is a choice between three actions corresponding to three types of requests from clients: a request to join the audio Multicast, a request to join the video Multicast, or a request to join both the audio and video Multicast.

```
HIDE isa, m, mgt_client IN
   c !Creq-a ?r_client:gid ?osa:gid
      ?delmin:time ?delmax:time
      ?jmin:time ?jmax:time5
   NEW
   ( < One_Ind_Flow [isa, osa, m]
   (empty, delmin, delmax, jmin, jmax) >
   | < Client_One_Flow_MGR
   [mgt_client, mma, m] (isa, osa, r_client) >
   | < r_client !mgt_client ; STOP > );
```

This is a request to join the audio Multicast. The request contains a return gate, r_client, of the requesting client and the gate that has to be bound to the audio stream, osa.

The MGR agent creates the gates isa, m and mgt_client and two objects, One_Ind_Flow and

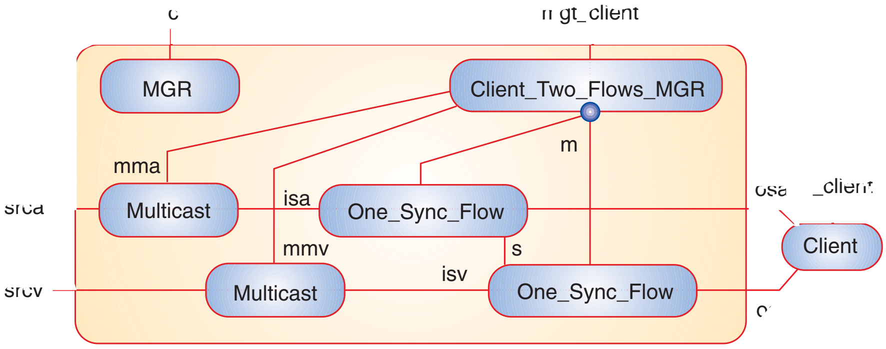


Figure 3 The binding object with one client connected to both the audio and video casts

Client_One_Flow_ MGR, and a return message <r_client! mgt_client;STOP>. Gate isa connects the Multicast object to One_Ind_Flow which conveys the audio stream to the client gate osa. Client_One_Flow_MGR manages through the access gate mgt_client the requests from the client concerning this stream. Client_One_Flow_MGR operates on One_Ind_Flow through gate m. Message <r_client!mgt_client; STOP> notifies the client of the success of the binding and provides him with the interface name mgt_client that has been created for him to manage his connection.

```
mma !Creq !isa ; MGR [c, mma, mmv]
(* MGR conveys the request, on behalf
of the client, to the audio Multicast
object, and provides the name of the
input gate isa *)

[]Hide isv, m, mgt_client IN
   c !Creq-v ?r_client:gid ?osv:gid
      ?delmin:time ?delmax:time
      ?jmin:time ?jmax:time
   NEW
      ( < One_Ind_Flow [isv, osv, m]
         (empty, delmin, delmax, jmin,
         jmax) >
      | < Client_One_Flow_MGR
      [mgt_client, mmv, m]
         (isv, osv, r_client)>
      | < r_client !mgt_client ; STOP> );
      mmv !Creq !isv ;
      MGR [c, mma, mmv]
```

This is the symmetric request for a video connection

```
[] Hide isa, isv, m, mgt_client IN
   c !Creq-av ?r_client:gid ?osa:gid ?osv:gid
   ?delmin:time ?delmax:time
   ?jmin:time ?jmax:time
   ?abv:time ?vba:time ?ar:time ?vr:time
NEW
   (RESTRICT s TO
      ( < One_Sync_Flow [isa, osa, m, s]
         (empty, delmin, delmax, jmin, jmax,
         0, 0, 0, 0, abv, vba, ar, vr, a, v) >
      | < One_Sync_Flow [isv, osv, m, s]
         (empty, delmin, delmax, jmin, jmax,
         0, 0, 0, 0, vba, abv, vr, ar, v, a) >)
      | < Client_Two_Flows_MGR
         [mgt_client, mma, mmv, m]
         (isa, isv, osa, osv, r_client) >
      | < r_client !mgt_client ;STOP >
   ); (mma !Creq !isa ; EXIT
   ||| mmv !Creq !isv ; EXIT )
      >> MGR [c, mma, mmv]
```

This third sub-expression of the choice represents the handling of a combined audio/video connection request. The client provides two gates to be bound, osa and osv, one for the audio and one for the video stream respectively. In this case, the objects that convey the streams are instantiated from the One_Sync_ Flow template: the two streams have to be synchronized. In this instantiation, we have used the following constants:

abv: Maximum advance of the audio stream on the video stream

vba: Maximum advance of the video stream on the audio stream

ar: Interval between two audio data packets

vr: Interval between two video data packets.

```
WHERE
PROCESS Client_One_Flow_MGR
[mgt_client, mm, m]
(ist, ost, r_client:gid) :noexit :=
(mgt_client ! Dreq ; (mm !Dreq !ist ; STOP
            ||| m !Dreq !ost ; STOP)
   [] m ? er:error!ost ; mm !Dreq !ist NEW
      <r_client!er!ost; STOP > ; STOP)
ENDPROC (* Client_One_Flow_ MGR *)
PROCESS Client_Two_Flows_MGR
[mgt_client, mma, mmv, m]
   (isa,isv,osa,osv,r_client:gid) :noexit :=
   mgt_client ! Dreq !a ;
   m!dis_other_stream!osv ;
      ( mma !Dreq!isa; EXIT
      ||| m!Dreq !osa ; EXIT)
      >> Client_One_Flow_MGR
      [mgt_client, mmv, m] (isv, osv, r_client)
   [] mgt_client ! Dreq !v ;
   m!dis_other_stream!osa ;
   (mmv !Dreq!isv; EXIT
   ||| m!Dreq !osv ; EXIT)
   >> Client_One_Flow_MGR
      [mgt_client, mma, m] (isa, osa, r_client)
   [] m ? er:error!osa ;
      (mma!Dreq !isa; EXIT
      ||| mmv!Dreq !isv; EXIT
      ||| m!Dreq!osv;EXIT)
   >>
   i NEW <r_client!er!osa ; STOP > ; STOP
   [] m ? er:error!osv ;
      (mma!Dreq !isa; EXIT
      ||| mmv!Dreq !isv; EXIT
      ||| m!Dreq!osa; EXIT)
   >>
   i NEW <r_client!er!osv ;  STOP > ;
   STOP
ENDPROC (* Client_Two_Flows_ MGR *)
ENDPROC (* MGR *)
ENDSPEC
```

5 Conclusion

Mobile-Timed-LOTOS extends the classical formal description technique LOTOS with two new features: *quantitative time* and *mobility.* These two enhancements make MT-LOTOS very well adapted to the specification of complex dynamic systems like an ODP binding object.

Quantitative time allows the precise description of real-time aspects. With the recent evolutions in networking, e.g. multimedia, such aspects become more and more crucial and being able to specify them is mandatory.

Mobility allows the writing of specifications whose structure can be dynamically reconfigured. New processes can be created as well as new connections between them, through new gates. This is typically the kind of flexibility required by a binding object which conveys interactions between interfaces that can change in time. Our binding object for example allows new customers to be connected on demand.

We have presented here a building blocks approach. Such an approach is made possible by the modularity that MT-LOTOS permits. We have defined a series of generic components that can be put together very easily to form various combinations. Note that LOTOS already offers a constraint oriented style. However, the structure of LOTOS specifications is static. The dynamic evolution of the binding object cannot be described as naturally as with MT-LOTOS. In particular, the inability to create new gates forces to resort to specification "tricks"; all the customers are connected to the same gate but each one is differentiated by a special attribute added to the gate. This results in less generic building blocks that cannot be reused easily in other contexts.

In the specification approach we have followed, we aimed at exhibiting the behaviour of objects that fulfill the required QoS constraints that are placed on the system. A complementary notation is needed which should formalize these requirements. Lamport's TLA [14] is one such notation which is based on temporal logic. Another notation, a Calculus for Object Contracts (COC), is proposed in [15]. This calculus is tailored to the ODP framework and deals with the deterministic real time QoS constraints. In particular, it formalises the important concepts of object contracts which is highly relevant not only for the advantages it brings in software development methodologies (abstraction and composition), but also, in an open multi vendor/ stakeholder environment, for its clarification of the relationships between the different actors involved. COC is based on an observer oriented approach: a contract is a (dynamic) configuration of observers which listen and analyze the events on the contractual interfaces and in case of error detection, emit error messages identifying the error type and the offending objects. Investigations are now made in order to define a framework whereby LOTOS based dialects are used to define object behaviours and COC and the QoS contracts that should be realised by these objects.

References

1 ISO/ODP. *ODP reference model, parts 1, 2, 3, 4.* (ISO/IEC 10746-1/2/3/4 or ITU-T X.901/2/3/4.)

2 Stefani, J-B. Open distributed processing : the next target for the application of formal description techniques. In: *3rd International Conference on Formal Description Techniques FORTE 90,* Madrid, 1990.

3 *Open Systems Interconnection, data magement and Open Distributed Processing. Working doucument on QoS in ODP.* (ISO/IEC JTC1/SC21/ WG7 N1192.)

4 Vissers, C A. FDTs for open distributed systems : a prospective view. In: *Proceedings 10th IFIP WG6.1 Workshop on Protocol Specification, Testing and Verification, Ottawa, Canada,* 1990.

5 ISO. *Revised working draft on enhancements to LOTOS.* May 1995. (ISO/IEC JTC1/SC21/WG7 N1001.)

6 ISO. *LOTOS : a formal description technique based on the temporal ordering of observational behaviour.* ISO, 1988. (International Standard 8807.)

7 Bolognesi, T, Brinksma, E. Introduction to the ISO specification language LOTOS. *Computer Networks and ISDN Systems,* 14, 1987, 25–29.

8 Milner, R, Parrow, J, Walker, D. A calculus of mobile processes, parts I & II. *Journal of Information and Computation,* 100, 1992, 1–77.

9 Léonard, L, Leduc, G. An enhanced version of timed LOTOS and its application to a case study. In: *Formal description techniques, VI.* Tenney, R, Amer, P, Uyar, Ü (eds.). Amsterdam, North Holland, 1994, 483–498.

10 Léonard, L et al. Belgian-Spanish proposal for a time extended LOTOS. In: *Revised draft on enhancements to LOTOS.* Quemada, J (ed.). 1995. (ISO/IEC JTC1/SC21/ WG7 N1001.)

11 Najm, E, Stefani, J-B, Février, A. Towards a mobile LOTOS. In: *Formal description techniques, VIII.* Bochman, G, Dsouli, R, Rafiq, O (eds.). Montréal, Canada, 1995.

12 Najm, E, Stefani, J-B, Février, A. Introducing mobility in LOTOS. In: *Revised draft on enhancements to LOTOS.* Quemada, J (ed.). 1994. (ISO/IEC JTC1/SC21/WG1 N1349.)

13 Stefani, J-B, Hazard, L, Horn, F. Computational model for multimedia applications based on a synchronous programming language. *Computer Communications,* 15, (2), 1992, 114–128.

14 Lamport, L. The temporal logic of actions. *Toplas,* 16, (3), 1994, 872–923.

15 Fevrier, A, Najm, E, Stefani, J-B. Contracts for ODP. *Fourth AMAST Workshop on Real-Time Systems, Concurrent, and Distributed Software. Towards a mathematical transformation-based development.* 21–23 May 1997, Ciudad de Mallorca, Mallorca. LNCS. Springer Verlag.

A user's perspective on Quality of Service activities in world telecommunications

BY ALBERT LEE[1]

Introduction

Dr W. Edwards Demming, the US born Quality "guru" who introduced the concepts of Quality Management to Japan shortly after the Second World War, is quoted as saying that "Survival is Not Compulsory". Just as Japanese Industry and later Western Industry, in response to the onslaught from Japan in the 1970s, embraced the concepts of Quality Management, in time the principles and all important attitudinal aspects were introduced to telecommunications and, in particular, to the organisations providing telecommunications service. Many monopoly telecommunications suppliers carrying, the baggage of old infrastructure and faced with a new operator deploying the very latest low cost technology and "cherry picking" the more profitable markets, have survived during the transition to freer markets by focusing on improving quality. This focus on quality has often enabled the "old" monopolies to compete whilst offering services which are priced higher than those from the new competitors. Increasingly as network operators move into the services provider role they have to take a wider view of quality than the traditional network operator role of network performance. They must view quality through the eyes of the customer, considering all aspects of providing service on an end-to-end basis.

More focus on overall Quality of Service is needed and less on network performance parameters

Prior to the 1980s the majority of telecommunications services were provided by government departments or state controlled monopolies. Many of these enjoyed the term in their title of "telecommunications service" but provided poor service to what most steadfastly called subscribers, often finding the concept of a customer alien. The inferior service was often founded in a combination of low capital investment and inferior understanding and, in some cases, a complete lack of understanding of the needs of subscribers or customers.

The late 1980s and early 1990s have seen a major change in the approach of telecommunications network providers (NP) and service providers (SP) towards quality. Traditionally, NPs have always committed considerable resource to monitoring switch and transmission performance to ensure their network capacity exceeded demand. In the past this has been very important in managing POTS, and although this is much less of an issue now in USA & Europe, it is still a concern in less developed networks.

There is still much activity focused in Europe in measuring and reporting aspects such as dial tone delay, although this is very rarely a constraint on the use of many of the European telecommunications services. However, in Europe, the relatively long lead times on repair of service and the substantial delays that still exist on the provision and support of many services are a real constraint on quality. After all, if you cannot receive the service in a reasonable time frame or have it repaired promptly, is this not the ultimate proof of poor quality of service? Most customers and users consider that the reporting of dial tone delay and related aspects are of no interest to them and would much prefer to see quality measured from a customer perspective and focused on activities which they see as important such as the provision of service, repair of service, accuracy of bills, the way complaints are dealt with and how easy it is to do business with their NP or SP; the measurement of network performance should be seen as only one aspect of the NP or SP quality of service.

Nevertheless, there is still a need to monitor certain network performance parameters to manage service quality across several networks and also to improve quality for less mature services. Despite the efforts of the "global" groupings there are few suppliers who can provide full end-to-end service and where they do so, it is often as a result of alliances which have to be founded on network service levels which must be measured. In less mature services users and customers do focus more on network performance, and thus in Europe this is currently strong in services such as GSM and ISDN and is likely to become an issue in POTS over the Internet, Frame Relay and ATM. In addition, this is still a very important aspect of quality activities and is becoming more so as service is increasingly provided across several networks. It is essential that an NP or SP is able to establish where service quality and transmission failures occur as customers are increasingly demanding end-to-end responsibility and effective management from their telecommunications supplier.

Other demands are being made by customers which are influencing regulators to take a much wider view of quality. This article provides an overview of some of those activities and in particular the involvement of user bodies, many of which still struggle to have their voices heard and encourage NPs and SPs to measure quality through the eyes and experience of their users rather than through internal network parameters. Let us begin our review by looking at bodies engaged in international activities and then look at country specific quality activities.

The International Telecommunications Users Group, INTUG

A grouping of international users and country user bodies with regional councils in Europe, USA, Africa and the Pacific Rim.

The European region has a regular quarterly meeting over two days at which a variety of matters are raised including Quality issues. During these meetings actions are agreed and reports prepared for issue. Work to date has focused on considering and developing Comparable Performance Indicators (CPIs) and Service Level Agreements (SLAs). The INTUG meetings are used to discuss input from various members which is often work done in the respective country user associations or, by a large corporate member; INTUG prepares an INTUG position with feedback from other countries and members, which it then publishes and uses to influence developments.

In addition, various articles are published in INTUG newsletters and presentations

[1] *The views expressed are the personal views of the author and have not been authorised by the bodies described in the article. This document is not to be duplicated or copied, in part or wholly, without the express permission of the author who holds copyright.*

are made by INTUG members and these have included articles by E Weiss on Measuring Quality and by the author on the CPI approach in the UK.

An example of the contribution made to quality improvement occurred at the Spring 1996 INTUG meeting in Brussels when EURESCOM informally sought attendees' views on their preferences for QoS reporting on pan-European services for a test bed, and the results, not in order of preference, were:

POTS, ISDN, GSM, Leased Lines and Frame Relay.

The European Virtual Private Network Users Association, EVUA

EVUA is a European based closed user group of around 30 large multinational users of Virtual Private Networks (VPNs). They are well advanced in preparing a draft SLA for their members' use with the suppliers with whom they currently have contracts: BT, AT&T & Unisource. This defines the activities and reports which they consider important and contains a wealth of material on service quality expectations of large multinationals.

Recently, EVUA have been developing a "request for information" for mobile voice services which they intend to use to influence the performance and pricing of these services in the same way as they influenced VPN services.

European ISDN Users' Association, Euro-ISDN, and the ISDN Memorandum of Understanding Implementations Management Grou, IMIMG

European ISDN Users Association and the IMIMG arm of ETNO work closely together and are concerned with QoS in ISDN service. Both of these bodies are strongly supplier based. The European ISDN Users Association is an umbrella organisation for the many European Country Organisations. IMIMG represents suppliers of ISDN service. Although described as user bodies they have a very strong supplier membership and relatively few users representing end users.

IMIMG is currently engaged in establishing a common approach to measuring QoS for ISDN and the EU has also funded research in this area, which is as yet unpublished.

The International Telecommunications Union, ITU

ITU has a Quality of Service Development Group (QSDG) which has helped to develop various definitions of Quality, e.g. E800 QoS definition and E.425/6 defining Answer Bid and Seizure ratios, etc.

In addition, QSDG have worked with the ETNO Working Group on Quality of Service to co-ordinate regular monitoring among ETNO/ITU members and subsequent reporting of Answer Bid Ratios (ABR) and Answer Seizure Ratios (ASR) in respect of international telephone service. Currently, 43 country destinations are reported on annually and data is available for a lesser number of countries for the last 6 years.

Several papers and summary reports on ASR and ABR have been prepared and published by these workgroups.

The European Telecommunications Network Operators, ETNO

This organisation comprises 34 (probably more by the time this is published) public network operators providing fixed public telecommunications. They are also considering permitting affiliation of mobile network operators.

In addition to the co-operative work with ITU mentioned above, the ETNO Working Group on Quality of Service run workshops to increase ETNO members' awareness of the growing needs for measurement and improvement in Quality of Service. Also the WG has defined a series of key QoS parameters:

- Supply time for initial network connection
- Response time for operator services
- Post dialling delay
- Successfully established call attempt rate
- Speech transmission quality
- Fault report per access line per year
- Fault repair time
- Availability of public pay-phones
- Billing complaint rate
- Dial tone delay.

The original five were expanded in July 1995 to incorporate the five CPI measures from the UK Telecommunications Industry Forum. The WG intends to use the detailed definitions developed by the UK Industry Forum.

The aim of the ETNO QoS parameters is to establish a harmonised definition and possible performance targets for pan-European services, in order to facilitate comparison with a view to improving service to customers. The WG is further studying the application of publication and comparison techniques and the feasibility of customer satisfaction surveys. These two activities are well advanced in the UK Industry Forum and the ETNO WG will use the developments as a basis for further study.

Another aspect under review by QSDG is the development of Service Quality Agreements (SLAs) among different PTOs.

European Union, EU

DG X111 is the prime body driving a growing interest in QoS, particularly as liberalisation and competition occurs in various services.

Various activities include:

- ONP Facilitation of a number of activities, recommendations, standards and directives covering a liberalised environment ensuring the market achieves "any to any and end to end" communication with a defined Quality of Service for all telecommunications services and networks of major public interest.
- Pre-liberalisation focused on guaranteeing access to services with a defined QoS (public service, usability, non-discrimination) and improving QoS to an EU harmonised high minimum level (pan-European service provision and harmonisation).

In the new competitive era emerging in Europe, it is expected to focus on guaranteeing access (universal service provided to defined minimum QoS), improving and maintaining certain pan-European services and availability of QoS targets and performance. It has issued various consultative papers and directives which specify and support QoS by the publication of data for leased lines and voice telephony, for example, with harmonised definitions.

The ONP voice telephony directive requires all member states to set and publish targets for eligible operators by early 1997, for supply time and QoS indicators for those operators covered by the directive in respect of:

- Supply time for initial connection
- Fault rate per connection, fault repair time & call failure rate
- Dial tone delay & call set up
- Transmission quality statistics
- Response time for operator services
- The % of coin & card operated public pay phones in working order
- Billing accuracy.

Further work is being considered to define visible and accessible QoS parameters, e.g. supply time, as well as less visible parameters such as availability in extreme circumstances. EU considers regulation will be the key to ensuring transparency and comparison of QoS with a well defined and harmonised minimum set of measures.

Service Quality activities in respect of European mobile services

The EC recently funded a study by OVUM UK Ltd and Intercai GmbH to establish a code of practice for Mobile Communications Service Provision.
A public consultation document has been issued by the consultants and this includes an interesting section on Service Quality. There is a high degree of commonality with other codes of practice currently in use in the UK by major terrestrial operators.

It proposes measures be established and monitored by a Code Authority for Europe which would address:

- Service connection
- Fault rates per connection
- Fault repair times
- Call failure rates for both call set-up and dropped calls
- Busy hour congestion
- Call set-up delay
- Billing accuracy
- Response times for customer support
- Response times for complaints and correspondence.

It is further proposed that the Code Authority will provide measurement methodology and targets. It will also publish the targets and review performance against targets, including an independent review.

ATM Forum

The ATM Forum is naturally at this early stage in the development of the services thin in user participation. However, there is a major activity funded by the EU to establish charging mechanisms for new ATM services. In order to agree a methodology it is necessary to define the service characteristics and, as a result of the study, the WG are considering definitions of ATM QoS to align with charging mechanisms. This activity has recently begun and TMA (the leading UK user body) is co-ordinating user input across Europe.

Country based activities

USA

The Federal Communications Commission (FCC), as regulator, has directed the Common Carrier Bureau to publish data which it had previously obtained from the Regional Bell Operating Cos. (RBOCs) in USA. The CCB has been monitoring QoS among RBOCs since 1983. The FCC began summarising and reporting on these in 1990 and continued through 1991 – 1994. The report for 1995 data is in the course of being published; all the published reports are available via the Internet.

The reports focus strongly on telephone service and the performance of key elements of each RBOC network component and presents these in a series of very complex league tables enabling comparison. The reports are regularly improved and there is a panel of users who provide input. Recent improvements have been stated as aiming “to focus much less on dial tone as modern digital switches are unlikely to be affected by slowed dial tone response”.

Each report includes data on key performance elements captured by the RBOCs as well as customer satisfaction surveys which they conduct. There is much misunderstanding among the traditional network performance adherents, who often consider surveys to be little more than opinion polls. It is true they may contain opinion, although this can be reduced by polling users who have recently had service experiences, but even where they do so they are still a useful indicator of customer perception. Customer perception is a major factor in making purchasing decisions. A selection of major measures includes:

- % of installation commitments met
- Average of missed installation days
- Average repair interval
- Trouble reports per 1000 access lines
- Trouble reports found per 1000 access lines
- Repeat trouble as a % of trouble reports
- Complaints per million access lines
- Number of access lines, trunk groups and switches
- Switches with downtime
- Average switch downtime in seconds per switch
- Unscheduled downtime over 2 minutes per occurrence
- Scheduled downtime over 2 minutes per occurrence
- Trunk groups with blocking over 3 month objective as a % of total trunk groups
- Reports on customer satisfaction in 3 categories: residential, small and large business.

The major user group in USA is the International Communications Association, and it has members of the supervising FCC panel which reviews and develops the measures. ICA is a member of INTUG.

Australia

The regulator is AUSTEL, and it has been publishing quarterly QoS data on the two operators in Australia, Telstra and Optus, since 1993. The data is not comparable but AUSTEL are working with both carriers to agree common definitions and reporting.

Currently they report in a variety of geographic categories – national, metropolitan, country, residential and business on the following aspects of QoS:

Telstra:

- Restoration of telephone service
- Service reports on telephone service
- % of faults cleared in 1 day and 2 days
- % Customer satisfaction with fault process
- Number of 2nd level faults re fault process
- % of 2nd level faults to all faults
- % of calls to fault help line answered within 15 seconds
- % of calls leaving fault help line without being answered
- % of calls lost by national, long distance and international by day and night
- Provision of service and upgrades to existing service as % of customer required dates
- Customer access lines installed
- A range of 7 reports on the performance of the operator services including one on customer satisfaction
- 4 reports on public pay-phones including one on customer satisfaction
- % of bills with itemised billing
- Total number of all types of complaints
- Total number of all 2nd level complaints
- Network loss for the public switched telephone network
- National call dropout rate for the Mobile telephone network
- Customer satisfaction by residential and business.

Optus:

- National and long distance services fault reports and clearance in a range of timing categories
- Number of total switch outages
- Number of major transmission line outages
- International call congestion reporting by ABR
- Residential customer satisfaction
- Service availability expressed as % total network coverage by population
- National call dropout rate for the Mobile telephone network
- Customer satisfaction by residential and business.

UK

In the UK there is an Industry Forum which comprises network operators, the regulator and user bodies, which has developed a process known as the Comparable Performance Indicators (CPI) process. Known as the UK Industry Forum, this voluntary body has agreed a set of definitions in respect of public switched telephone service and leased lines and reports separately on performance in residential and business categories for 5 key measures:

- % of service provided to agreed dates
- % of faults to 1000 access lines
- % of faults cleared to published target times (each operator offers several grades)
- % of complaints resolved within 20 days
- Number of complaints about billing accuracy per 1000 lines.

In addition to reporting on actual “hard” performance based upon data provided from the operators measurement systems, the qualifying members of the Industry Forum also conduct “soft” customer satisfaction surveys which canvass customers who have recently experienced service to determine their degree of satisfaction with the above processes. The surveys establish a critical “check and balance” effect in respect of the above five measures enabling both “hard” and “soft” measures to be published for competing operators.

The processes by which operators gather data on actual performance in the above categories and services, as well as customer satisfaction data against each service and category, are subjected to independent audits. The UK regulator Office of Telecommunications (OFTEL) reports on the data half-yearly. The reports began Jan 1996 and enable a true comparison of quality in a very competitive market, since all major NPs and SPs use the same definitions and processes.

Germany

The regulator, the Federal Ministry of Posts and Telecommunications, defines certain performance parameters for Deutsche Telekom the monopoly supplier in respect of:

- Installation time
- Repair time
- Quality of speech
- Operator service
- Billing process
- Public pay-phones
- In addition, the installation times and availability of leased lines are defined.

Financial penalties are paid by Deutsche Telekom to customers where it does not meet the parameters in respect of installation and repair. However, the parameters are very broad with delays typically measured in terms of days for repairs and weeks for service provision. In addition, the regulator conducts ad-hoc surveys to establish user opinion and perception of services.

Netherlands

A similar approach to that of Germany is adopted by the regulator, the Ministry of Transport Public Works and Water Management – Telecomms and Post Dept, in what is still a monopoly by PTT Telecom. The regulator defines the following QoS obligations:

- Publish telephone directories on a regular basis
- Respond to service applications in 10 days
- Provide standard facilities in 3 months
- Provide special facilities in 12 months
- Investigate malfunctions in 2 days
- Restore normal service in 5 days
- Operate a complaints commissioner
- Operate public pay-phones.

Where these targets are not met the customer is financially compensated.

For GSM and analogue mobile services the regulator requires:

- Provide nation-wide service
- Provide QoS data as will be defined by regulator (except the next one)
- Guarantee a certain call-blocking ratio
- Operate a complaints commissioner
- Provide a help desk /subscriber database
- Free service to emergency calls.

It is questionable as to whether several of these indicators would be considered quality indicators by customers.

Sweden

Tele 2, possibly the second largest operator in Sweden, has a pilot scheme to make use of Call Detail Records (CDR) as well as data contained in SS7 signalling to monitor network performance. Tele 2 are considering facilitating switch manufacturers to agree a standard CDR format. They also have some interesting ideas on the use of automated customer databases of events to automatically report on customer needs and performance.

Norway

Telenor, the current monopoly, has developed a pilot system of automatically gathering live traffic monitoring using SS7 signalling and an independent system of creating artificial calls which are then monitored and reported upon. These focus upon:

- Successfully established call ratios
- Waiting time for dial tone
- Waiting time for ring tone
- Properly placed calls
- Transmission quality measured by echo, delay, distortion, noise, etc.
- Service charges.

Telenor is currently testing the independent system with an international field trial with Netherlands, Denmark, Finland, Iceland and Sweden. In addition, Telenor are piloting in Norway a trial of advanced SS 7 signalling monitoring.

France

France Telecom makes extensive use of customer satisfaction surveys and feeds the data gathered back into the France Telecom organisation to drive QoS improvements. It has adopted a standard survey technique and set of questions and conducts 1 million customer surveys per year, for which participants are paid a small fee. Customers rate services in 5 grades.

Residential and business customers are surveyed in respect of:

- Installation of telephone lines, leased lines and ISDN
- Repair of above
- Perceived call quality of above
- How FT commercially deal with customers. This involves a complex set of questions on receipt of enquiry, processing and implementation
- How customers rate FT with others who are able to offer competitive services.

Some future developments

In the true spirit of quality all these efforts are subject to continuous improvement. Limitations of space do not permit me to describe more than the future directions that the UK Industry Forum is considering for CPI requirements in 1997. The first step has been for the various user bodies, including the TMA on behalf of the business community, to review progress against their previous requests. The following is a "shopping list" being submitted to other user bodies to agree a common user body approach:

Additional measures being sought for 1997:

All categories & all services

The publication in all categories of business and residential, direct and indirect, dedicated and switched of the operators' own target times for all of the measures, i.e. provision, repair, fault reporting, complaint handling and bill accuracy.

The reason for this is that it will add considerably to business and residential consumer usefulness of the current data if, when reviewing the performance in % terms, the consumer can see the scaleability to aid making a judgement on the comparability of the performance. For example: one operator's 100 % achievement against internal targets of 3 months elapsed time is not as effective a performance as 99 % achievement for the same task, against internal targets of 5 days. There are many other examples where adding data on the actual targets and publishing it in the reports in close proximity to the % performance will aid usefulness.

In requesting this, TMA is not moving away from the position that the performance assessment is above all about meeting promises, nor does it wish to see reporting move towards proscribed elapsed times. But it considers that with the wide range of actual targets the CPI process may have a built-in incentive for operators to set themselves low targets in order to achieve high %.

ISDN

This is an area of increasing importance to business and residential alike. It is available in both basic and primary rate service. TMA would like to see a set of measures added for each of the primary and basic ISDN services. As primary rate service is widely available from all UK operators, and they are currently measuring it as part of the switched voice service and encountering some difficulty with understanding the previously agreed definition, it may be appropriate to separate this measure using the ISDN definitions.

Regarding the basic rate service; although BT is the major supplier there is a growing number of Other Licensed Operators (OLOs) offering basic rate ISDN at tariffs which differ, and it is becoming increasingly important to prospective users to have performance data to aid informed choice in this area.

Requirements for beyond 1997

TMA future requirements can be further divided into two sets:

1. The addition of "Managed Services" to Switched and Dedicated measures
2. Extensions to the existing switched and dedicated measures.

Additional Measures

TMA believes that the following topics must be put forward for consideration for both soft and hard measures in the future:

- Provision of upgrades to existing services as % of customer required dates
- Provision of new services as % of customer required dates
- Average of missed installation days/month or 6 months
- Faults found per 1000 lines
- Faults reported as % of faults found per 1000 lines
- Repeat faults reported as a % of fault reports
- Repeat faults found as a % of faults found
- Average repair interval
- % of faults repaired in 1 day and 2 days
- Number of access lines, trunk groups and switches
- Number of switches with downtime
- Average switch downtime in seconds per switch
- Unscheduled downtime over 2 minutes per occurrence
- Unscheduled downtime as % of total downtime (over 2 minutes per occurrence)
- Scheduled downtime over 2 minutes per occurrence
- Scheduled downtime as % of total downtime (over 2 minutes per occurrence)
- Accuracy of inventory
- Complaints per thousand access lines
- Total number of all types of complaints
- Total number of 2nd level complaints
- How operators deal commercially with customers (ease of doing business)
- Reports on customer satisfaction to be expanded to three categories: Residential, small and large businesses.

Managed Services Matrix

VOICE	DATA	MULTIMEDIA
Centrex	Frame Relay	Internet (IP)
VPN	ATM	ATM
Charge Card	SMDS	Video (Video Conferencing)
	X.25	Broadcast
	FBS	Dark Fibre

Conclusion

NPs and SPs are shifting the focus of their quality management systems to have a stronger customer perspective. Users' and customers' views are receiving more credibility and attention. The better our perceived quality of service becomes, the higher our expectations.

Network interworking concepts for the competitive multiple service deliverer environment

BY PETER GERRAND[1]

Monopoly telecommunications carriers have traditionally taken full responsibility for end-to-end performance of national calls, and set all tariffs. Neither principle is practicable or desirable in a multiple Service Deliverer[2] environment. This paper describes some innovative concepts in interconnection policy achieved in Australia through the conceptual framework of the Regulator (AUSTEL)'s *Interconnection Model* and the use of a peak industry forum, the Network Interworking Industry Forum (NIIF), to develop a consensus framework based on that Model to support the multiple Service Deliverer environment. The Model's general principles may be useful in other national regulatory environments as well, to support the introduction of intelligent network services (and other advanced services) across distinct network boundaries.

This paper was presented at Networks'96, the 7th International Network Planning Symposium, "Planning Networks & Services for the Information Age" held in Sydney, Australia, 24 – 29 November 1996.

1 Introduction

The major liberalisation of the US telecommunications market in February 1996, the further deregulation of the Australian market in July 1997 and the planned liberalisation of the European market in 1998, will create regulatory environments with new challenges for industry stakeholders, especially end users.

The 'up-side' for end users is the benefit of greater choice amongst Service Deliverers (SDs), and the cheaper prices that usually arise as a result of competition. The 'down-side' for users is that the new environments will be characterised by

- Volatility of the market, as new SDs enter and exit the market
- "Invisibility" of many Service Deliverers, other than Access SDs, to the end users
- Potential confusion as to which SD has responsibility for end-to-end performance of a particular service carried across several networks
- Potential confusion as to which SD will bill a Calling Party for a service, and
- Lack of transparency of advanced services across other SDs' access networks.

The challenge for responsible industry participants, not just the industry regulators, is to ensure that end users receive the benefits of competition without the 'down-side' elements listed above. The industry as a whole will get a bad name if users' experiences with telecommunications services, and the management of these services, deteriorate as a result of providing services across several SD networks. In Australia, the industry and the national government alike have decided that it is in the industry's interests to provide, as far as possible by consensus, an agreed framework for network interworking in the post-July 1997 regulatory environment. This goes far beyond agreement on technical standards, and deals with the rights and responsibilities of all Service Deliverers of public network services, both to each other and to the end users of the services they will share the responsibility for delivering.

2 A future network interworking scenario

The complexity of future scenarios for service interworking is illustrated in Figure 1 – the example of a call from a mobile phone to an Intelligent Network-based 'intelligent routing' global number service, using a '13' code in Australia – similar to the 1-900 number services in North America.

For this kind of service, it is quite likely four or more – in fact up to seven [3] – different Service Deliverers can be involved in jointly delivering the service between certain pairs of customers. The hypothetical example in Figure 1 assumes that a Service Provider (AAPT) provides a '13' intelligent routing service on behalf of a contracting customer (e.g. Qantas Airways). It is assumed that AAPT has decided to provide the intelligent routing and service management functions using a database connected to its own switch, which is connected to the Optus network. Furthermore, AAPT has contracted to connect to the other two carriers (Telstra and Vodafone) where calls to this service are going to originate or terminate on those other carriers' networks.

Figure 1 shows the case where the calling party (A-Party) is connected to the Vodafone mobile network. The Vodafone mobile switch and gateway analyse the B-Party '13' number (131313) dialled by the A-Party, deduce that it is allocated to the Optus network, and route the call – together with the Calling Line Identity (CLI) number – accordingly. A switch in the Optus network analyses the B-party number further, and deduces that this number has been allocated to its customer AAPT, and routes the call – together with its CLI number – to the relevant AAPT switch. The AAPT switch deduces that the B-party number is Qantas' 13-number, and follows Qantas' instructions, embodied in the AAPT database software, in order to perform the number translation from the B-Party number to the desired terminating number (B') for the call, which will depend upon the CLI, Time of Day, Day of Year and any other significant information.

Let us assume for simplicity that the terminating number is 'geographical', meaning that its geographical destination can be deduced from the first five digits of this network number, e.g. (03) 9344 9305. The terminating number – e.g. that of a Qantas enquiries office in a different city – will be passed to the Optus switch, which will then route the call – with its CLI – via its long-distance network to the Telstra gateway switch closest to the terminating number. The Telstra access network will then endeavour to complete the call connection.

[1] *Special consultant to AUSTEL (the Australian telecommunications regulator) for the development of AUSTEL's Interconnection Model for the introduction of new telecommunication services.*

[2] *Because the legal distinctions between licensed carriers and non-carrier service providers differ, not only from country to country but also between different regulatory timeframes in the same country, the general term Service Deliverer was coined in [1] to refer to any carrier or non-carrier service provider.*

Several interworking issues are thrown up by this example, which need to be – and are – resolved in the Australian Interconnection Model [1, 2]:

- Who is responsible for the end-to-end performance of the call?
- Given that the A-Party will usually be ignorant of who the Service Deliverer is for the call (a situation which will grow even more volatile once single-number portability between Service Deliverers is introduced), to whom should the A-Party report service difficulties, and who should take action on the complaints?
- Who decides where the Originating Access Carrier should hand-over the call to the next carrier (Optus) – Vodafone or Optus? And should this be as close as possible to the A-Party, or as close as possible to the AAPT switch?
- Does the Service Provider (AAPT) for this '13' service have the right to set an access (A-Party) tariff that is different from the usual flagfall charge set by the Originating Access Carrier (Vodafone) – or only the right to set the tariff for the contracting customer (Qantas)?
- Does an Access Carrier have the right to improve its competitive position by charging an excessive access interconnect fee to other carriers or Service Providers for calls to or from new services (not covered by the existing Government-determined interconnect fee agreements)?

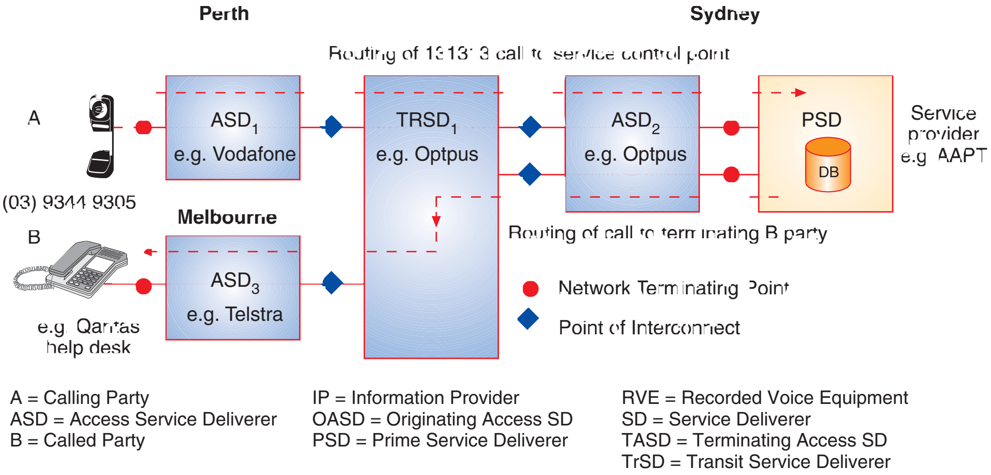

Perth
Sydney
Routing of 131313 call to service control point
A
ASD$_1$
e.g. Vodafone
TRSD$_1$
e.g. Optpus
ASD$_2$
e.g. Optpus
PSD
DB
Service provider e.g AAPT
(03) 9344 9305
Melbourne
B
ASD$_3$
e.g. Telstra
e.g. Qantas help desk
Routing of call to terminating B party
Network Terminating Point
Point of Interconnect
A = Calling Party
ASD = Access Service Deliverer
B = Called Party
IP = Information Provider
OASD = Originating Access SD
PSD = Prime Service Deliverer
RVE = Recorded Voice Equipment
SD = Service Deliverer
TASD = Terminating Access SD
TrSD = Transit Service Deliverer

Figure 1 Interconnection of an Intelligent Network service across multiple Service Deliverer networks

3 Overview of the interconnection model

To achieve sufficient generality and simplicity, a set of eleven *Definitions* of new terms has been introduced, together with a set of seventeen *General Principles* which provide the operational basis for service interworking when introducing new services (or extending existing services via a new carrier or Service Provider). A number of *Observations* have also been included, to provide additional insights.

The new Interconnection Model consists of the Definitions, General Principles and Observations, together with an AUSTEL Endorsed Process for conducting commercial negotiations on interconnection, using the Model, and reporting on successful implementations.

Since mid-1995, and industry-wide forum, the Network Interworking Industry Forum (NIIF) has taken over from AUSTEL the task of developing a consensus framework, based on the Model, to support service and network interworking in the future multiple Service Deliverer environment.

To encourage the role of Service Providers in providing a full range of services, the Interconnection Model begins by defining the term *Service Deliverer* (SD) to describe both licensed carriers and non-carrier Service Providers, and then proceeds to express most of its General Principles in terms of the generality of Service Deliverers (SDs), rather than having separate rules for carriers and SPs. This has proved advantageous given continuing uncertainty in Australian government policy as to the proposed distinctions between carriers and SPs.

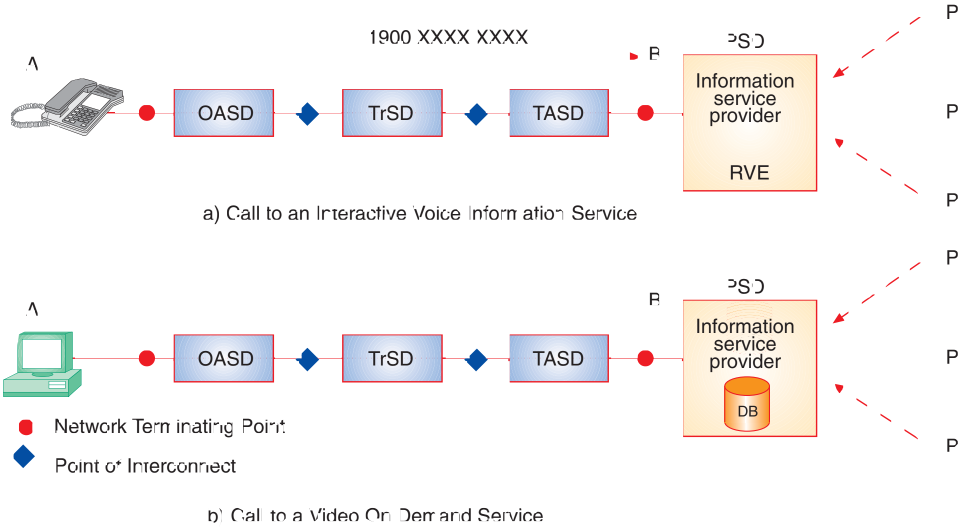
The Model is intended to be general enough to apply to the entire range of foreseeable telecommunication services that will require interconnection across multiple Service Deliverers (SDs). However, many of the General Principles are specific to switched (call-based) services, e.g. PSTN services, because of their importance, e.g. in respect to preselection.

The Model distinguishes between the *Contracting Customer*, who contracts with a Service Deliverer (SD) to provide a particular service, and *other end-users* of that service.

In the above example based on Figure 1, Qantas is the Contracting Customer for a '13' service supplied by AAPT. Any other PSTN user may call that '13' number, but they do not normally have a direct contractual relationship with AAPT. Instead, they have a contractual relationship with their Access SD – e.g. Vodafone or Telstra – which in turn has a contractual relationship with Optus for the purpose of interconnecting calls to this service. Optus has a contractual relationship to AAPT as its Access SD, which completes the chain of contractual relationships from the A-Party making the call to AAPT. A similar chain of relationships extends from AAPT as provider of this '13' service to the B' Party to whose network number the call is terminated.

AAPT is defined to be the *Prime Service Deliverer* (PSD) for the Contracting Customer's '13' service, and the other SDs who support this service through interconnecting their networks – Vodafone, Optus and Telstra – are defined to be the *Supporting Service Deliverers*. The PSD has the right to determine the tariff for the Contracting Customer (CC), and also to negotiate with the relevant Access SDs to determine the tariffs for other participating users of the service – the A-Parties. The PSD also carries the prime responsibility for the end-to-end performance of the service – whether the PSD is a carrier or not.

Negotiations for setting tariffs could bog down in stand-offs if more than one SD were to be assigned an equal Prime


1900 XXXX XXXX
A
B
PSD
Information service provider
RVE
IP
IP
IP
OASD
TrSD
TASD
a) Call to an Interactive Voice Information Service
A
B
PSD
Information service provider
DB
IP
IP
IP
OASD
TrSD
TASD
Network Terminating Point
Point of Interconnect
b) Call to a Video On Demand Service

A = Calling Party
ASD = Access Service Deliverer
B = Called Party

IP = Information Provider
OASD = Originating Access SD
PSD = Prime Service Deliverer

RVE = Recorded Voice Equipment
SD = Service Deliverer
TASD = Terminating Access SD
TrSD = Transit Service Deliverer

Figure 2 Interconnection of interactive Information Services across multiple Service Deliverer networks

Service Deliverer role for a given service. For this reason the Model assigns the PSD role to a single SD – the SD that provides the service on behalf of the Contracting Customer (CC). This modelling decision deliberately favours service innovation, whereas the alternative choice – to assign the PSD role to the Originating Access SD – would give too much power to the incumbent Access SDs to frustrate the introduction of new services by their competitors.

The Interconnection Model is intentionally *user-oriented*. Consideration of the contractual relationships between SDs starts with the relationship between the CC and PSD, then introduces the further conceptual building-blocks of the Access SD and its customers, and the Transit SD that interconnects other SDs. By contrast the earlier AUSTEL Interconnect Model – and those of many Models used overseas – tend to deal exclusively with the relationship between Access Carriers and Interconnecting Carriers.

In future service scenarios, the ordinary user is unlikely to be aware of the identity of the PSD for any of a whole range of services they invoke (e.g. by dialling the numbers they see in advertisements). On the other hand, they would normally be able to ascertain who their Access SD is – the SD who provides their wireline telephone, their mobile phone or pager (or their PABX or LAN if part of a private network service), from the evidence of invoices received or a copy of the service agreement. Accordingly, General Principle G4 of the Model assigns a *particular responsibility on the Access Service Deliverer* for responding to user complaints on any services to which they provide access, by co-ordinating actions to fix the problems with the PSDs.

Lastly, the Model identifies which of the SDs is the PSD, in the case of a wide range of current and imminent service classes. In the case of switched services, the Model also indicates the point in the network at which calls should be handed over by the Originating Access SD – “unless otherwise agreed between the SDs”. The Model introduces the distinction between “global numbers”, “geographical numbers” and carrier-specific mobile numbers, in order to logically deduce which SD should be the PSD for a range of PSTN services.

Two further applications of the Model are given in Figure 2, showing the versatility of the Interconnection Model when applied to current and future on-line services. In Figure 2A an interactive voice recorded information service is accessed by dialling a 1900 number, and in Figure 2B a Video-on-Demand service is accessed via a yet-to-be determined numbering plan for broadband services. In both these cases the Contracting Customers are various Information Providers, who supply information to the Information Service Provider (ISP).

The Model deduces that the PSD is not the Originating or Terminating Access SD but the ISP, and assigns privileges and responsibilities accordingly. If the ISP numbers are global numbers, handover of calls should take place at the POI closest to the A-Party – “unless otherwise agreed between the SDs”.

4 Service-specific factors

To permit the practical interworking of a new or existing telecommunication service between several SDs, the General Principles need to be fleshed out with practical details, specific to each proposed service type. The NIIF’s expectation is that each significant public network service will be defined in a Service Definition Document. An example is given in page 76–77 of [1].

5 Concluding remarks

The broader work of the NIIF, including the collaborative development of a National Network Performance Plan, are described in an accompanying paper by Peter Darling of Telstra.

6 References

1 Gerrand, P. AUSTEL’s new interconnection model for the multi service deliverer environment. BTCE Communications Research Forum 1995, Sydney, 26 September 1994; extended and reprinted in *Telecommunications Journal Australia,* 44, (3), 1994, 59–74.

2 *AUSTEL interconnection model for the multi service deliverer environment : final report.* Melbourne, AUSTEL, March 1995.

3 Gerrand, P. Network interworking post-1997. In: *Telecommunications Journal Australia,* 46, (2), 1996, 41–50.

Measuring Quality of Service in the public switched telephone network

BY EDDIE ULRIKSEN

Intrusive and non-intrusive methods of measuring the quality of service in a telecommunications network are discussed. This includes supervision of test traffic or monitoring live traffic, either by built-in-test functions in the network elements, signal monitoring equipment or by separate, independent test generated traffic equipment.

It is concluded that independent systems can give quite complete and realistic measurements of the end-to-end quality as perceived by the user and the network performance as engineered by the network operator.

Two types of independent systems in use by Telenor networks today, one based on an artificial testing method and one for live traffic monitoring of SS No 7 are presented.

1 Introduction

Service and network providers are in a much more competitive situation than before. To the subscribers and the users, quality of service becomes an important criterion when choosing among several service providers. In some cases the service provider may introduce a quality-dependent price level as an aspect of the service offerings. National and even international regulatory bodies may in the near future define certain minimum figures for service quality that have to be fulfilled by approved service providers.

With today's large traffic flow, even a relatively small increase in traffic carrying capacity of an existing network can mean a significant increase in the total income. By analysing the results of quality measurements, e.g. quality degradations due to long delays or congestion, the network operator can optimise the maintenance procedures, and thus increase the utilisation of the existing network resources.

When the users get accustomed to high quality services, their basic attitude towards telecommunications will be to increase the quality demand for both old and new services. On this background, it is important both to the service provider and the network operator, to continuously monitor the quality of the services delivered to the customers.

Therefore, serious considerations must be given to choose the best possible principle for measuring quality. Keeping in mind ITU's definition of quality of service (E:800): – "The collective effect of service performance which determine the degree of satisfaction of a user of the service" – it seems obvious that at least three basic requirements should be fulfilled:

1. Quality of service (QoS) is defined from the user's point of view, and therefore, it should be measured accordingly. This implies that end-to-end measurements will give the most accurate results.
2. According to a recognised principle in the quality assurance field, quality of service should be measured by equipment that is independent of the traffic carrying elements of the network.
3. The method should be universal in the sense that it may produce comparable results for all parts and regions of a network, domestically and internationally.

A variety of measurements methods and tools have been used in the past. Basically, there are two different approaches:

(1) Monitoring live traffic, or **(2)** Monitoring small, controlled volumes of test traffic.

In either case, measurements can be made by built-in facilities in the network elements, or by independent, external equipment.

2 Reasons for Telenor's selection of measuring equipment

Telenor's present maintenance philosophy is "Controlled Corrective Maintenance", also called "automated fault repair". It is imperative to have equipment that accurately facilitate evaluation of attained goals and objectives.

It is further understood that focus must be on the services that are provided and not primarily on the technical equipment. Services must therefore be specified and measured in accordance with the way our customers (subscribers and users) perceive them.

Another major element in Telenor's strategy for operation and maintenance of the digital network is that technical implementations, as far as possible, should be independent of the telecommunications and maintenance systems. Basis for this strategy is:

- Supplier and systems independence
- Better economy when the same O&M system can be used in different telecommunication systems
- More stable and unified routines for maintenance crews
- Systems should be controlled by other independent systems.

3 ITU-TS recommendations on Quality of Networks and Telephone Services

The definition of Quality of Service in rec. E.800 is wide and includes subjective customer satisfaction. In this paper the aspects of QoS covered are focused on the identification of parameters that can be directly *observed and measured* at the point where the service is accessed by the user, i.e. end-to-end measurements and/or monitored by retrieving data from different network elements, (non-intrusive measurements).

The network performance (NP) part of E.800 is measured in terms of parameters meaningful to the network provider and are used for system design, configuration, operation and maintenance. NP is defined independently of terminal performance and user actions.

Although E.800 is a complete system where QoS is the ultimate result of also the network performance characteristics, the QoS- and NP points of view are different and outlined in Table 1. E.800 is given a more comprehensive treatment in other articles in this issue.

Subscriber-to-subscriber service quality measurements on the public switched telephone networks are specified in ITU-T (CCITT) recommendations: E.422 Observations on international outgoing telephone calls for quality of service, E.434 Subscriber-to Subscriber measurement of the public switched telephone network, and E.435 End-to-End Transmission Measurement Techniques. Table 2 lists SS No 7 – TUP/telephone user part- cause values categorised in "fate" of the call classes as: Successful calls, unsuccessful calls due to: Network, customer, etc. Similar listing for ISUP-ISDN user part of SS No 7 is found in Table 3. Table 4 (Appendix A) shows listing of cause values definitions (for ISUP) in rec. ITU-T E.850.

E.435 notes that to permit quality parameters, (as loss Q.551, attenuation distortion, group delay O.81, total distortion 141, idle channel noise O.41, impulse noise O.71, round trip delay, echo and clipping) to be compared with *QoS targets* or for comparisons to be made between administrations, the measurement technique must be defined.

Table 1 Distinction between quality of service and network performance

Quality of Service	Network Performance
User oriented	Provider oriented
Service attribute	Connection element attribute
Focus on user-observable effects	Focus on planning, development (design), operation and maintenance
Between (at) service access points	End-to-end network connection elements capabilities

Table 2 Observations of international outgoing telephone calls for quality of service (SIGNALLING SYSTEM No. 7 – Telephone User Part)

Country of origin . . . .
Outgoing international exchange . . . .
Group of circuits . . . .
Service automatic/semi-automatic
Period: from . . . . to . . . .

Point of access:
National side . . . .
Link circuits . . . .
Outgoing side . . . .
Time of observations . . . .

Category	Number			Percentage	
	UBM	Sub-total	Total	Sub-total	Total
C.1 Calls successfully put through			-		-
C.2 Unsuccessful calls					
C.2.1 Due to the Customer			-		-
2.1.1 Ring tone received but no answer		-		-	
2.1.2 Busy tone received		-		-	
2.1.3 Subscriber busy	SSB	-		-	
C.2.2 Due to the Network			-		-
2.2.1 Circuit Group Congestion	CGC	-		-	
2.2.2 Switching Equipment Congestion	SEC	-		-	
2.2.3 National Network Congestion	NNC	-		-	
2.2.4 Digital Path Not Provided	DPN	-		-	
C.2.3 Due to the Customer and/or Network					
2.3.1 Unallocated Number	UNN	-		-	
2.3.2 Line Out of Service	LOS	-		-	
2.3.3 Address Incomplete	ADI	-		-	
2.3.4 Call Failure	CFL	-		-	
C.3 *Total calls monitored* (categories 1–2)			-		100
C.4 Unsuccessful calls: Positive indication of failure from outgoing international exchange			-		
4.1 Congestion on outgoing international circuits		-			
4.2 All other indications		-			
C.5 Successful calls with defects. These calls all included in category 1					
5.1 Non reception of ANC on chargeable calls		-			
5.2 Other calls with defects		-			
C.6 Unsuccessful calls due to the signalling system failure. These calls are not included in the previous categories			-		-
6.1 Protocol failure		-		-	
6.2 Signalling network failure		-		-	

Table 3 Observations of international outgoing telephone calls for quality of service (SIGNALLING SYSTEM No. 7 – ISDN User Part)

Country of origin . . . .
Outgoing international exchange . . . .
Group of circuits . . . .
Service automatic/semi-automatic
Period: from . . . . to . . . .

Point of access:
National side . . . .
Link circuits . . . .
Outgoing side . . . .
Time of observations . . . .

Category	Number			Percentage	
	UBM	Sub-total	Total	Sub-total	Total
D.1 Calls successfully put through			-		-
D.2 Unsuccessful calls: normal class (CV 1-to-15, 17-to-31)					
D.2.1 Due to the Customer			-		-
2.1.1 Ring tone received but no answer	-	-		-	
2.1.2 Busy tone received	-	-		-	
2.1.3 User busy	17	-		-	
2.1.4 No user responding	18	-		-	
2.1.5 No answer from user	19	-		-	
2.1.6 Call rejected	21	-		-	
D.2.2 Due to the Customer and/or Network			-		-
2.2.1 Unallocated number	1	-		-	
2.2.2 No route to destination	3	-		-	
2.2.3 Send special information tone	4	-		-	
2.2.4 Number changed	22	-		-	
2.2.5 Destination out of order	27	-		-	
2.2.6 Address incomplete	28	-		-	
2.2.7 Facility Rejected	29	-		-	
2.2.8 Normal, unspecified	31	-		-	
2.2.9 Other causes		-		-	
D.3 Unsuccessful calls: Resource unavailable class (CV 34-to-47)			-		-
3.1 No circuit available	34	-		-	
3.2 Network out of order	38	-		-	
3.3 Switching equipment congestion	42	-		-	
3.4 Other causes		-		-	
D.4 Unsuccessful calls: Service or option not available class (CV 50-to-63)			-		-
D.5 Unsuccessful calls: Service or option not implemented class (CV 65-to-79)			-		-
D.6 Unsuccessful calls: Invalid message class (CV 87-to-95)			-		-

Table 4 (Appendix A) Cause values definitions (Rec ITU-T E.850)

a)	**Normal class**
1	- Unallocated (unassigned) number
2	- No route to specified transit network (national use)
3	- No route to destination
4	- Send special information tone
5	- Misdialled trunk prefix (national use)
6	- Channel unacceptable
7	- Call awarded and being delivered in an established channel
8	- Preemption
9	- Preemption – circuit reserved for reuse
16	*- Normal call clearing*
17	*- User busy*
18	*- No user responding*
19	- No answer from user (user alerted)
20	- Subscriber absent
21	- Call rejected
22	- Number changed
26	- Non selected user clearing
27	- Destination out of order
28	- Invalid number format (address incomplete)
29	- Facility rejected
30	- Response to STATUS ENQUIRY
31	*- Normal, unspecified*
b)	**Resource unavailable class**
34	*- No circuit/channel available*
38	- Network out of order
39	- Permanent frame mode connection out-of-service
40	- Permanent frame mode connection operational
41	*- Temporary failure*
42	*- Switching equipment congestion*
43	- Access informatin discarded
44	- Requested circuit/channel not available
46	- Precedence call blocked
47	- Resource unavailable, unspecified
c)	**Service or option unavailable class**
49	- Quality of Service not available
50	- Requested facility not subscribed
53	- Outgoing calls barred within CUG
55	- Incoming calls barred within CUG
57	- Bearer capability not authorized
58	- Bearer capability not presently available
62	- Incosistency in designaed outgoing access information and subscriber class
63	- Service or option not available, unspecified
d)	**Service or option not implemented class**
65	- Bearer capability not implemented
66	- Channel type not implemented
69	- Requested facility not implemented
70	- Only restricted digital information bearer capability is available (national use)
79	- Service or option not implemented, unspecified
e)	**Invalid message (e.g. parameter out of range) class**
81	- Invalid call reference value
82	- Identified channel does not exist
83	- A suspended call exists, but this call identity does not
84	- Call identity in use
85	- No call suspended
86	- Call having the requested call identity has been cleared
87	- User not member of CUG
88	*- Incompatible destination*
90	- Non existent CUG
91	- Invalid transit network selection (national use)
95	*- Invalid message, unspecified*
f)	**Protocol error (e.g. unknown message) class**
96	- Mandatory information element is missing
97	- Message type not existent or not implemented
98	*- Message not compatible with call state or message type non-existent or not implemented*
99	- Information element/parameter not existent or not implemented
100	- Invalid informatin element contents
101	- Message not compatible with call state
102	- Recovery on timer expiry
103	- Parameter non existent or not implemented-passed on (national use)
110	- Message with unrecognized parameter discarded
111	- Protocol error, unspecified
g)	**Interworking class**
127	- Interworking, unspecified

This is essential for comparing numeric data and for the compatibility between different manufactures and different Public Network Operators and Carriers.

A model is necessary to relate the various serveability measures to the different call phases and identify the elementary causes. Rec. E.820 shows a proposed call model to be used by network operators. (Figure 1 shows a model of the serveability performance on a basic call in the telephone network.) It enables cooperating PNO/TOs to:

1) Make comparison between the same parameters measured on different network services

2) Evaluate the offered QoS-level

3) Identify, locate and remove by appropriate O&M actions, service/network problems.

Customer interviews can be used to assess- (a) QoS perceived by the customer on calls generally and specifically on calls recently performed, (b) the customer's perception of the improvement required. Interviews must distinguish between:

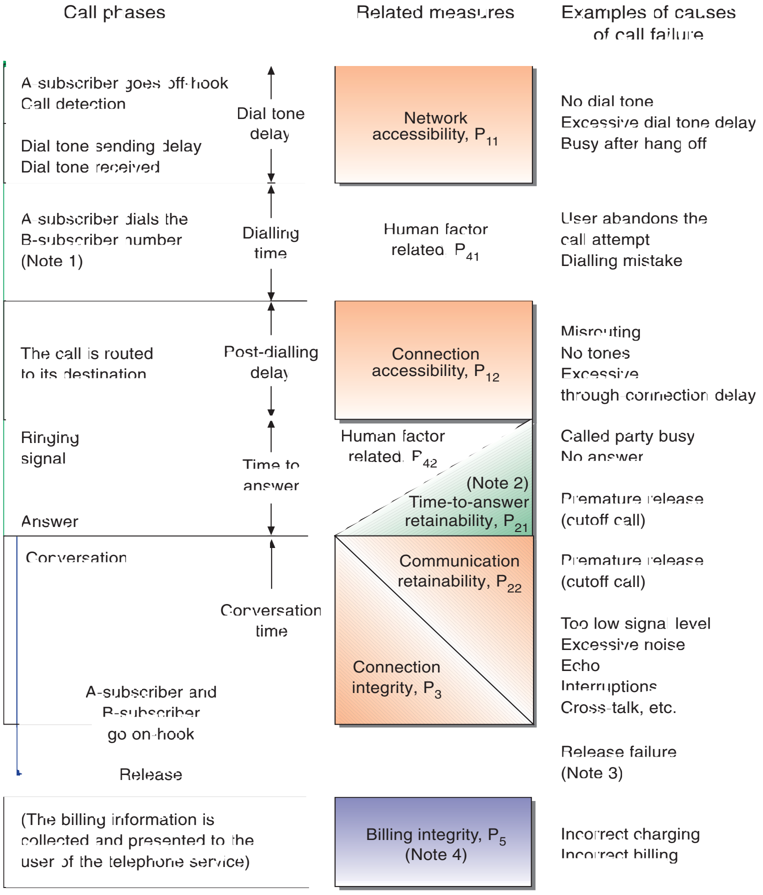
- National and international calls
- Residential and business calls
- Small/low(infrequent) users and heavy/big frequent users, etc.

Interview questions are normally related to the last call made, but some questions involve the more general or cumulative experience of the telephone user. Complaints, interviews results and network performance measurements should be collected in the same period and analysed collectively in order to reach a general overview of the current QoS – ref. Figure 2 from ETNO.

ITU's general definition of quality of service was stated in the introduction of this article. However, to enable quantitative measurements in the network the following constrained definition has in the context of this article so far been used by Telenor:

> *Quality of Service – the percentage of calls that, with a defined transmission quality (in both directions) within a defined time, arrives at the correct called party, and are charged correctly.*

In order to evaluate the quality of service according to this definition, a number of



Call phases
Related measures
Examples of causes of call failure
A subscriber goes off-hook
Call detection
Dial tone sending delay
Dial tone received
Dial tone delay
Network accessibility, P_{11}
No dial tone
Excessive dial tone delay
Busy after hang off
A subscriber dials the B-subscriber number (Note 1)
Dialling time
Human factor related, P_{41}
User abandons the call attempt
Dialling mistake
The call is routed to its destination
Post-dialling delay
Connection accessibility, P_{12}
Misrouting
No tones
Excessive through connection delay
Ringing signal
Time to answer
Answer
Human factor related, P_{42}
(Note 2) Time-to-answer retainability, P_{21}
Called party busy
No answer
Premature release (cutoff call)
Conversation
Conversation time
Communication retainability, P_{22}
Connection integrity, P_3
Premature release (cutoff call)
Too low signal level
Excessive noise
Echo
Interruptions
Cross-talk, etc.
A-subscriber and B-subscriber go on-hook
Release
Release failure (Note 3)
(The billing information is collected and presented to the user of the telephone service)
Billing integrity, P_5 (Note 4)
Incorrect charging
Incorrect billing

Note 1 The routing of the call may start before all digits have been received.
Note 2 Normally, the measure referred to time-to-answer retainability is not assessed because negligible.
Note 3 The release of a call is not a separate phase in this model. A release failure may result in network inaccessibility for a new call.
Note 4 The billing integrity has been shown for completeness, but is not a part of serveability performance.

Figure 1 Model of the serveability performance on a basic call in the telephone network

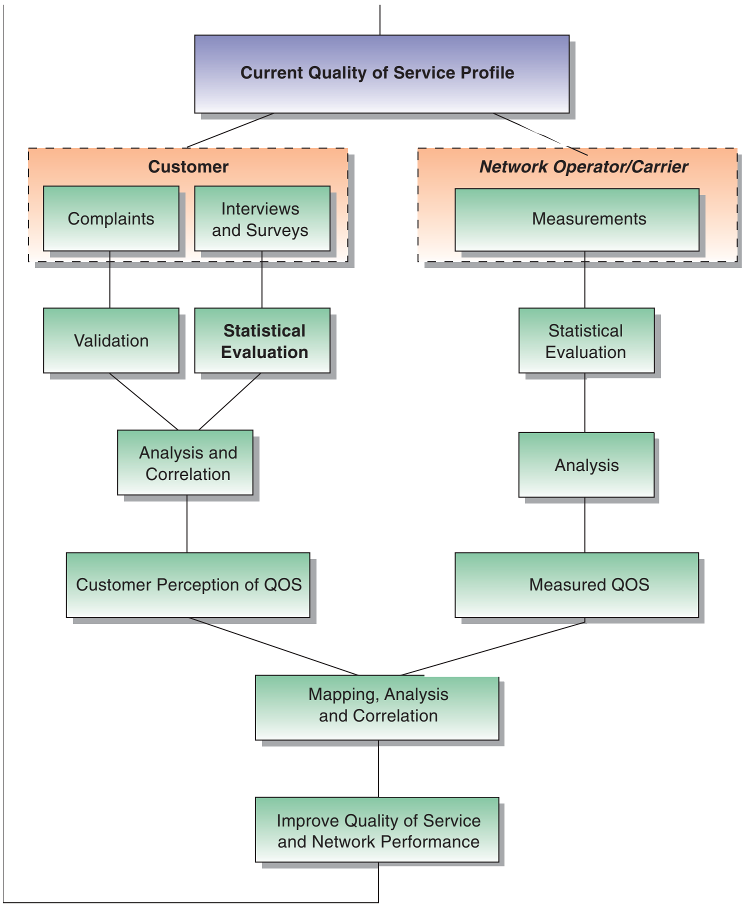


Figure 2 General procedures for analysing and relating data

parameters must be measured. Examples are:

- Successfully established calls, peak traffic vs. idle time
- Waiting time for dialling tone
- Waiting time for ringing tone
- Properly placed calls
- Transmission quality control in both directions as, echo, delay, distortion, noise, etc.
- Service charges.

A TRT (Traffic Route Test) system can make a whole range of these measurements and therefore give effective product/service control.

4 Experience gained with use of a Traffic Route Test (TRT) system in the Nordic (and European) public switched telephone network

4.1 Historical background

Telenor has used TRT systems in the Norwegian national telecommunications network (local/regional/long distance network), as well in the intra Nordic network for some years. A Nordic discussion started in the late 80s / beginning of the 1990s to measure the end to-end quality between; Denmark, Finland, Norway and Sweden. Iceland joined the Nordic Group on Quality of Service Improvement in 1991. PTT Telecom BV, the Netherlands, was the first company outside the Nordics to join the “Group” in 1993. The aim was to get a picture of the quality of service of the telephone network as perceived by the customers and users.

Figure 3 presents the service loss (defined by Telenor as loss) in busy hours in the Norwegian trunk network. The “fast” reduction in measured service loss was due to several maintenance activities with focus on the network parameters measured by the TRT system. The “trend” describes the “target” level set by Telenor. Target today is 1.5 % for intra regional test generated traffic measured by the TRT.

Table 5 Examples of technical problems uncovered by independent monitoring of telecommunications systems

Transmission problems:

- Oscillator fault in radio link station 1800 channels (noise)
- Ice problems on radio link antenna
- Antenna coverage (random) fault, 140 Mbit system, bit fault
- Noise on 60-groups in a 300 group
- High room temperature in a radio link station
- Fading on radio links (over lakes in special weather conditions)
- Bad connectors in transmission equipment
- Bit faults on 2-, 8- and 34 Mbit systems
- Faults on hybrids
- Faults on coaxial cables
- Impedance mismatch
- No voice transmitted in a timeslot in a 2 Mbit system
- Coil loading (pupinisation)
- Humid and wet cables in the ground
- Bad/missing solderings
- Cross-talk (noise) caused by bad insulation in jumper fields and distribution frames
- Attenuation/frequency distortion (highest frequency in the speech band is too low)
- Noise on some coax cables caused by amplifiers

Dial tone problems:

- Missing dial tone (fault in registers or other equipment in local exchange)
- Unknown tone (incorrect frequency or other tones on the line)

Metering faults:

- No B-answer signal on certain lines
- Metering pulse before B-answer signal
- Too long/short metering intervals
- No metering for a new area code
- No metering for certain area codes due to software fault in digital exchange

Congestion and miscellaneous faults:

- Faults on relays, selectors and MFC-equipment
- Faults in digital trunk modules
- Faults in wiring and cable connections
- No alternative route for certain area codes
- Misrouting
- Sticking (remanence) of relays and selectors
- Faults on SS No 7 signalling route
- Vibration in building caused by street traffic (interrupting calls)
- Bundles too small in busy hour
- Silicon (insulation from coils) on relay contacts
- 2 Mbit system out-of-service
- 2 Mbit system in loop

Wrong B-number reached:

- Misrouting in selectors
- Register faults (counting chain/code receivers)
- Faults in MFC equipment

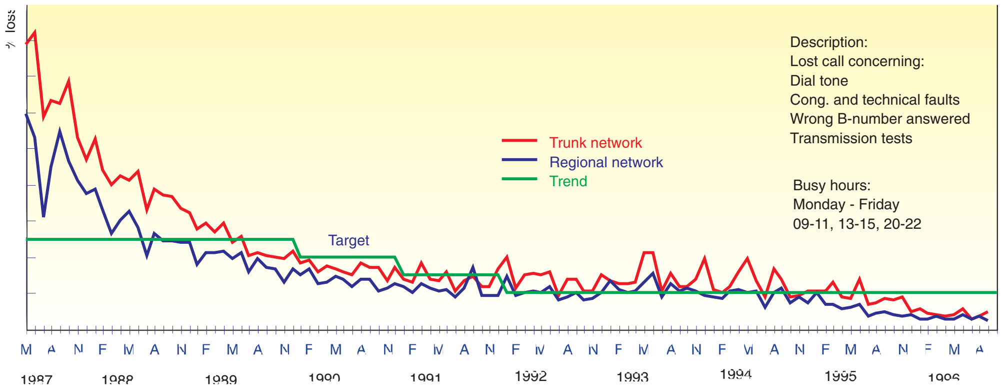


Figure 3 Loss in busy hours, trunk network

Since the introduction of TRT system in 1987 we have experienced increased revenues of about 0.4 % due to a higher percentage of successful call attempts.

Table 5 lists examples of technical problems uncovered by independent monitoring using TRT.

4.2 Nordic field trial

A Nordic field trial was formed in the spring of 1989. Two traffic route test units were placed in each country, one in an analogue exchange and one in a digital exchange. This to check the quality in the old analogue network, the rapidly expanding digital network as well as between the two “networks” including the mixed analogue – digital network. As the TRT units are connected to the exchange as a normal subscriber (detecting only tones and incidents which can be detected by the subscribers) the units needed some new software to be able to detect the tones used in the different countries. The TRT centre (the Master unit TRTC) was located in Oslo, Norway and the Nordic ISCCs (International Service and Coordinating Centres) had the responsibility of the operation and the following up of the TRT results.

The first full scale test programme was performed in October 1989. The results from this test were discussed in the Nordic Quality of Service Group and considered so useful that a second test period was agreed upon in December the same year. The following weekly test periods (lasting several months) underlined the importance and necessity of continuing the TRT measurements in the Nordic network.

4.3 Description of the TRT programme

The TRT system is controlled by the TRTC, initially loaded with information about the different tones, test numbers, configuration of the TRT test units etc. in the Nordic network. The data is normally used for computing test traffic matrices for the entire network and for special defined network groups. As test calls within each country were not required, each TRT unit was defined as a network group and the number of test calls from each TRT unit per hour was set to 36 in busy hours and 48 in low traffic hours. The possibility of measuring within one country was introduced later in the field trial period. Busy hours, were defined as: Monday – Friday 0900–1100, 1300–1500 and 2000–2200 (all other periods regarded as low traffic hours). Test patterns and test types were decided by the operator in the TRTC on a test by test basis. The following test types were scheduled in the field trial (using Central European Time):

1) Service loss (S-sequences busy hours): Monday – Friday 0900–1100, 1300–1500, 2000–2200
2) Network test (N-sequences non busy hours): All weekdays 0030–0300, 0400–0700
3) Extended transmission test (T-sequences): All weekdays 1600–1900.

Service loss and network test are identical tests. The only difference is the time of the tests (service loss in busy hours and network test in low traffic hours). These tests are normal calls where the following parameters are measured: Correct address, transmission quality, post dialling delay, dial tone delay and any abnormal termination of the call.

The extended transmission test is a normal service loss/network test extended with attenuation measurement on 3 frequencies (400, 820, 2800 Hz) and noise measurements. All the parameters measured in service loss and network tests are measured as well.

The TRTC generates detailed test patterns for each TRT unit (TRTU) and sends the test packets to the respective TRTUs. The TRTUs are connected to the exchange as normal subscribers and generate the test calls as specified in the packet. The test calls are made in specific time slots ensuring that the called TRTU is always idle. Each TRTU is able to simultaneously generate and receive one call in the same time slot.

As indicated in Figure 4 there are two sets of thresholds, one for defining faults (poor quality) and one set for defining lost calls. Threshold 1 can for example be used to report specified service norm violations. Threshold 2 can represent the level where a subscriber would consider a call to be lost. A call can in principle, have numerous faults, but only one loss reason. In other words, a test call is aborted when a loss threshold is exceeded. Detailed data are stored in the TRU only for calls exceeding the set fault/loss threshold.

Table 6 is a list of loss/fault categories in TRT.

4.4 Procedures for running the tests

At the end of the test, TRTC collects the results from the TRTUs. The results are processed with separate set of thresholds for classifying lost calls and calls that exceed quality targets. After each test a sequence report gives an overall list of fault types and detailed fault report available for each test unit.

The results from each test are accumulated in numerous statistics. The statistics are separate for busy hour and low traffic hours giving the possibility to, for example, analyse the amount of congestion and technical faults. In addition there are statistical reports for degrading quality such as long dialling delay, too high attenuation, etc.

Congestion will normally not occur in the low traffic hours, hence the loss will be due to technical faults. As a "rule of thumb", the increase in loss in busy hour is due to a congestion situation in the network.

In the field trial a selection of agreed reports were sent to the Nordic ISCCs operations centre by facsimile every Monday, during the test periods. If necessary, detailed reports were retrieved from the database and distributed to all parties involved.

A considerable amount of test calls were generated during the field trial months giving a good overview of the call quality in the intra Nordic network. In one month some 40,000 test calls were generated and classified as: Service tests, network tests and transmission tests. Overall service loss (on all traffic relations) was at first approximately 12 % and was reduced to 5 % in busy hours during 8 months. Similar results were obtained for service loss outside busy periods of the day. A 12 % service loss measured the first month of the field trial was reduced to 4.5 % after 8 months.

Although maintenance work during day time, restoration of transmission systems, bringing new circuits and testing of switching equipment and other types of "network activities" effected the TRT-results, it reflected the true quality of service (and network performance) given to the users of the telecommunications services during the measured time. We therefore requested the ITMC (International transmission) and ISMC (International switching) to report all maintenance activities that might degrade the traffic and transmission quality. The information included:

1. Date and time a fault was reported and actually occurred
2. The nature of the reported fault
3. The reporting location
4. The location of the fault, when found
5. The actual fault condition found and the temporary corrective or permanent action taken.

5 Being high quality network operators

To be a trustworthy network operator it is important to have a good view of the quality of service that is provided and offered to the customers. The use of the TRT system has contributed to an increase in the quality in the intra Nordic network and is still considered a useful tool for keeping a high QoS level. The allocation of appropriate human resources (technicians, engineers, signalling experts, etc.) to follow up the results was considered very important by the carriers involved in the tests.

5.1 Multi carrier/network operator environment

The multicarrier situation gives the consumer the choice of selecting carrier depending of destinations needed. This new situation puts focus on the coordination of Quality of Service programmes between the network operators and international carriers.

The Nordics QoS Group, was one of the first in the world to introduce artificial test generated calls across country borders on a large scale.

However, other carriers around the world also had developed measurements programmes and test-systems (mostly for domestic coverage). They saw the need for a standard in this area. The idea was supported by ITU's Telecommunication Standardization Sector, and a working group in Study Group 2 in the ITU-TSS developed a recommendation for international Quality of Service measurements, E.435: End-to-end transmission measurements techniques.

5.2 Further development

The new generation of the TRT system from Ericsson named, NEAT (Network Evaluation and Test System) has made possible advanced transmission tests following the ITU's recommendations E.434 and E.435. These tests will be used in an international field trial during 1997. See Figures 5 and 6.

Members of the field trial are: Telecom Finland, Telia Sweden, Telecom Iceland, Tele Denmark, PTT Telecom the Netherlands, Telefonica Spain, Swiss Telecom and Telenor Norway. The NEAT system is already in use in some countries, but this is the first time an international cooperation of this size is formed.

Each of the members will locate two units in major cities in their country, with the centre unit place in Oslo. Several test matrices will be designed to measure the transmission quality (call quality) and the establishment of the call. The primary objective of this international cooperation is to improve the quality of the international telephone service between members network, (cross border traffic) through:

1) Monitoring the QoS level
2) Distribution of the information obtained at regular intervals (e.g. once a month, quarterly, etc.).
3) Cooperating in the identification of any network problem and its resolution
4) Pursuing other activities which may improve the QoS, as a result of the measurement programmes agreed.

5.3 General test functions to be used in the field trial in 1997

The main categories of test functions are:

- Quality tests
  - Network Group Tests (between and within a specified number of units in a network group)
- Fault Detection Tests
  - Network group tests
  - Single call tests
- Advanced transmission tests
  - Echo measurements
  - Round trip propagation delay

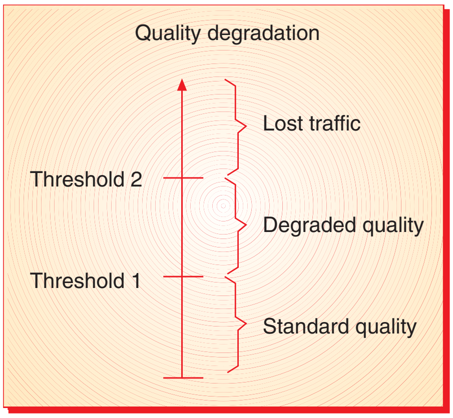


Figure 4 Thresholds for reporting quality degradations

Table 6 List of loss/fault categories in TRT

Category	Type	Fault category
DT	Loss	No dial tone before timeout No dial tone was detected before the waiting time expired
	Loss	Undefined tone instead of dial tone Another tone was detected instead of the specified dial tone. This could be either a tone with wrong frequency or a tone with a “known” frequency but wrong cadence
	Fault	Dial tone delay limit exceeded Dial tone was detected later than the specified limit
RT	Fault	No ringing tone No ringing tone was detected before the TT-tone
	Fault	Tone delay limit exceeded A tone was detected later than the specified limit. This could be ring tone or any of the other specified tones
C & T	Loss	Unexpected new dial tone A new dial tone was detected when this was not supposed to happen
	Loss	Congestion tone A congestion tone or busy tone was detected after dialling the B-number. (Busy tone has the same spec. as cong. tone)
	Loss	Information tone Information tone detected after dialling the B-number
	Loss	Unknown tone An unknown tone was detected instead of ring tone. This could be either a tone with the wrong frequency or a tone with a “known” frequency with the wrong cadence
	Loss	No answer Silent after dialling
	Loss	Too short time slot The time slot expired before the test was completed
BNO	Loss	Wrong B-No. / no answer Nothing is heard after detecting the ring tone
TSM	Loss	TT-tone too low + TT-tone too high The attenuation of the TT-tone is too large to continue the call. (Greater than 22 dBm)
	Loss	Communication error The transmission quality is too bad to continue communication
	Fault	Error in TT-level A to B Attenuation on TT-tone is above defined limits
	Fault	Error in TT-level B to A Attenuation on TT-tone is aobve defined limits
	Fault	Error in transmission/noise A to B Attenuation on any of the three transmission tones is above deinfed limits. Or the idle channel noise measurement exceeds the specified limits
Noise	Fault	Error in transmission/noise B to A Attenuation on any of the three transmission tones is above deinfed limits. Or the idle channel noise measurement exceeds the specified limits
	Fault	S/N to B Signal to noise ratio is below specified limits
	Fault	S/N B to A Signal to noise ratio is below specified limits
	Fault	Telegram errors A to B Number of telegram retransmissions has exceeded the specified limit
	Fault	Telegram errors B to A Number of telegram retransmissions has exceeded the specified limit
	Fault	Low level dial tone Dial tone level below specified limit
	Fault	Low level tone Dial tone level below specified limit
OTH	Loss/ Fault	Wrong B-ID The NTU has reached another NTU (B-side) than expected
	Loss/ Fault	Wrong B-access position The B-side answers the call on another line than expected
	Loss/ Fault	A-line busy No resources are available for the NTU to make a test call

- Clipping measurements
- Call continuity measurements
- Impulse noise measurements/Idle channel noise.

In addition, there are some optional test functions:

- Billing integrity
  - Charging tests
  - Toll billing tests.

These optional tests will be performed in the field trial to a limited extent. Measurements will only be made on request and not be a part of the regular matrices.

The preliminary testing of functions and applications has already started and continued throughout 1996. The regular tests are expected to start in the first months of 1997. After a short field trial period, and evaluation having been made, targets for service levels will be defined for the different traffic relations. Offers to join the IQWF (International QoS Working Forum) will be sent to interested network operators throughout the world. Also the use of mobile units will addressed by the Forum. A final evaluation of the complete system will be done during 1997.

5.4 The next phase, testing of the Integrated Service Digital Network, ISDN

With the rapid growth in demand for ISDN there is an associated demand for accurate, fast and reliable testing and verification of the services offered by the Integrated Service Digital Network, ISDN.

Most European countries have signed an MOU (memorandum of understanding) and thereby agreed to supply Euro-ISDN services. Today, most countries have declared that they will meet (or already meet) the MOU requirements, although Euro-ISDN is being implemented at different rates in different countries.

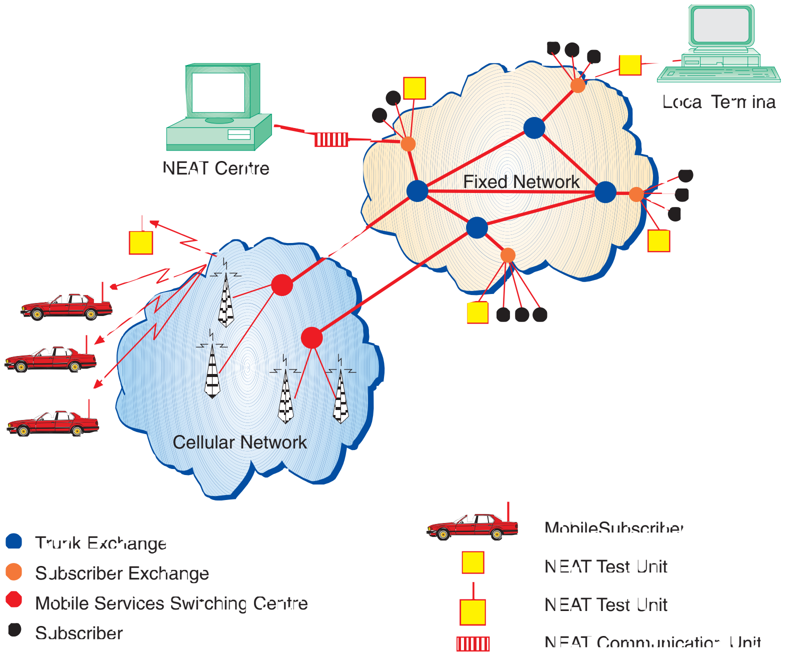


Figure 5 System overview containing fixed and cellular networks

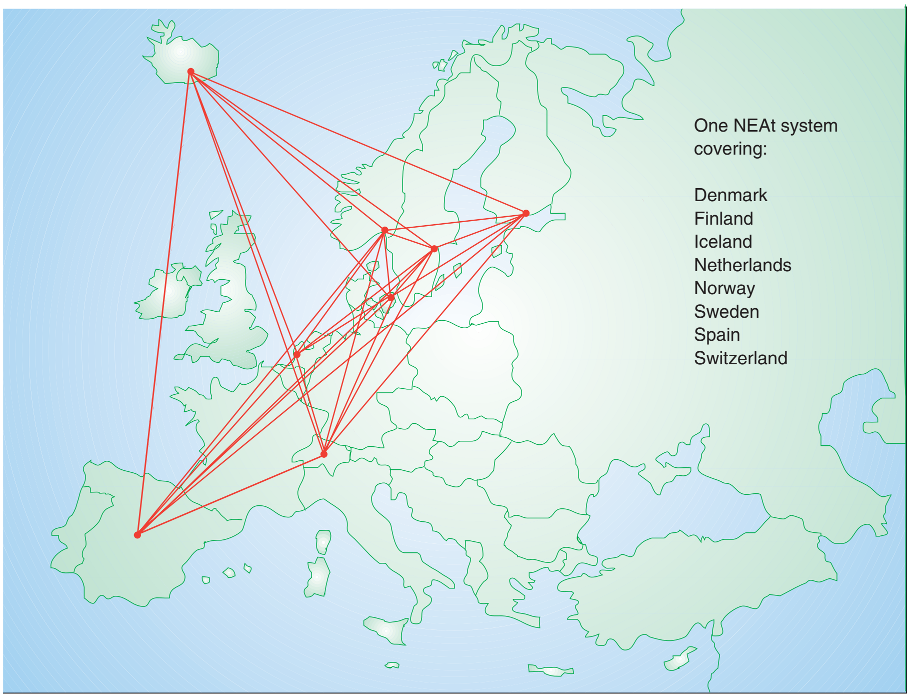


Figure 6 Testing of international networks

A prime requirement to ISDN testing equipment is versatility, as it has to meet the needs of a variety of public network operators, private networks and service providers. On the technical side a lot of areas have to be covered by the equipment: The physical quality of the line must be measured to ensure that there is no degradation of the signal. With the use of existing copper wires found in the local telecommunications loops (designed for analogue transmission) there is a need for testing a range of factors, such as impedance, continuity, signal balance and electrical loading. There are several levels of protocol testing to be carried out on new ISDN installations. The interoperability of different ISDN products and/or services, terminals and switching equipment also needs to be checked. Etc.

5.5 Network Quality of Service parameters for ISDN

The relevant QoS parameters for ISDN bearer services have been identified by

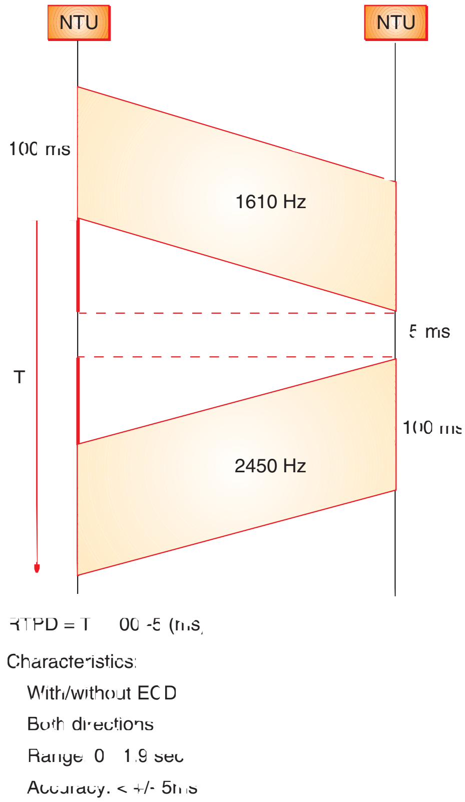


Figure 7 Example of a round trip propagation delay measurement

investigations conducted by users, equipment and network providers, and by collecting recommendations and standards issued by international groups and standardization bodies. Due to the different characteristics of the bearer services used by ISDN applications, the related QoS parameters have accordingly been classified in two main classes (a, b) namely:

a) Parameters for applications using single B-channel circuit switched ISDN connections

1) Bearer service 64 kbit/s unrestricted:

Connection establishment
- Successful call ratio for ISDN/ISDN calls
- Connection establishment delay
- Reliability of calling line identification

Information transfer
- Propagation delay (one way)
- Errored seconds
- Severely errored seconds
- % error free calls of a certain time duration
- Premature release ratio

Disconnection
- Connection release delay
- Release failure probability

2) Bearer services "speech" or 3.1 kHz audio

Connection establishment
- Successful call ratio for ISDN/ISDN calls
- Successful call ratio for ISDN/IPSTN
- Connection establishment delay
- Reliability of calling line identification

Information transfer
- Propagation delay (one way)
- Premature release ratio
- Speech transmission quality

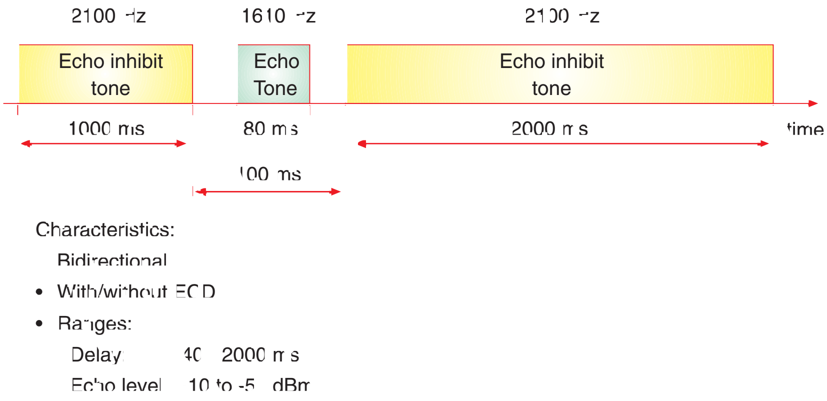


Figure 8 Example of an echo measurement

Disconnection
- Connection release delay
- Release failure probability

3) Definition of the application groups, AG

All the previous QoS bearer parameters have different meaning and relevance to the user depending on the user characteristic and the application. In order to simplify the process of mapping the appropriate QoS bearer parameters to each ISDN application, the following four application groups have been identified:

AG1. Transport of signals in the audio frequency band:
Characteristics: transport, for interactive or non-interactive use, of an analogue audible signal; in most cases the applications will be using the bearer services speech or 3.1 kHz audio; this application group includes the ISDN to PSTN calls. Applications include:

- Speech telephony
- File transfer in the voice band
- Fax Group 3
- Audio conferencing
- Radio 7 kHz

AG2. Highly available data transmission:
Characteristics: data-transport with very high probability of successful connection establishment; the most important requirement that these applications put on the network is that once a connection is needed, it can be trusted that it is there, and that it is there fast. Applications include:

- Back-up for leased lines
- Telemaintenance, tele-surveillance, tele-alarm

AG3. Transfer of files and documents:
Characteristics: reliable transfer of data with moderate propagation delay and error performance requirements; these applications occur when the initiating user wants to transmit data from his location to another user; because the use is not really interactive, the requirements on propagation delay are less severe than for certain other applications; also, this allows time for (forward or backward) error correction, which puts less severe requirements on the error performance of the connection. Applications include:

- Fax Group 4
- File transfer
- Image transfer.

AG4. Interactive telecommunication applications and database access: Characteristics: data-transport in an interactive environment; this application group includes interactive digital telecommunications between two or more users, and also applications where a user has (quick) access to a database. Applications include:

- Access to Internet, ordering systems, electronic payment, ISDN videotext
- Teleteaching, remote assistance
- Video-telephony, (multi-point) videoconferencing, multimedia applications
- Joint document editing, joint Computer Aided Design

Remarks:

1. The applications of the AG4 group can further be splitted in 2 subgroups, corresponding to applications involving single or multiple B-channel circuit switched ISDN connections respectively.
2. Some "general" applications, for instance LAN-LAN interconnect, belong to more than one application group. This is because these "applications" in fact comprise several types of ISDN applications.

4) Relevance of the measurable items for each application groups

Table 7, Importance level of the NP-parameters, indicates, for each parameter what is, with respect to each application group, the importance level of the parameter (High {H}, Moderate {M} or low {no indication}.

b) Parameters for applications using multiple B-channel circuit switched ISDN-connections (from 2 to 30 channels)

Connection establishment

- Successful call ratio for multiple channel ISDN/ISDN calls
- Connection establishment delay (excluding time taken by the terminals)
- Reliability for calling line identification (for all B-channels)

Information transfer

- (One way) propagation delay (for all B-channels)
- Errored second (for all B-channels)
- Severely errored seconds (for all B-channels)
- % errored seconds (for all B-channels)
- Propagation delay difference between B-channels
- Premature release ratio (for all B-channels)
- Speech transmission quality

Disconnection

- Connection release delay (for all B-channels)
- Release failure probability (for all B-channels).

6 Non-intrusive method for live traffic monitoring

The importance of understanding the customers perception of service has led to many telecom companies undertaking market research to indicate potential areas for improvement. Recent work by several companies has confirmed a very simple but important relationship between the customer's reception of call connection and network performance. For example, a customer measure "percentage of calls connected first time", correlates to a collection of network parameters such as, answer, ring tone no reply, customer busy and originating customer replacement, and not directly to an answered response.

Results that are based upon answer responses, do not provide a clear indication of network performance and can therefore lead to inappropriate investigation and/or action being taken. However, "network delivery" parameters, e.g. cause values form the TUP in SS No 7 such as *un-allocated number, call failure, ring tone no reply, address incomplete/ insufficient digits dialled, national network congestion, destination out of order,* etc., give a clearer indication of performance dropping below an acceptable level and are more accurate measurements of the network performance.

Use of "network delivery" parameters also provides valuable data for setting network performance targets. By inspecting the success or inversely the failure of the network, the margins of improvement can be determined and appropriate target set.

Bearing in mind the so far lack of information of network delivery parameters, telling *the fate of the call,* the idea to a non-intrusive measuring equipment was born. Telenor has developed such a non-intrusive system for measuring the network performance in the digital network. This new monitoring concept is called S-MAN (a Signalling Monitoring system for Analysis of Network performance).

6.1 S-MAN, a Signalling Monitoring system for Analysis of Network performance

The purpose of the S-MAN system is to gather information on teletraffic by monitoring the signalling messages in SS No 7. The data base is a relation database which can store huge number of call data record, CDR's (call detailed records). It can combine several types of reports and the user/operator indicates which time period he wants to look into/or observe

S-MAN is connected to the STPs (signalling transfer points) which transits SS No 7 messages between the different telecom regions, including all numbering areas and services.

This is done by monitoring the SS No 7 signalling link based on a measuring equipment from ELMI and Hewlett &Packard. Telenor has developed software for further processing the signalling information, from such user parts as TUP(telephone user part), ISUP(ISDN user part), INAP(Intelligent Network Application Part).

The system provides information and produces reports in the 3 following main areas:

1. Network performance
2. Correctness of the signalling and network elements.
3. Fraud detection.

Furthermore, detailed searching criteria can be specified and detailed reports produced. Some examples are:

1. Dedicated service reports, (e.g. international report)
2. Listing of one specific country (e.g. USA country code #1) and/or area code/city code in the same country (e.g. code 212 or 718 for New York).
3. Originating number report, (e.g. calls from Oslo/city code 22 to London/city code 171 or 181).

4. Terminating number report (e.g. all traffic to Italy Rome, country code 39 + city code 6).
5. Sectional "routing" report, specifying the 2 Mbit/s traffic stream.
6. User defined Query.
7. "Top 10 report" or "bottom 10 report"
8. Geographical international report, e.g. Africa, America, USA/Canada, Asia, Australia, Europe etc.
9. Matrix report, e.g. traffic performance between regions and specified services report.

A live traffic matrix presents accessibility parameters between all number areas. The presented parameters are accessibility/Answer Bid Ratio, response time and holding time of a call. Historical data for required numbering areas or required originating or terminating number series can also be produced.

Figures 9 to 12 show examples of "grouping" of TUP cause values and are explained below.

S-MAN/1 Weekly UNN (unallocated number/wrong or not existing number), Address incomplete (insufficient digits dialled), CLF(clear forward/caller stops the call attempt for any reason). The UNN+ADI+CLF can be seen as user oriented problem or loss. Use of recorded announcement can in some situations reduce this "loss" significantly.

S-MAN/2 shows the ABR (answered calls) + SSB (subscriber busy) + Ring no answer/reply). These cause values gives us an indication of the "optimal" answer rate that can be accomplished to a destination. In this example: ABR figure of approximately 90 %. The last 10 % is found in the S-MAN/1 figure as user oriented loss. Use of "different answering functions" available in digital exchanges can reduce the "ring no answer" part about 15–20 %, as shown in this example. The SSB part (subscriber busy) can also be reduced, for instance with optimal use of "business and residential" telephone lines. As an example: It's quite common for small and medium business companies to use the fax machine both for outgoing as well as for incoming messages. "Setting off" one (or more) fax machines/numbers for only incoming facsimile, can improve "access" to company and tribute to an increase in profits. At the same time, the call completion will be improved both for the business and the network operator indicated as a higher efficiency of available circuits.

Table 7 Importance level of the NP parameters for each application group

This table has been filled by application experts, according to their experience in this area. A detailed analysis for each combination of parameter and application group (AG), concerning the importance of the parameter for the application group, and the possibility to derive target values for the parameter, has been carried out. The group AG4 has been splitted in applications requiring 1 or n B-channels, as mentioned in the previous chapter, to show the different relevance of the QoS parameters within the two subgroups.

Connection establishment	**AG1**	**AG2**	**AG3**	**AG4** **1 B ch.**	**AG4** **n x B ch.**
Successful call ratio for ISDN/ISDN calls	M	H	M	H	
Successful call ration for ISDN/PSTN calls	M				
Successful call ratio for multiple cahnnel ISDN/ISDN calls					M
Connection establishment delay	M	H		M/H	M/H
Reliability of Calling Line Identification	M	H	H		for all B-ch.

Information transfer	**AG1**	**AG2**	**AG3**	**AG4** **1 B ch.**	**AG4** **n x B ch.**
One way propagation delay	M	H		M	for all B-ch. H
Propagation delay difference between B-channels					H
Errored seconds		M	M	M	for all B-ch.
Severely errored seconds	H	M	M	H	for all B-ch.
% error free calls of a certain time duration		M	H	H	
Premature release ratio	H	H	M	H	for all B-ch. H
Speech transmission quality	H				H

Connection release	**AG1**	**AG2**	**AG3**	**AG4** **1 B ch.**	**AG4** **n x B ch.**
Connection release delay	M	H		M	for all B-ch. H
Release failure probability	M	H	H	H	for all B-ch. H

S-MAN/3 shows the response time for a call, (also called post dialling delay) which is the time from the caller has dialled the last digit until response is

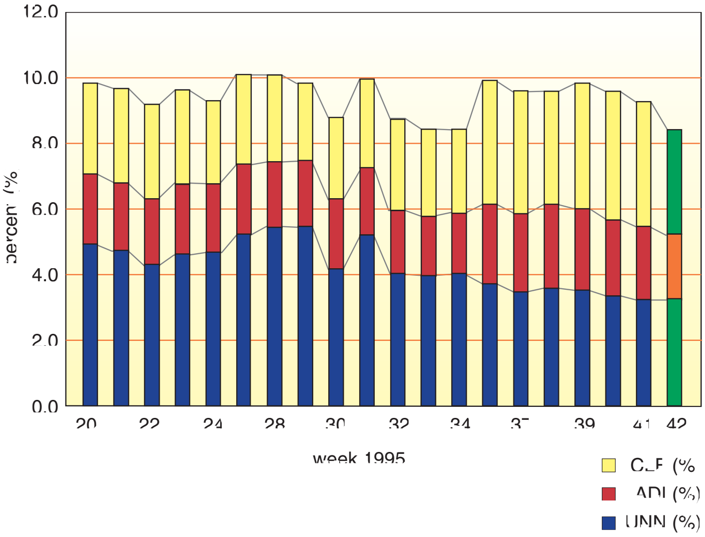


Figure 9 Weekly UNN, ADI and CLF

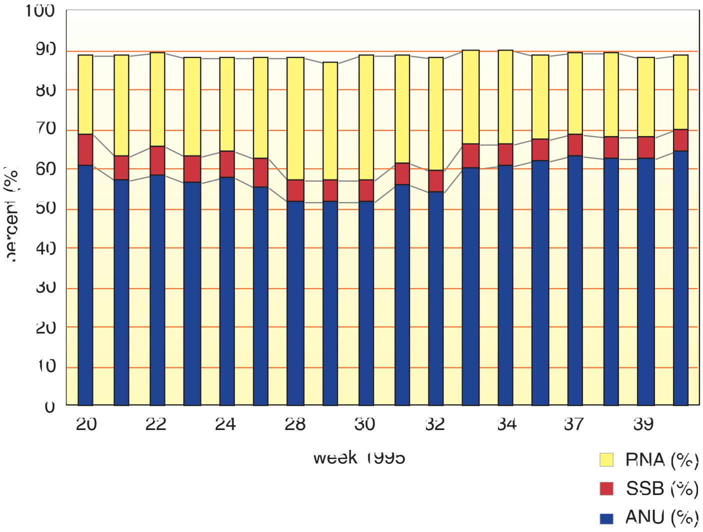


Figure 10 Weekly ABR compared with SSB and RNA

received from the network (an answer, subscribe busy, ringing tone, a recorded announcement etc.) The response time gives a good indication of the "far end network", being an old analogue one or a modern digital network. Also technical related problems in setting up the call through the network can be detected by the different response times measured with S-MAN.

S-MAN/4 shows the average HT (sec.) holding time of all calls (including call attempts failing) and the CT (sec.) the average conversation time. Short holding times can indicate, extremely heavy traffic loads in the network or faults in the network (e.g. killer circuits). Short conversation time can indicate the type of traffic carried over the network, referring to transfer of data. facsimile messages and a normal speech call. The CT parameter tends to fluctuate in accordance with the time of the day, busy hour vs. slack hours, as traffic during day time exists of a mixed business and residential/private traffic.

The objective of using S-MAN is to observe all services from all geographical areas in Norway based on the numbering series. All traffic relations and traffic streams can in principle be monitored. This means for example that calls originating in Oslo and destined to any other place nationally or internationally can be monitored. A "large" scale implementation of monitoring international SS No 7 links was made during 1996. During the first years of running the S-MAN our efforts were mostly focused on the domestic market. With the development of the S-MAN system through the 1990s and new services introduced in the telecommunications market a more "product" oriented approach was adopted. Numerous number of service reports were designed to meet the increasing demand for product control of the new services including: 1) Service reports for premium services, 2) Freephone services, 3) Information/Teletorg services, 4) Public emergency services, 5) Mobile to mobile and mobile to fixed network services, 6) Maritime satellite services, 7) Directory services, etc., 8) Special international service reports.

S-MAN has been very promising for detecting irregular traffic patterns and fraudulent misuse of signalling systems and teleservices. S-MAN is able to present a very good survey and overview of all numbering areas and services offered by service providers and network operators domestically and internally, (originating or transmitting the Telenor's network). Due to number transfers in the network we are also able to observe the origin of the call, as well as the termination. Typical traffic (call) patterns pointing at possible irregularities are:

- Extremely short or long holding times towards specified services or numbers
- Certain heavy or focused traffic interest to specific numbers and/or services
- Calls at certain hours/period of the day.

The S-MAN system produces relevant traffic information for use in the daily operation (and maintenance) of the network. The use of the data from S-MAN has directly contributed to an increase in the overall quality of the services. Use of the real time applications of S-MAN enables an immediate controlled action from the central network operations centre in situations with cable failures, general network overloads, focused traffic congestions, etc. This can be measured from an increase in the call completion and thereby directly seen in the amount of the paid traffic.

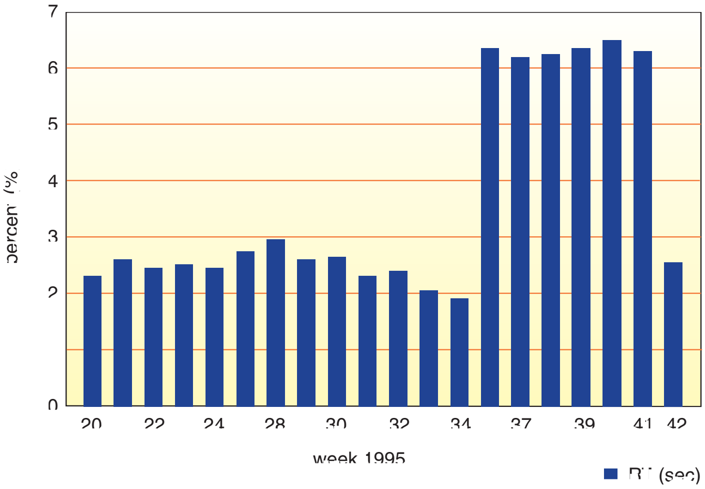


Figure 11 Weekly UNN, ADI and CLF

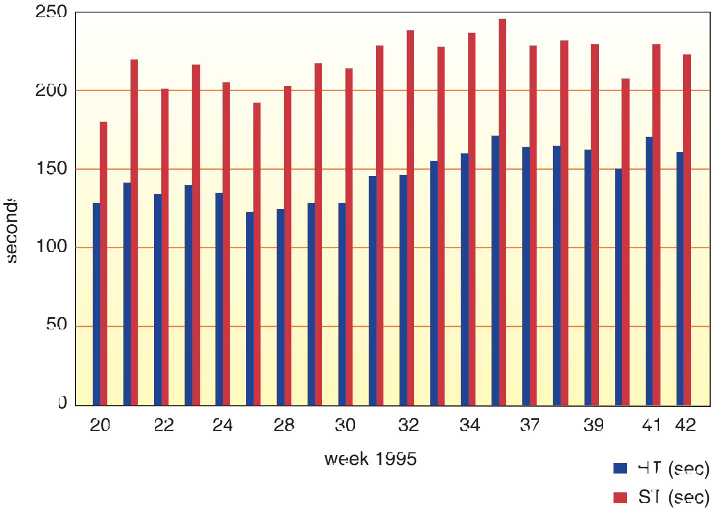


Figure 12 Weekly HT compared with ST, S-MAN/e

Today, Telenor is aiming at monitoring the performance of both national and international traffic relations, traffic streams and international destinations, covering, originating, terminating and transit traffic.

7 Summary and conclusion

Important issues to consider when deciding on the implementation and use of an intrusive "test system instrument" are:

- The measurement objective, whether it is for performance measurement or for the diagnostic capability or both.
- The use of portable units, or a complete fixed test network.
- The location of the test-units (at international centres or at the local exchange or at other points within the network)
- The parameters of importance to the customer.
- Sufficient number of test calls (call samples) collected for performance measurements to ensure that the results are statistically significant.

Measurement programmes should be carried out under controlled network conditions making it possible to compare results. If the network (equipment or traffic route) changes during measurement programmes the results may be distorted, making comparison with previous results difficult.

The use of currently available non-intrusive SS No 7 monitoring equipment gives the operator an objective evaluation of the telephone services, to a large degree in accordance with how the customer perceives the QoS received.

The SS No 7 monitoring system is a very strong tool to "identify and trace" network irregularities, e.g. irregular network access, manipulation of signalling systems, misuse of network facilities and services.

References

1 Huslende, R. *Proc. 14th International Telecommunication Symposium,* Taipei, Taiwan, 27–29 Sept. 1994.

2 Friend, T. Intrusive versus non-intrusive measurement devices. *QSDG,* Chester, UK, June 1994. (Doc. QSDG 11/94-17.)

3 Bostica, P. *Proposal for revision of rec. ITU-T E.820.* ITU/SG 2-contribution, Bethesda, April 1995.

4 Tommassen, P. *QSDG Task Group on QoS aspects of ISDN including IN.* Cape Town, April 1996. (Doc. QSDG 13/96/40.)

5 *Handbook on QoS.* Geneva, ITU, 1984.

6 Ulriksen, E. *Measuring Quality of Service.* ACECO/ETNO seminar on Telecommunication Quality. Bled, Slovenia, February 1996.

Measurement-based software process improvement

BY TORE DYBÅ AND ØYSTEIN SKOGSTAD

With increasingly more of the functionality in today's telecommunication systems implemented in software, one of the most challenging tasks for PNOs is to assess the quality of the software in these systems and the capabilities of potential suppliers. In this article we describe the use of goal-oriented measurement and assessment as the basis for systematic quality improvements in software development, maintenance and acquisition.

Introduction

The Public Network Operators (PNOs) acquire a large quantity of software every year. The role of software in telecommunications systems is growing and it has become increasingly important for a PNO to be able to assure the quality of software being acquired. Like all members of the software community, PNOs are potentially vulnerable to the general lack of attention given to software quality. Particular areas of concern are software reliability, maintainability and the quality of the suppliers' software development processes.

The IEEE defines a process as "a sequence of steps performed for a given purpose" [19]. Correspondingly, a software development process can be defined as "a set of activities, methods, practices, and transformations that people use to develop and maintain software and the associated products (e.g., project plans, design documents, code, test cases, and user manuals)" [25].

The underlying premise of software process management is that the quality of a software product is largely determined by the quality of the process used to develop and maintain it. Thus, the role of process is to tie together the people developing the software and the technology used (e.g. tools and methods), as well as the product complexity and environmental characteristics (e.g. schedule pressure) as shown in Figure 1.

To observe and quantify the impact of software process improvement, we must measure the performance of a software organization over time. Thus, measurement plays a key role for incremental improvements to the software development process.

The intent of software process improvement is:

- Improving software product quality,
- Increasing productivity, and
- Decreasing the lead time for product development.

In addition, for some organizations, improvement of the software process to meet international and contractual standards is becoming a necessity for doing business. In other words, software process improvement is a critical research and business priority aiming at achieving competitive advantage.

The Plan-Do-Check-Act (PDCA) cycle (Figure 2), developed by Dr. Walter Shewhart in the 1920's, provides the basic philosophy for a disciplined, cyclical approach to continuous improvement. PDCA is also referred to as the "scientific method" and the "Shewhart cycle". The cycle was later introduced by Edward Deming in his work with the Japanese industry after World War II.

The essential message of the cycle can be viewed in Figure 2.

While there are important differences, the ideas of quality or process improvement is just as applicable to software development as they are to manufacturing. Software process improvement is the application of these concepts to software development.

There are several on-going process improvement programs, both in national industry and research projects, and in international projects co-sponsored by the European Community's European Strategic Program for Research in Information Technologies (ESPRIT).

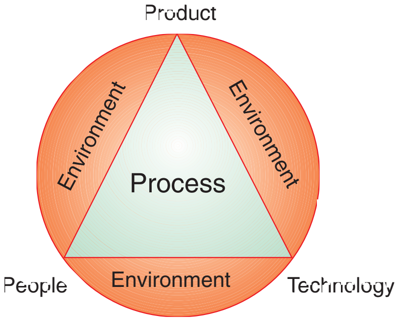


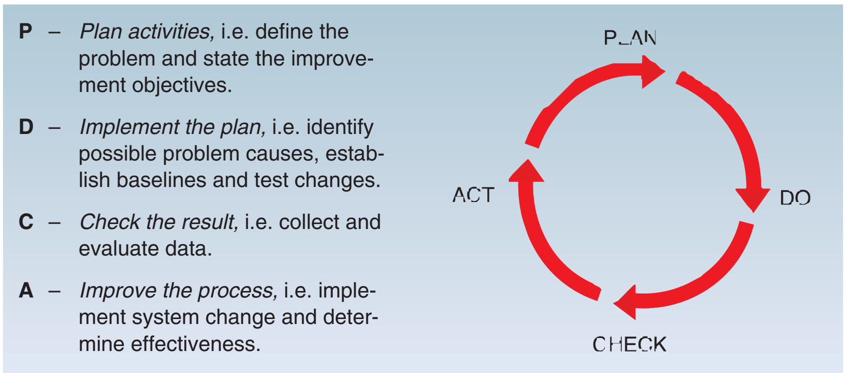
Figure 1 Process as an integrating factor

More information about these projects may be attained from the authors.

The next section outlines an overview of software measurement in general, and in particular the concepts of goal-oriented measurement and the Goal-Question-Metric approach.

Then, we give an introduction to software process assessment and capability evaluation, and a description of the Software Engineering Institute's Capability Maturity Model and the SPICE project's proposal for a new international standard on software process assessment.

Finally, the last section provides an overview of software process improvement and the principles of the Quality Improvement Paradigm and the Experience Factory.


P – *Plan activities,* i.e. define the problem and state the improvement objectives.

D – *Implement the plan,* i.e. identify possible problem causes, establish baselines and test changes.

C – *Check the result,* i.e. collect and evaluate data.

A – *Improve the process,* i.e. implement system change and determine effectiveness.

Figure 2 Shewart improvement cycle [12]

Software measurement

Measurement overview

To observe and quantify the impact of software process improvement, it is necessary to measure the performance of a software development organization over time.

In addition to the Shewhart cycle, Deming based his work with the Japanese industry on the concept of statistical process control. A process is said to be stable or under statistical control if its future performance is predictable within established statistical limits [11].

The basic principle behind statistical control is measurement. As Lord Kelvin said a century ago [13]:

> *"When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the stage of science."*

There are a number of approaches for defining measurable goals that have been described in the literature and that have been applied in practice: the Quality Function Deployment (QFD) approach [38], the Goal-Question-Metric (GQM) approach [7], [6], and the Software Quality Metrics (SQM) approach [22]. Most prominent amongst these is the GQM approach, which has been selected by many organizations because of its flexibility and adaptability to any type of organization and metrication objective.

Goal-oriented measurement

In goal-oriented measurement, the primary question is not, "What metrics should I use?" but, "What do I want to know or learn?" The goal-driven process begins with identifying business goals and breaking them into manageable subgoals. It ends with a plan for implementing well-defined measures and indicators that support the goals. Along the way, it maintains traceability back to business goals, so that those who collect and process measurement data do not lose sight of their business objective.

Goal-oriented measurement helps ensure adequacy, consistency and completeness of the measurement plan and the data collection procedures. It helps manage the complexity of the plan, and it helps stimulate a structured discussion and promote consensus about measurement and improvement goals. In turn, this helps define widely accepted measures and models within an organization, a crucial prerequisite for measurement success.

Goal-Question-Metric (GQM) is an approach for goal-oriented measurement in software projects that has proven to be a particularly effective approach to selecting and implementing measures and indicators [7], [6]. It represents a systematic approach for tailoring and integrating goals with models of the software processes, products and quality perspectives of interest, based upon the specific needs of the project and the organization.

The GQM model has three levels [3]:

1. *Conceptual level* (**goal**): A goal is defined for a variety of reasons with respect to various models of quality, from various points of view, relative to a particular environment.

   The fundamental types of measurement fall into three categories [14]:

   - *Processes:* collections of software-related activities, e.g. specifying, designing, testing, interviewing.
   - *Products:* any artefact, deliverable or document that results from a process activity, e.g. specifications, designs, programs, test suites.
   - *Resources:* entities required by a process activity, e.g. personnel, hardware, software, office space.

2. *Operational level* (**question**): A set of questions is used to characterize the way the assessment/achievement of a specific goal is going to be performed, based on some characterizing model. Questions try to characterize the object of measurement with respect to a selected quality issue, and to determine its quality from the selected viewpoint.

3. *Quantitative level* (**metric**): A set of data is associated with every question in order to answer it in a quantitative way. The data can be

   - *Objective:* If they depend only on the object that is being measured and not on the viewpoint from which they are taken; e.g. number of versions of a document, staff hours spent on a task, size of a program, delivery time.
   - *Subjective:* If they depend on both the object that is being measured and the viewpoint from which they are taken; e.g. readability of a text, level of user satisfaction.

A GQM model is a hierarchical structure starting with a goal as shown in Figure 3. The goal is subsequently refined into a set of questions, and each question is then refined into metrics. The metrics reflect the actual data needed to answer the questions. The same metric can be used in order to answer different questions under the same goal.

As an aid in the process of defining the goals, questions and metrics, the CEMP-project has provided a process model and a series of templates [15]. As an example, suppose your overall goal is to understand the causes of reliability. That is, you want to derive a quality model

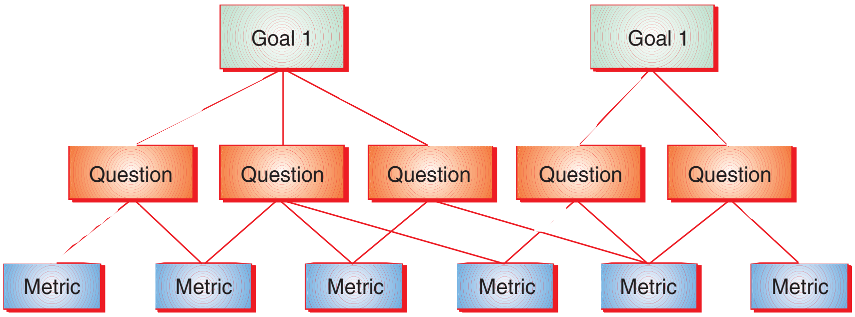


Figure 3 GQM hierarchical structure

and influencing factors for the goal “reliability”. To achieve this understanding, you must ask several key questions. First, it is important to define questions relating to the quality focus of the goal. For goal “reliability”, this could be questions relating to the number of failures and faults, and the cost of defects. Second, questions need to be asked regarding possible variation factors such as process conformance.

Once these questions are identified, you must analyze each question to determine what must be measured in order to answer the question. Likewise, you need to establish baseline hypotheses regarding the answers of the questions, and how the variation factors will impact on those baseline hypotheses.

In this way, we only generate those measures that are related to the goal. In many cases, however, several measurements may be needed to answer a single question, and a single measurement may apply to more than one question.

Parts of the resulting GQM model, i.e. questions and metrics related to the quality focus, is shown in Figure 4, which is an example of how a GQM hierarchy may be worked out. The example shows how metrics can be established for faults and failures before delivery. A *fault* occurs when a human error results in a mistake in some software product. On the other hand, a *failure* is the departure of a system from its required behaviour. The underlying hypothesis in the example being that several faults and failures found *before* delivery, implies that the software may be failure prone also after delivery.

Benefits of measurement-based process improvement

The implementation of a measurement-based process improvement program, such as GQM, will result in numerous benefits, both for software developers and purchasers.

The benefits of measurement to *software developers* improves the ability to:

- Decide whether to adopt a particular process or method, knowing the possible improvements it would bring in terms of applicability, appropriateness to the practitioners, strength, robustness, efficiency, etc.
- Decide whether to adopt a tool, knowing (or being able to estimate) the likely effect it would have on efficiency of the software development process.
- Compare the software development process and its quality characteristics against industry norms, and thereby to find more areas for improvement.
- Make trade-offs of process effort against product quality.
- Predict the outcome of developing software in a particular way with greater accuracy.

The benefits of measurement to *software purchasers* improves abilities to:

- Differentiate between suppliers on the basis of the efficiency of their software development process, or the competence of their staff.
- Choose between alternative products according to the process used to develop them.
- Require the use of a particular process from a supplier to achieve some desired effect.
- Accept product from a supplier on a quantified basis of quality.
- Improve the performance of its suppliers.

These abilities would make purchasers more discerning and discriminating, leading to better specified products, and products that better meet their specifications. In other words, they improve their procurement process. Purchasing organizations that also have software development and support operations can apply the same measurement techniques internally and gain additional benefits.

The role of measurement in software development

Measurement offers visibility into the way in which the processes, products, resources, methods, and technologies of software development relate to one another. Measurement can help answer questions about the effectiveness of techniques or tools, the productivity of devel-

Goal reliability

Analyze the development processes	[object]
for the purpose of reducing	[purpose]
the causes of unreliability	[quality focus]
from the viewpoint of the software development team	[viewpoint]
in the following context: company X, site Y, division Z.	[environment]

Question:

What is the number of failures reported before delivery?

Metrics:

Number of failure reports turned in before delivery.
Classification of failures by severity.

Question:

What is the number of faults detected before delivery?

Metrics:

Number of fault reports turned in before delivery.
Name of module the fault was isolated in.
Life cycle phase the fault was detected in.

Question:

What is the effort working on defects?

Metrics:

Effort in hours to isolate the faults that caused the failure.
Effort in hours to correct the faults that caused the failure.

Figure 4 Example GQM model

opment activities, the quality of products, and more. Measurement also allows us to define a baseline for understanding the nature and impact of proposed changes. Finally, measurement allows managers and developers to monitor the effects of activities and changes on all aspects of development.

The three key reasons for software measurement are thus to:

- *Understand* what is happening during development and maintenance, by assessing the current situation and establishing baselines that help in setting goals for future behaviour.
- *Control* what is happening in the projects. Baselines, goals and understanding of relationships is used to predict the future behaviour and make changes to processes and products that help us meet our goals.
- Guide *improvements* in software engineering processes and products.

Any of these reasons should be enough to motivate an organization to implement a measurement program. The underlying purpose of any such program, however, must be to achieve specific results from the use and application of the measurements; collecting data is not the objective. Furthermore, choosing metrics, collecting data, analyzing the results and taking appropriate action require time and resources; these activities are cost effective only if they are directed at specific improvement objectives. Without such objectives, no benefit will be derived from the measurement effort.

In order to understand when measurement is appropriate, the notion of process maturity is often helpful. The more mature the processes are, the more is visible and therefore measurable. The next section will take a closer look at some of the more popular process maturity assessment frameworks.

Software process assessments and capability evaluations

The basic concept of a maturity framework was inspired by Crosby's quality management maturity grid and its five evolutionary stages in adopting quality practices [9]. This maturity framework was adapted to software by Radice et al. at IBM [31]. Humphrey brought the maturity framework from IBM to the Software Engineering Institute (SEI) in 1986, adding the concept of maturity levels [17], [18]. After several years of experience with the initial process maturity framework, the SEI evolved the framework into the Capability Maturity Model (CMM) for software [25], [26].

A maturity model supports measurement of the software process by providing a framework for performing appraisals. Although human judgement cannot be removed from these process appraisals, the use of a maturity model provides a basis for objectivity.

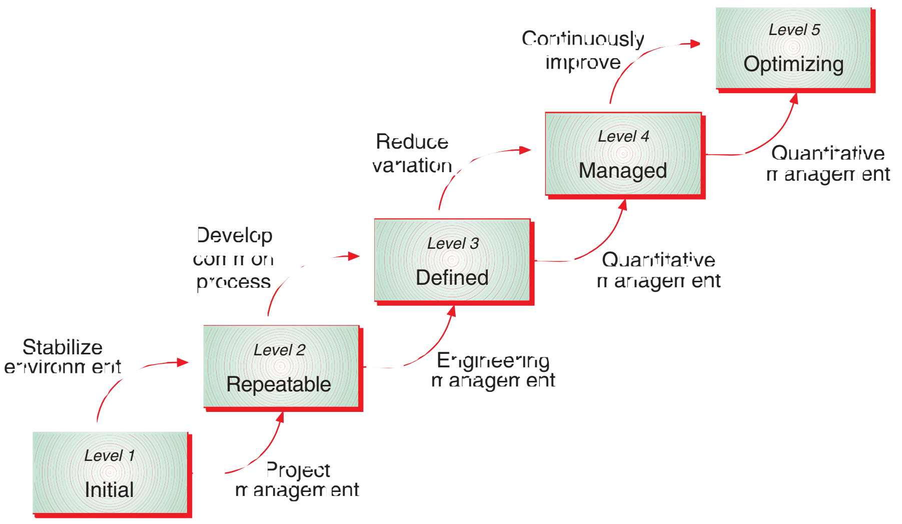


Figure 5 The five levels of the CMM

There are two general classes of appraisal [30]:

- *Software process assessments* are used to determine the state of an organization's current software process, to determine the high-priority software process-related issues facing an organization, and to obtain the organizational support for software process improvement.

  The most valuable outcomes of an assessment are: identifying the software process issues facing the organization, the buy-in to improvement, the organization-wide focus on process, and the motivation and enthusiasm in executing an action plan.
- *Software capability evaluations* are used to identify contractors who are qualified to perform the software work or to monitor the state of the software process used on an existing software effort. The evaluations are performed in an audit-oriented environment, and the emphasis is on a documented audit trail that reveals the software process actually implemented by the organization.

The capability maturity model for software

The CMM describes the principles and practices underlying software process maturity and is intended to help software organizations improve the maturity of their software processes in terms of an evolutionary path from ad hoc, chaotic processes to mature, disciplined software processes.

The CMM is organized in five levels as shown in Figure 5.

The following characterizations of the five maturity levels highlight the primary process changes made at each level [25]:

1. *Initial*
   The software process is characterized as ad hoc, and occasionally even chaotic. Few processes are defined, and success depends on individual effort.
2. *Repeatable*
   Basic project management processes

are established to track cost, schedule, and functionality. The necessary process discipline is in place to repeat earlier successes on projects with similar applications.

3. *Defined*
The software process for both management and engineering activities is documented, standardized, and integrated into a standard software process for the organization. All projects use an approved, tailored version of the organization's standard software process for developing and maintaining software.

4. *Managed*
Detailed measures of the software process and product quality are collected. Both the software process and products are quantitatively understood and controlled.

5. *Optimizing*
Continuous process improvement is enabled by quantitative feedback from the process and from piloting innovative ideas and technologies.

With the exception of Level 1, each maturity level is associated with a set of key process areas, and each process area is organized into five sections called common features. The common features contain the key practices. The key practices are supposed to provide evidence that the process area is effective, repeatable, and long-lasting [30]. An organization is said to satisfy a key process area only when the process area is both implemented and institutionalized.

The CMM's implications for measurement are clear: It emphasizes quantitative control of the process, and the higher levels direct the organization to use measurement and quantitative analysis.

A software organization operating at level 1 is likely to do very little measurement, level 1 measurements will, however, provide a baseline for comparison as an organization seek to improve its processes and products. At level 2, the organization will use a minimum set of data needed to control and manage a software project, while level 3 measures the intermediate and final products produced during development. The measurements at level 4 capture characteristics of the development process itself in order to allow control of the individual activities of the process. At level 5 the process is so mature and carefully managed that it allows measurements to provide feedback for dynamically changing the process during a particular project's development.

A GQM analysis of a project using the CMM framework can derive many of the goals for the analysis, and questions directly from the key process areas and practices. Such an analysis has been performed, and the SEI provides a technical report that discusses the metrics implications of the CMM [5].

The CMM has popularized the notion of measuring the software process maturity of organizations. Based on the CMM and other process assessment models, such as Trillium [37], Bootstrap [21] and many others, the International Organization for Standardization (ISO) is developing a suite of standards on software process assessment under the rubric of SPICE [35], in an attempt to harmonize existing approaches.

SPICE

One of the objectives of the ISO effort is to create a way of measuring process capability, while avoiding a specific approach to improvement such as the SEI's maturity levels. ISO selected an approach to measure the implementation and institutionalization of specific processes; a process measurement rather than an organization measurement. Using this approach, maturity levels can be viewed as sets of process profiles.

The SPICE framework is built on an assessment architecture that defines desirable practices and processes. Two different types of practices are distinguished:

- *Base practices,* which are essential activities of a specific process.
- *Generic practices,* which implement or institutionalize a process in a general way.

The architecture thus defines a two-dimensional view of process capability, as shown in Figure 6.

The left-hand side of Figure 6 represents the functional, base practices, involved in software development and management. This functional view addresses five general areas of activity:

1. *Customer–supplier*
This process category consists of processes that affect the customer directly, support development and delivery of the products to the customer, and ensure correct operation and use.

2. *Engineering*
This process category consists of processes that specify, implement or maintain the system and its documentation.

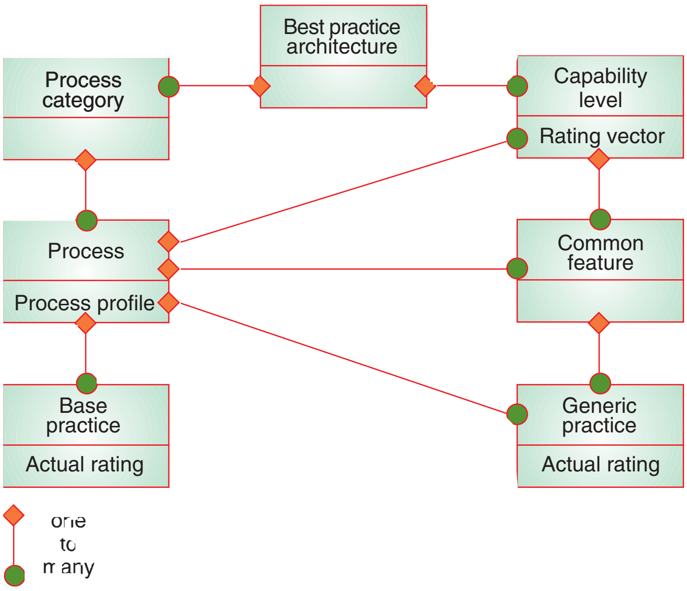


Figure 6 SPICE architecture for process assessment [32]

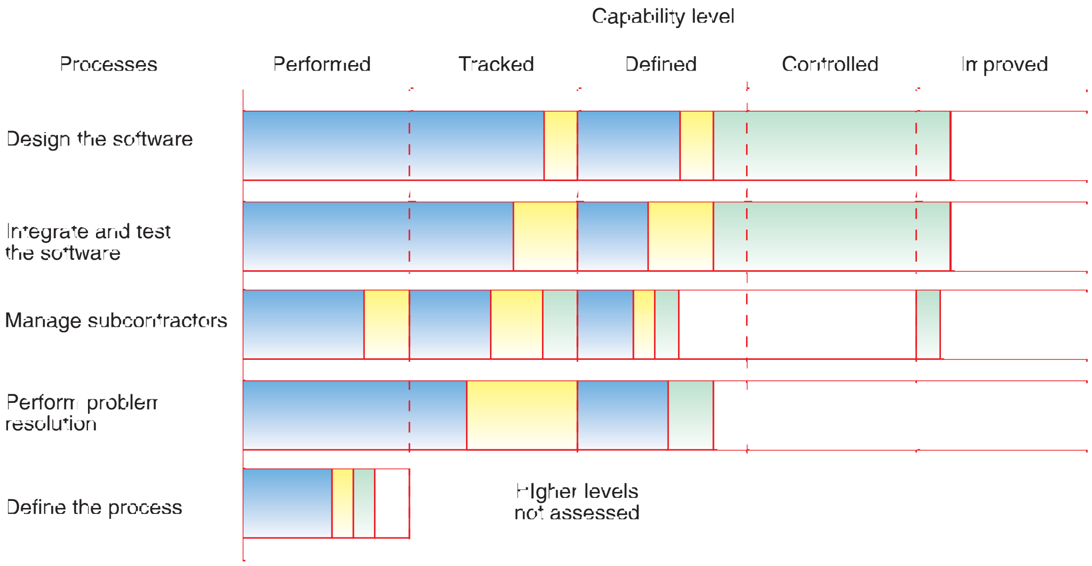


Figure 7 SPICE assessment profile [32]

3. *Project*
   This process category consists of processes that establish the project, coordinate or manage resources, or provide customer services.
4. *Support*
   This process category consists of processes that enable or support performance of the other processes.
5. *Organization*
   This process category consists of processes that establish business goals, and develop assets to achieve those goals.

The right-hand side of Figure 6 represents the management, generic practices, applicable to all processes. These practices are arranged into six capability levels:

0. *Not performed*
   There is a general failure to perform, and there are no identifiable workproducts.
1. *Performed informally*
   Performance is not planned and tracked, it depends on individual knowledge and effort, but workproducts are identifiable.
2. *Planned and tracked*
   Performance is verified according to specified procedures, and the workproducts conform to specified standards and requirements.
3. *Well-defined*
   Performance according to a well-defined process using approved, tailored versions of standard, documented processes.
4. *Quantitatively controlled*
   Detailed measures of performance are used for predictions, objective management, and quantitative evaluation of workproducts.
5. *Continuously improving*
   Quantitative process effectiveness and efficiency targets are established based on business goals, quantitative feedback from defined processes, and from piloting innovative ideas and technologies.

An assessment report is a profile where each process area is evaluated and reported to be at one of the six capability levels. Figure 7 shows an example of how the profile is reported. The shading indicates the degree to which the activities were satisfied at each level. Thus, whereas the CMM addresses organizations, SPICE addresses processes.

Of special interest to the telecom community is the Trillium model [37] which is a telecommunications product development and support capability model.

TRILLIUM

The Trillium model is used in a variety of ways:

- To benchmark an organization’s product development and support process capability against best practices in the industry,
- In self-assessment mode, to help identify opportunities for improvement within a product development organization, and
- In pre-contractual negotiations, to assist in selecting a supplier.

Product in the Trillium context refers to what the customers receive, use and perceive. For an embedded telecommunications product, this would typically include hardware, software, documentation, training and support services.

The Trillium model is based on the CMM v1.1. In addition, it also incorporates the intent of ISO 9001 (and its companion guide 9000-3), the Malcolm Baldrige examination criteria, Bellcore’s TR-179 and IPQM TA-1315[1], the IEEE Software Engineering standards and IEC 300.

The reliability of measuring maturity

The maturity models and their assessment methods are becoming de facto standards in many organizations. For example, minimum CMM scores are expected to be required for some US Air Force software contracts [34], and the scores have a significant effect on US Navy contract decisions [33]. But ever since the introduction of the SEI’s original process maturity model, there have been objections to its applications and use.

As the maturity models help in identifying strong and weak points, organizations make major business and technical deci-

[1] *Since the release of Trillium, the Bellcore document is upgraded to GR-1315.*

sions based on the assessment results. However, if the models and measurements are incorrect or misguided, the result can be misallocation of resources, loss of business, and more. Inconsistent results have been reported from CMM assessments of the same organization by different teams, thus raising the issue of reliability of assessments shortly after the CMM was introduced [8]. Here, reliability refers to the extent to which the same measurement procedure yields the same results on repeated trials.

Thus, there are serious measurement questions to be addressed in considering the use of these process and organizational frameworks for software process assessment and capability evaluation. When using such frameworks, organizations must ensure that the models are appropriate (e.g. for understanding, control or improvement), and that they understand how reliable and valid the measurements and models are. Organizations must know what entities and what attributes of those entities that are being measured, and they must test the relationships between the maturity scores and the behaviours that "maturity" is supposed to produce or enhance.

Combining measurement with process maturity

All process assessment models share a common goal and approach of process visibility as a key discriminator among a set of maturity levels. That is, the more visibility into the overall development process, the higher the maturity and the better managers and developers can understand, control and improve their development and maintenance activities.

Software engineers have detailed insight into the state of a project because they have first-hand information on project status and performance. However, in large projects, their insight is usually drawn only from their personal experience in their area of responsibility. Those outside the project without first-hand exposure, such as senior managers, lack visibility into the project's processes and therefore rely on periodic reviews for the information they require to monitor progress.

Each succeeding maturity level incrementally provides better visibility into the software process [30]. At the lowest levels of maturity, the process is not clear and repetitive. As maturity increases, however, the process is better understood and better defined. At each maturity level, measurement and visibility are closely related: a developer can measure only what is visible in the process, and measurement helps to increase visibility [14].

Using goals to suggest a measurement program has been successful in many organizations and is well-documented in the literature [16], [28]. The GQM approach, together with process maturity models, has also been widely used. The ami (application of metrics in industry) approach is a good example of a combination of GQM and CMM [29]. Furthermore, the ami approach is recommended by the EURESCOM project P227 [23][2].

A frequently asked question by organizations using models like the CMM, Trillium and SPICE to assess their process maturity is, however: "What should I do, after the assessment, to start an improvement program and what activities will the program entail?" [27]

Measurement, like assessment, does not create improvement. It just makes it possible and supports it. Hence, the next section describes how the combination of a process maturity framework and a comprehensive measurement program can be used as the basis for software process improvement.

Software process improvement

Process improvement overview

World-wide, there are various approaches to implementing software process improvement. Most continuous improvement approaches, however, are based on the PDCA cycle. There is a general consensus regarding the need for software quality improvement, and that measurable process improvement is possible. However, there are different opinions on specific issues such as the best way to proceed.

There are two basic approaches to process improvement [36]. The top-down approach compares an organization's process with some generally accepted standard process (best practices). Process improvement is then the elimination of differences between an existing process and a standard one. The assumption is that, once the process is changed the generated products will be improved – or at least the quality risks of generating new software will be reduced. Examples of such top-down approaches are the already presented assessment models CMM, Trillium, SPICE and others.

The bottom-up approach assumes that process change must be driven by an organization's goals, characteristics, product attributes, and experiences. Change is defined by a local domain instead of a universal set of accepted practices. For example, an organization whose primary goal is improving time to market may take a significantly different approach to process change than one whose primary goal is to produce defect-free software. Examples of such bottom-up approaches are the Quality Improvement Paradigm [2] and the IDEAL model [27].

This section will take a closer look at the Quality Improvement Paradigm which offers a framework based on an evolutionary quality management paradigm tailored for software business, and the Experience Factory [1] which is an organizational approach for building software competencies and supplying them to projects. Finally, we take a look at Crosby's approach to quantifying the benefits of improvements [10].

Quality Improvement Paradigm

The Quality Improvement Paradigm (QIP) developed by Basili et al. [2], is the result of the application of the Shewhart cycle to the problem of software quality improvement. It consists of the following six steps [4], as shown in Figure 6:

1. *Characterize*
   Understand the environment based upon available models, data, intuition, etc. Establish baselines with the existing business processes in the organization and characterize their criticality.
2. *Set Goals*
   On the basis of the initial characterization and of the capabilities that have a

[2] *These recommendations are now being evaluated and updated in EURESCOM project P 619 [24].*

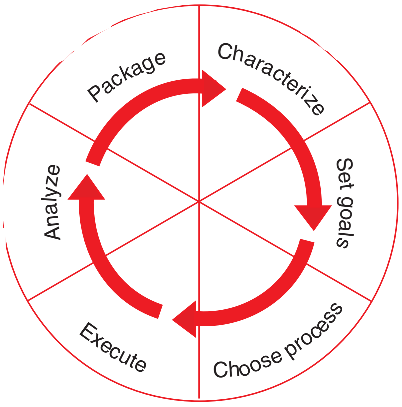


Figure 8 The Quality Improvement Paradigm

strategic relevance to the organization, set quantifiable goals for successful project and organization performance and improvement. The reasonable expectations are defined based upon the baseline provided by the characterization.

3. *Choose Process*
   On the basis of the characterization of the environment and of the goals that have been set, choose the appropriate processes for improvement, and supporting methods and tools, making sure that they are consistent with the goals that have been set.

4. *Execute*
   Perform the processes constructing the products and providing project feedback based upon the data on goal achievement that are being collected.

5. *Analyze*
   At the end of each specific project, analyze the data and the information gathered to evaluate the current practices, determine problems, record findings, and make recommendations for future project improvements.

6. *Package*
   Consolidate the experience gained in the form of new, or updated and refined, models and other forms of structured knowledge gained from this and prior projects, and store it in an experience base so it is available for future projects.

Within an overall improvement framework such as the QIP, a maturity model like the CMM may be used as the basis for diagnosing and characterizing an organization's software processes, establish priorities, and acting on them.

The Quality Improvement Paradigm implements two feedback cycles [4]:

- The project feedback cycle (control cycle) is the feedback that is provided to the project during the execution phase.
- The corporate feedback cycle (capitalization cycle) is the feedback that is provided to the organization and has the double purpose of:
  - Providing analytical information about project performance at project completion time by comparing the project data with the nominal range in the organization and analyzing concordance and discrepancy.
  - Accumulating reusable experience in the form of software artefacts that are applicable to other projects and are, in general, improved based on the performed analysis.

The GQM approach, as discussed in the previous section, is the mechanism used by the Quality Improvement Paradigm for defining and evaluating a set of operational goals using measurement.

In addition to the GQM approach, the Quality Improvement Paradigm uses the Experience Factory organization for building software competencies and supplying them to projects.

Experience Factory

The concept of the Experience Factory [1] was introduced to institutionalize the collective learning of the organization that is at the root of continual improvement and competitive advantage. The Experience Factory is an organizational structure that supports the activities specified in the QIP by continuously accumulating evaluated experiences, building a repository of integrated experience models that projects can access and modify to meet their needs.

The project organization offers to the Experience Factory their products, the plans used in their development, and the data gathered during development and operation [2], as shown in Figure 7. The Experience Factory transforms these objects into reusable units and supplies them to the project organizations, together with specific support that includes monitoring and consulting [2], as shown in Figure 8.

The Experience Factory can be a logical and/or physical organization, but it is important that its activities are separated and made independent from those of the project organization.

The Experience Factory can package all kinds of experience [4]. It can build resource models and baselines, change and defect models and baselines, product models and baselines, process definitions and models, methods and technique models and evaluations, products and product models, a library of lessons learned, and a variety of quality models.

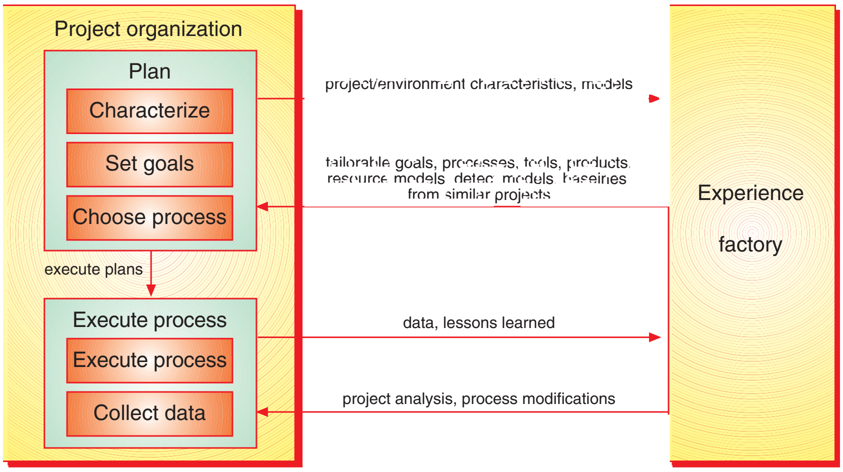

Project organization
Plan
Characterize
Set goals
Choose process
execute plans
Execute process
Execute process
Collect data
project/environment characteristics, models
tailorable goals, processes, tools, products, resource models, defect models, baselines from similar projects
data, lessons learned
project analysis, process modifications
Experience
factory

Figure 9 Project organization functions

The main product of the Experience Factory is the experience package [4]. The content and the structure of an experience package vary based upon the kind of experience clustered in the package. There is, generally, a central element that determines what the package is: a software life-cycle product or process, a mathematical relationship, an empirical or theoretical model, a data base, etc. Examples of experience packages are:

- Product packages (e.g. programs, architectures, designs)
- Process packages (e.g. process models, methods)
- Relationship packages (e.g. cost and defect models, resource models)
- Tool packages (e.g. static analyzer, regression tester)
- Management packages (e.g. management handbooks, decision support models)
- Data packages (e.g. project databases, quality records).

Thus, the Experience Factory aids in the formalization of both management and development processes, and it forces research to focus on corporate needs and technology transfer.

Calculating savings

Philip Crosby's approach [10] can be of valuable help in quantifying the benefits of software process improvements. Crosby's approach differentiates the cost of doing it right the first time from the cost of rework, and categorizes the costs associated with any process as:

- *Performance.* The costs associated with doing it right the first time, including elements such as developing the design and generating the code
- *Appraisal.* The costs associated with testing the product to determine if it is faulty
- *Rework (non-conformance).* The costs associated with fixing defects in the code or design
- *Prevention.* The costs incurred in attempting to prevent the fault from getting into the product.

The sum of appraisal, rework, and prevention costs is what Crosby calls "the cost of quality." The total project cost is, thus, simply the cost of quality plus the performance costs.

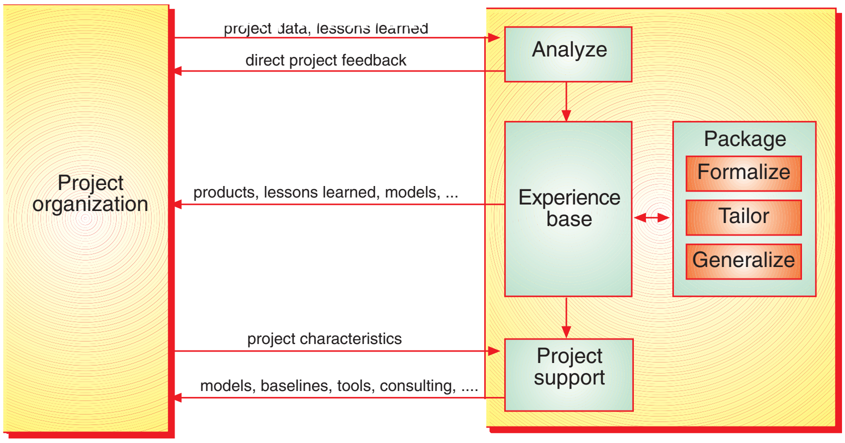


Figure 10 Experience Factory functions

Conclusions

In this article, we have presented two basic approaches to software process improvement; the top-down approach using models like the CMM and SPICE, and the bottom-up approach, which assumes that process change must be driven by the organization's goals, characteristics, product attributes, and experiences. Both approaches, however, need to be supported by measurements in order to quantify the improvements. Setting up a successful measurement program for process improvement is in other words a necessity, but it is a challenging undertaking.

For this reason, goal-oriented measurement combined with explicit modelling (e.g. process, quality, etc.) can greatly help structure and provide rigor to the measurement plan. Furthermore, the integration of the Quality Improvement Paradigm with the Goal-Question-Metric approach and the Experience Factory, provides a comprehensive framework for software engineering development, maintenance and research, that takes advantage of the experimental nature of software engineering.

In summary, we would like to conclude with Robert B. Grady's analogy to emphasize the importance of measurement [16]:

> *"A software project is like a train in a tunnel rushing toward a light. With knowledge of its speed and schedule, knowing that it is on the right track, we can be reasonably assured that it's daylight we see at the end of the tunnel. Without such quantifiable facts, it is just as likely that the light we see is another train (or disaster) rushing toward us on the same track."*

References

1 Basili, V R. Software development : a paradigm for the future. *Proceedings of the 13th Annual International Computer Software & Applications Conference (COMPSAC),* Keynote Address, Orlando, FL, September, 1989.

2 Basili, V R et al. The software engineering laboratory : an operational software experience factory. In: *Proceedings of the 14th International Conference on Software Engineering,* 1992.

3 Basili, V R, Caldiera, G, Rombach, H D. Measurement. In: Marciniak, J J (ed.). *Encyclopedia of software engineering,* vol 1, 528–532. John Wiley, 1994.

4 Basili, V R, Caldiera, G, Rombach, H D. Experience factory. In: Marciniak, J J (ed.). *Encyclopedia of software engineering,* vol 1, pages 528–532. John Wiley, 1994.

5 Baumert, J H, McWhinney, M S. *Software measures and the capability maturity model.* Software Engineering Institute, CMU/SEI-92-TR-25, 1992.

6 Basili, V R, Rombach, H D. The TAME project : towards improvement-oriented software environments. *IEEE Transactions on Software Engineering,* SE-14(6), 758–73, 1988.

7 Basili, V R, Weiss, D. A methodology for collecting valid software engineering data. *IEEE Transactions on Software Engineering,* SE-10(6), 728–38, 1984.

8 Card, D N. Capability evaluations rated highly variable. *IEEE Software,* September, 1992.

9 Crosby, P B. *Quality is free.* New York, McGraw-Hill, 1979.

10 Crosby, P B. *Quality without tears.* New York, McGraw-Hill, 1984.

11 Deming, W E. *Quality, productivity, and competitive position.* Camebridge, Mass., MIT Center for Advanced Engineering Study, 1982.

12 Deming, W E. *Out of the Crisis.* Camebridge, Mass., MIT Center for Advanced Engineering Study, 1986.

13 Dunham, J R, Kruesi, E. The measurement task area. *IEEE Computer,* November, 1983.

14 Fenton, N E, Pfleeger, S H. *Software metrics : a rigorous and practical approach.* International Thomson Computer Press, 1996.

15 Gresse, C, Hoisl, B. Wüst, J. *A process model for GQM-based measurement.* STTI-95-04-E, Software Technology Transfer Initiative Kaiserslautern, Germany, 1995.

16 Grady, R B. *Practical software metrics for project management and process improvement.* Englewood Cliffs, N.J., Prentice-Hall, 1992.

17 Humphrey, W S. Characterizing the software process. *IEEE Software,* March, 1988.

18 Humphrey, W S. *Managing the software process.* Reading, Mass., Addison-Wesley, 1989.

19 IEEE Std 610.12-1990. *Glossary of software engineering terminology.*

20 ISO 9000-1. *Quality management and quality assurance standards, part 1 : guidelines for selection and use.* Geneva, 1994.

21 Kuvaja, P et al. *Software process assessment and improvement : the bootstrap approach.* Blackwell Business, 1994.

22 McCall, J A, Richards, P K, Walters, G F. *Factors in software quality.* Rome Air Development Center, RADC TR-77-369, 1977.

23 Eurescom Project P227. *Quality assurance of software.*

24 Eurescom Project P619. *PNO : suppliers technical interfaces.*

25 Paulk, M C et al. *Capability maturity model for software, version 1.1.* Software Engineering Institute, CMU/SEI-93-TR-24, 1993

26 Paulk, M C et al. *Key practices of the capability maturity model, version 1.1.* Software Engineering Institute, CMU/SEI-93-TR-25, 1993.

27 Peterson, B. Software engineering institute. *Software Process : Improvement and Practice,* 1(1), August, 1995.

28 Pfleeger, S L. Lessons learned in building a corporate metrics program. *IEEE Software,* May, 1993.

29 Pulford, K, Kuntzmann-Combelles, A, Shirlaw, S. *A quantitative approach to software management : the ami handbook.* Addison-Wesley, 1996.

30 Paulk, M C et al. *The capability maturity model : guidelines for improving the software process.* Addison-Wesley, 1995.

31 Radice, R A et al. A programming process study. *IBM Systems Journal,* 24, (2), 1985.

32 Rout, T P. SPICE : a framework for software process assessment. *Software Process : Improvement and Practice,* 1(1), August, 1995.

33 Rugg, D. Using a capability evaluation to select a contractor. *IEEE Software,* July, 1993.

34 Saiedian, H, Kuzara, R. SEI capability maturity model's impact on contractors. *IEEE Computer,* January, 1995.

35 The SPICE Project (ISO/IEC JTC1/SC7/WG10). Software process assessment, parts 1–9. *Consolidated Product : Version 1.00,* June 1995.

36 Thomas, M, McGarry, F. Top-down vs. bottom-up process improvement. *IEEE Software,* July, 1994.

37 *Trillium : model for telecom product development and support process capability, release 3.0,* Bell Canada, 1994.

38 Uchimaru, K, Okamoto, S, Kurahara, B. *TQM for technical groups : total quality principles for product development.* Productivity Press, 1993.

Tool based in-process software Quality Analysis

BY SYED R. ALI

This article describes an innovative approach for controlling the software development process and software product quality through industry standard software metrics and a state-of-the-art metrics tool.

Introduction

After extensive research, Bellcore has developed a set of software metrics called IPQM™ ℠ that spans the entire software development life cycle, covering requirements, design, code, test and deployment. This advanced methodology is independent of any particular software development model and is prominent in the telecommunications industry as the de facto standard for software metrics. It is also under consideration in Europe as an international standard for software metrics.

This is the first time Bellcore, a world class leader in software development is sharing and marketing its software metrics tool that is based on "best in class", extensive research and industry buy-ins. IPQM generic requirements are now being implemented by leaders in the software industry.

Bellcore developed the Generic Requirement for software development organizations to measure and control software quality during software development.

The process

The IPQM process focuses on three major software measurement processes:

1. **Project Management** – Process requirements and metrics used by project managers to plan, monitor, and control the software development of individual projects or releases.
2. **Process Effectiveness** – Process requirements and metrics used to assess the effectiveness and adequacy of the defined software development life cycle processes.
3. **Process Compliance** – Process requirements and metrics used to identify areas in the software development cycle where defined processes are not being followed.

Each of the above processes is part of the development cycle of a software product. These processes are measured in each of the life cycle phases of software development:

- Requirements
- Design
- Coding
- Integration test
- System test
- Deployment.

IPQM is a Trademark and Service Mark of Bellcore

Table 1 Project Management Metrics

Key Metrics	Function	Software Life Cycle					
		Requirements	Design	Coding	Integration Test	System Test	Deployment
Software System Size	Tracks Software resources. Major component of cost and schedule estimation models	X	X				X
Personal Resources	Assess the effectiveness of staffing estimation & forecasting mechanism	X	X	X	X	X	
Project Planning	Tracks and measures the progress of tasks and overall project toward the completion of schedules	X	X	X	X	X	
Milestone Delay	Tracks and controls milestone slips pro-actively	X	X	X	X	X	X
Phase Transition	Defines and tracks life cycle phase overlaps. Controls deliverables that may impact product delivery	X	X	X	X	X	X
Requirements & Design Stability	Monitors changes to requirements and design after associated baselines have established	X	X				
Test Tracking	Assess the progress of internal test planning test activities. Tracks testing activities during the integration, regression and system test life cycle phases			X	X		

IPQM is instrumental to help software developers establish a measurement framework to collect, analyze, and use software development data to meet the objectives of the customer or software product they are developing. IPQM helps to guide developers through the following framework for their development efforts:

1. An established and manageable software development process that meets the requirements and objectives of process maturity.
2. An established measurement of key software development and quality indices.
3. Models of historical project and process performance that can be used to establish standards, objectives, performance profiles and relationships to "downstream" quality objectives.
4. In-process metrics to depict current performance and to compare with profiles to identify deviations and initiate corrective actions.
5. Defect models to identify opportunities to modify software development processes necessary to improve quality.

Key Metrics are defined by the user and consist of three basic elements:

- **Software Metrics** – Plan for implementation based on internal and/or customer requirements.
- **Threshold** – An objective set for each key metric for use as a control criteria and used when process stability has been achieved.
- **Action Plan** – Plan used when a threshold criteria has been exceeded.

Each metric needs to be implemented to provide a degree of process stability and measurement methodology which are necessary conditions to collect data, compute metrics, and plan action based on these metrics. It is the successful completion of these element metrics in each life cycle phase that will result in the release of a software product for deployment. Each of the software measurements are further described below.

Project Management

Projects planned in advance will help to ensure the timely delivery of software releases resulting in increased customer satisfaction. Projects are divided into smaller manageable entities for better control which require identification, responsibility assignments, and established milestones. Note: Each key metric applies to specific life cycle phases (i.e. Software System Size ... Requirements, Design, and Deployment only). IPQM uses seven key metrics for project management. See Table 1 for details.

Process Effectiveness

Assesses the effectiveness and measures control of the software quality plan. Once again, each key metric applies to specific life cycle phases (i.e. Requirements

Table 2 Process Effectiveness Metrics

Key Metrics	Function	Software Life Cycle					
		Requirements	Design	Coding	Integration Test	System Test	Deployment
Requirements Traceability	Measures and controls requirements and design traceability to system test. It ensures completeness of requirements implementation during design & system test	X	X			X	
Defect Removal and Tracking	Tracks and controls software defects during system test and deployment					X	X
Defect Detection Effectiveness and Profiling	Assess software developing organization's defect prevention and detection activities for all life cycle phases. Aids in Continuous Improvement Program (CIP)	X	X	X	X	X	X
Defect Prediction and Estimation	Establishes a software developing organization's maturity in predicting the number of software defects in advance and taking actions to contain it	X	X	X	X	X	X

Traceability ... Requirements, Design and System Test only). IPQM uses four key metrics for measuring process effectiveness. See Table 2 for details.

Process Compliance

Measures the compliance of scheduled audits. Audits determine where corrective actions may be needed. Auditing results have the greatest potential for far-reaching quality improvement benefits. When audit results are classified, it will further provide the supplier with feedback on previously implemented corrective actions to verify that the process improvements reduce or eliminate the deficiencies across development groups.

The Tool

The IPQM Tool is a quality metrics tool (incorporates Bellcore's GR-1315 IPQM Generic Requirements) that supports improvements in project management, process management, and process compliance in the software development life cycle process. In the early stages of development, it assists and alerts users of process problems and provides them with better control of project schedules and software quality. It can also help to identify, remove, and lower defects in the software prior to release. In addition, it can help the user anticipate defects and propose corrective actions before they occur. Users have the flexibility to use the tool's default metrics and thresholds, or users can define their own metrics and thresholds along with IPQM. When a metric does not meet the threshold, the IPQM Tool keeps track of all out of range metrics and support the implementation of action plans. The tool also generates metric reports for different levels of management and line organizations as well as graphs and charts for all implemented metrics. IPQM has a Graphical User Interface and on-line help to support ease-of-use.

Features

- Automatically incorporates In-Process Quality Metrics
- Supports automatic generation of metrics report for different levels of management and line organizations
- Allows incorporation of user defined metrics along with In-Process Quality Metrics
- Allows planning groups to establish thresholds for all metrics dynamically with user defined names and ranges
- Keeps track of all out of range metrics and supports implementation of action plans
- Provides Graphs and Charts for all implemented metrics
- Provides an intuitive Graphical User Interface (GUI) and on-line help that makes it easy to use.

Product Platform

The IPQM Tool:

- Unix based supports
  - Sun Sparc
  - IBM RS/6000
  - DEC Alpha
  - DEC VAX
  - HP 9000/700
- NT and Windows 95 (available 2Q 1997)
- Uses Fourth Generation Language (4GL) for template support.

Who will benefit

The IPQM Tool is intended to benefit both the customers and suppliers of software products. In particular, Project Managers, Senior Management, Software Procurement, Quality Assurance, Software Development and Software Maintenance Organizations.

Benefits – intended to:

- Improve software engineering practice by establishing an uniform software metrics requirement throughout an organization
- Identify software process problems earlier in the life cycle
- Provide better control of project schedules and software quality
- Provide automatic implementation of industry and internal standards
- Reduce cost for the development of new tools
- Aid in continuous improvement program
- Support higher levels of Capability Maturity Model (SEI's CMM)
- Improve customer satisfaction
- Support baselining software process
- Support benchmarking activity
- Is a low cost tool based on industry standards.

EIRUS – a user group for quality measurements

BY ROLF JOHANSSON

Background

Quality measurements and how to measure in-service performance has always been a big issue for telecom operators. Different measurement systems have been used among the operators in Europe, and it has not always been easy to communicate between operators in the matter of quality aspects and what to monitor in order to get a good and real picture of how the different systems are performing.

A task under a EURESCOM project started in December 1993. The aim of this task was to find out what was important to measure and also if there were any existing metrics that could be used among the operators in Europe. Questionnaires were performed in order to get information into the project on what the operators considered was of both interest and importance.

After the work with the questionnaires was finished, the task group looked into existing quality systems world-wide. They found that a system which would very much fulfil the requirements of Europe had been invented by Bellcore in the USA. A relation between the task group and Bellcore was established and some rules on how the task could use the Bellcore documentation were formed.

After that, the task group organised a seminar in Heidelberg in March 1995. To that seminar both operators, suppliers and of course Bellcore were invited and asked to give feedback. The outcome of the seminar was that the operators were asked to establish a user group for this measurement system. At the seminar the task group suggested an implementation in two steps, but the suppliers advised that the implementation should have a pilot phase so that a more general understanding of this system could be obtained.

Set-up meeting

The first meeting was held in Stockholm 20–21 June 1995 and was hosted by Telia. The purpose of the meeting was to form the first draft rules and procedures of this new user group. Discussions among the users initiated a common understanding of E-IPQM and E-RQMS.

Now, what is E-IPQM and E-RQMS? The measurement system that the operators had chosen was IPQM (In Process Quality Metrics) and RQMS (Reliability and Quality Measurements for Telecommunication Systems). The group also formed rules for membership and stated the number of ordinary meetings per year and rules for meetings in-between.

After voting among the existing members the logo type for this user group was decided. The name EIRUS is an acronym for:

European In Process Quality Metrics
Reliability and Quality Measurements
for Telecommunication Systems
Users

This means that the group will use Bellcore standards modified for European PNOs (Public Network Operators).

The main objectives of EIRUS are:

- Implementation of E-IPQM and E-RQMS
- Contact forum for PNOs and suppliers
- Retain uniformity of measurements
- Establish a common attitude towards the suppliers
- Establish and retain interaction with Bellcore.

The second meeting was held in Norway and was hosted by Telenor Research and Development. During this meeting the final version of rules and procedures was formed and rules for documentation within EIRUS were decided.

The users also produced a document for upper management to justify the need for participating in EIRUS.

The second meeting also produced a communication plan. EIRUS needs to have contact with the following organisations and projects, or is considering this:

- The user group for several telecommunication systems
- EURESCOM as such
- EURESCOM project P619 (1996–1997 retain and develop the metrics)
- Bellcore
- The European Union
- The European Foundation for Quality Management (EFQM).

For the system user groups, one EIRUS contact person per user group will be needed on a voluntary basis. In principle, a formal contact will be made at the next meeting of a group, with a presentation of EIRUS by the contact person for the user group. The following user groups have already been identified:

- AXE User Forum (Ericsson)
- S12 User Club (Alcatel)
- EWSD User Club (Siemens)
- 5ESS User Conference (AT&T)
- Nokia.

In the future, also transmission, broadband and mobile systems should be considered, and contacts with other system user groups may be established if an EIRUS member volunteers to be the contact person for a system or a group.

EIRUS will also work actively on promoting itself and the following will be done to this aim:

- A newsletter, EIRUS News, will be issued from end-1996
- EIRUS seminars will be held from 1997
- EIRUS has an homepage on the Internet under the EURESCOM server
- Publications in the EURESCOM newsletters
- Standard information package for potential members.

During the second meeting members of the user group also started with presentations in order to share experiences and plans regarding the implementation of these measurements. Finally, venues for the next three meetings were decided, and that concluded the second meeting in this user group.

EIRUS meetings 1995 – 1996

- Nuremberg, 22–24 November 1995
- Leidschendam, 24–26 April 1996
- Rovaniemi, 17–18 June 1996
- Athens, 15–17 October 1996.

So far, the meetings could only be attended by the PNOs, but as from the third meeting the suppliers were invited. The third meeting was held in Nuremberg and was hosted by Deutsche Telekom. The dates for this meeting were 22–24 November 1995.

During the first day only PNOs were present due to the Rules and Procedures of EIRUS. Items on the agenda were different reports from meetings with Suppliers, Experiences from Pre-study Phases on implementation.

The chairman gave a report from the seminar in Heidelberg and said that there was a positive response from the EURESCOM staff and that EIRUS could count on support regarding future work with EURESCOM.

On the second day the suppliers participated in the meeting and there were a lot of discussions regarding the use of "all these metrics". The message from the suppliers was at first that this will cost you a lot of money. Later on in discussions between different PNOs and their suppliers most of this has been solved. The general response for Phase 0 was positive and constructive. They agreed with the work done in the EIRUS group and they also welcomed the idea of uniformity in the quality measurements. There were some doubts about the possibility of having this uniformity on all metrics, but as time goes by, I think all this will be solved.

The milestones defined for the practical implementation of Phase 0 are shown in Table 1.

Figure 1 is an example of the "Patches" measurement plot in RQMS.

A review of the day with the suppliers was done on the third day of this meeting, and the venue for the next meeting was agreed. Since the next meeting was to be held in April the following year, most of the members would hopefully have gained experiences from the implementation of E-IPQM and E-RQMS.

The EIRUS group met again in Leidschendam on 24–26 April 1996. Bellcore also attended this meeting, and a delegation from EIRUS gave a presentation on contacts with Bellcore made in the USA when a group of people from Bellcore met with a delegation from EIRUS.

Table 1 Process for implementation of Phase 0

Milestone	Date	Description
M1	23-11-1995	Common agreement between EIRUS and the suppliers
M2	15-12-1995	Implementation plan Phase 0 for projects agreed between each individual PNO and the supplier (as many as possible)
M3	15-01-1996	Supplier and PNO are ready to start Phase 0 (people informed, methods and tools available, ...)
M4	01-02-1996	Start monthly metric reporting of Phase 0 to the PNO
M5	25-04-1996	Analysis of EIRUS ready and feedback from EIRUS to suppliers Set program for Phase 1 with suppliers
M6	xx-06-1996	Evaluation of Phase 0 End of Phase 0
M7	01-07-1996	Start of Phase 1

Report on contacts with Bellcore

In order to strengthen the relation between EIRUS and Bellcore and also to acquire more knowledge and experience on RQMS and IPQM, a group of people from EIRUS and EURESCOM Project P619 visited Bellcore, Bell Atlantic and Ericsson in the USA.

Some relevant points for EIRUS and the EIRUS members are:

- Due to the positive results for quality (trends) regarding telecommunication software in the USA after obtaining management commitment for the implementation of RQMS, EIRUS is strongly advised to continue the implementation of IPQM and RQMS in its processes and organisation.
- The use of IPQM and RQMS ought to become a requirement for EIRUS' suppliers through contracts.
- EIRUS must advise its suppliers to seek assistance (process definitions, tools, etc.) from their US counterparts for the implementation of IPQM and RQMS.
- The Bellcore ideas on RQMS are now daily practice on the US market. The same will soon happen with IPQM now this has become a Bellcore Generic Requirement. Training by Bellcore can be seen as an option for PNOs. Offers for this will be presented by Bellcore on the EIRUS meeting in October 1996.
- Operation support systems and transmission systems could now be measured by IPQM and RQMS.

There was a lot of feedback from the suppliers during this meeting. Some of the suppliers presented their remarks on the EIRUS work, and not everything was positive. At the end of the day, the message both from the suppliers and from the EIRUS group was clear, so some actions had to be taken before the next meeting.

→ World Wide Web
EIRUS has a homepage on the Internet

→ Location
http://www.eurescom.de/public/newpub.htm

→ Information
Members (PNOs, suppliers, contact persons)
Chairman & Secretary
Phases ...

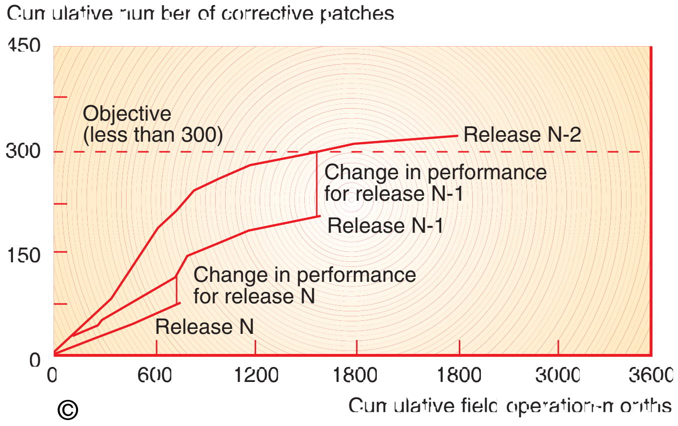


Figure 1 Examples of the format of one of the metrics in Phase 0 (© Bellcore)

EIRUS promised to be more clear in the way we communicate with our suppliers and in the future, we also want to see more interaction between the PNOs and the suppliers. Outside the meeting with the suppliers a lot of work was carried out by different groups within the EIRUS membership.

One group took care of the preparation on Phase 1 and another started with the homepage on the World Wide Web. A third group prepared an information package to be sent out to anyone interested in the work of EIRUS.

Facts and figures (Jan 1997)

- **PNOs with full membership in EIRUS:**

  Belgacom
  British Telecom
  Deutsche Telekom
  Hellenic Telecom
  PTT Telecom
  Swiss Telecom
  Telecom Italia
  Telecom Finland
  Telia

- **PNOs not yet members of EIRUS:**

  France Telecom
  TeleDanmark
  Hungarian Telecom
  Telenor
  Telefonica

- **Suppliers supporting EIRUS:**

  Alcatel
  Bosch Telecom
  DeTeWe AG
  ECI Telecom GmbH
  Ericsson Telecom
  GPT
  ITALTEL
  Lucent Technologies
  Nokia Telecommunications
  Siemens AG
  Siemens ATEA
  Siemens ÖN
  Siemens Schweiz

Today, EIRUS has a homepage on the Internet, and PNOs or suppliers who wish to make a contact will find some useful information there on how that could be obtained.

It was decided that we needed an extra meeting in June in order to make a decision on extension of Phase 0.

A meeting with the PNOs only was held in Rovaniemi, on the Arctic circle, 17–18 June 1996. Some jokes about long days were heard, but we made it in reasonable time. It was decided that we will proceed with Phase 1 and NOT, as many suppliers wanted, extend Phase 0. The plenary came to the conclusion that we want to “show some muscle”. It is very important to carry on with the implementation and show both our management and our suppliers that we are in control of the situation.

At the meeting in Rovaniemi we also had information from the EURESCOM project P619. EIRUS decided that a formal relationship between EIRUS and P619 is not needed, but a flow of information between both groups is accepted. The P619 project concludes at the end of 1997. Some examples of information flow:

- Phase 0 reports will be given to P619
- Task 4 of P619 will further inform EIRUS about the ongoing work in the project.

The next meeting took place in Athens 15–17 October 1996, and at that meeting we gave more time to the suppliers.

I have a positive outlook for the future of EIRUS and think that the climate at the meetings with our suppliers will improve.

I also look forward to extended future interaction between the PNOs and the suppliers. It is very important to make clear to our suppliers that we *don’t* discuss commercial business in this forum. All contractual things must be brought up between each PNO and its supplier. It is also important to clarify that we don’t want to compare our suppliers – that is not the goal of EIRUS. But using uniform measurements will help us, operators and suppliers in performing our services to the customers with a better quality and in time.

Why EIRUS?
– 5 years quality assurance by Betatest NKÜ 2000

BY HELMUT IMLAU

Introduction

A central part of Deutsche Telekom's project SYNET *(SYnchronous NETwork)* is the transmission network node *NKÜ 2000* (in German: Netzknoten der Übertragungstechnik). *NKÜ* is the product name of a DXC 4/1 (Digital Cross Connect System), specified by Deutsche Telekom. At the beginning there were three different types from different suppliers, now there are about 90 network nodes from two different suppliers in operation.

It was the first complex software controlled system of Deutsche Telekom's transmission network, which was one reason for establishing a System Management Group for Transmission Network Nodes in Bremen.

One of the groups' tasks is software quality assurance combined with software testing during the development process. The testing, carried out in parallel with the development, is called *Betatest NKÜ 2000*.

This article gives an overall view of the system DXC 4/1, the work of the System Management Group for Transmission Network Nodes, methods and results of the tests.

After a short introduction into EIRUS (European IPQM and RQMS Users) and the Bellcore document based metric systems E-IPQM (European In-process quality metrics) and E-RQMS (European Reliability and quality metrics in telecommunications) the relations between *Betatest NKÜ 2000* and E-IPQM are shown.

A further part describes first experiences with EIRUS and the usage of the received metrics.

The author of this article started his work within the new System Management Group for Transmission Network Nodes in a team which dealt with one of the DXC 4/1 in 1991. Since 1995 he has been working as assistant head of department and is responsible for system independent tasks. One of these tasks is the work within EIRUS for implementation of E-IPQM and E-RQMS into development, operation and maintenance transmission systems. The following statements are often personal experiences.

1 NKÜ 2000 = Deutsche Telekom's DXC 4/1

The DXC 4/1 provides the cross connect levels

- AU (Administrative Unit) -4
- TU (Tributary Unit) -3, -2, -12 and the ports:
- STM1 (Synchronous Transport Module 1 = 155 Mbit/s), and
- E4 (140 Mbit/s). E1 (Mbit/s), E3 (34 Mbit/s).

The operation can be done via

- OAMT (Operation, Administration and Maintenance Terminal), or
- via Q3-interface.

For the NKÜ 2000 history, please refer to Table 1.

The DXC 4/1 physically consists of ports, cross connection matrix and controller. The controller architecture is hierarchical with 3 levels: on-board software, boards controlling software and node control software. For realisation of the

Table 1 History of NKÜ 2000

1989	Start of project with 3 suppliers
1991	Start of development sample tests, start of Betatest NKÜ 2000 Establishment of the system management group for transmission network nodes
1992	Installation of 3 different systems, start of field test (17 DXC 4/1)
1993	21 additional DXC 4/1
1994	Start of operation with 2 different systems via local and remote OAMT
1995	Start of operation with management system (65 DXC 4/1)
1996	Delivery of additional package performance management, about 90 DXC 4/1 in operation

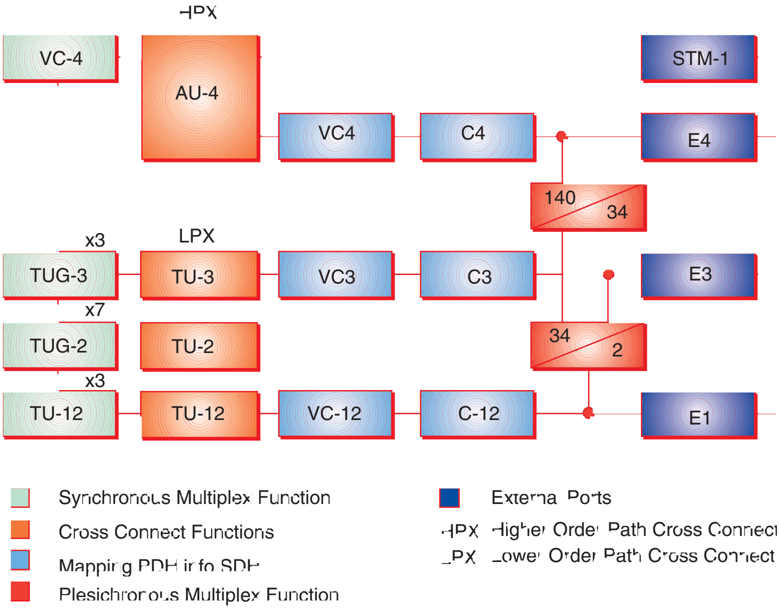


Figure 1 Multiplex Structure of the DXC 4/1

management fields configuration management (CM), fault management (FM) and performance management (PM) the network nodes need complex software systems. According to supplier's statements the software part needs 70 – 80 % of the overall effort for development of the NKÜ 2000.

2 Work of the System Management Group for Transmission Network Nodes

Our task is to deliver reliable systems to our operational staff.

We are responsible for

- Fault management
- Modification
- Further development

of the systems. An important task is the test of new or modified versions.

In the field of transmission technology only the following mostly hardware oriented tests were used by Deutsche Telekom up to 1990:

- Development Sample Test
- Acceptance Test
- Field Test
- Type Sample Test.

Due to the large software systems the decision was taken, that in the field of software for NKÜ 2000 more effort for software quality assurance was necessary. Now, an additional number of quality test activities are being carried out. They are described in the commercial contracts with the suppliers.

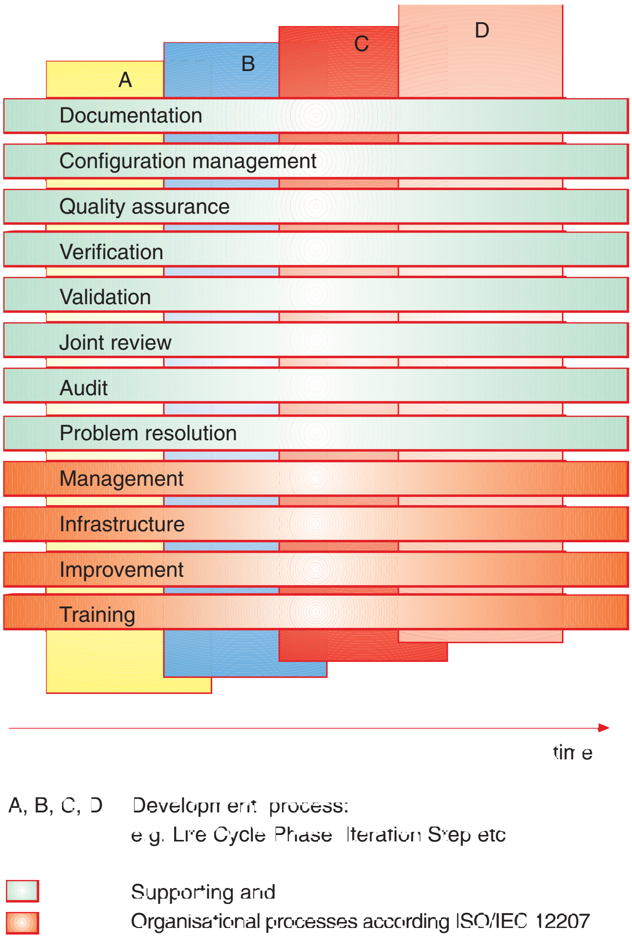


Figure 2 Process model

3 Quality assurance by Betatest NKÜ 2000

Quality assurance in EN ISO 8402 [1] is described as *"all the planned and systematic activities implemented within the quality system and demonstrated as needed, to provide adequate confidence that an entity will fulfil requirements for quality. There are both internal and external purposes for quality assurance: ... b) external quality assurance: in contractual or other situations, quality assurance provides confidence to the customers or others."*

The following quality assurance methods have been chosen:

- Participation, modification and execution of early software tests of the development process (module- or class-, integration- and system test). That is the original reason for the name *Betatest NKÜ 2000*.
- Software quality assessment including development process quality with judgement of the supporting and organisational life cycle processes. A main point of interest during the test execution is the practical realisation of the planned supporting and organisational processes.

These two kinds of tests are our way of realising quality assurance, we call this combination *Betatest NKÜ 2000*.

An example for another way of carrying out quality assurance is the "Glass Box Test" by KPN Research, Netherlands.

The following points explain scope, activities, test planning, test execution, evaluation of the results and experiences from our tests. During the test of the first system versions we got a good impression of the ongoing development. For the next versions the test effort could be reduced. A possible influence of EIRUS is described in Chapter 6: Usage of the received metrics.

3.1 Scope

According to ISO/IEC 12207 (Information technology – Software life cycle processes) the software life cycle processes can be divided into primary, supporting and organisational processes. The primary life cycle process includes the engineering view, e.g. with the development process and its life cycle model (e.g. waterfall or spiral model). The supporting process includes the quality management view with for example the quality assurance process. The organisational process includes for example the management and infrastructure process [2].

Figure 2 shows a model of the development process with supporting and organisational processes. These processes are subject of the quality assurance by Betatest NKÜ 2000.

Our tests include

- Test realisation activities in one life cycle phase of the development (e.g. design phase or class test)

and at the same time

- Test of suppliers practical realisation of the planned supporting and organisational processes within the chosen life cycle phase.

To choose the real test object within the concerning life cycle phase, we take a random sample. A combination of software quality knowledge and system knowledge is necessary to carry out these tests.

The tests are performed separate for each system, sub-project and life cycle phase inside each sub-project.

3.2 Test planning

Basis of the test planning is a Supplier Independent Strategic Test Plan. It describes the goals of the tests with regard to the required quality characteristics like capability for maintenance, code modification and test. Criteria are user-friendliness, efficiency, correctness and robustness.

The Supplier Independent Strategic Test Plan has to be transferred to the Supplier Specific Strategic Test Plans for each type of DXC 4/1. These test plans aimed to reflect the real system structure, e.g. the distribution into sub-projects.

The third step of the test planning is the Operative Test Plan with the details of testing for each subsystem.

Figure 3 shows test planning and the evaluation of the test results.

3.3 Test execution

The test execution consists of the following activities:

1. Introductory meeting concerning the life cycle phase. It takes place at the development location of the supplier. The emphasis is on realisation, quality assurance, reviews, and tests.
2. Ordering of the current project documentation with description of life cycle phase, development- and test tools, documentation of achieved results, executed tests, source code, etc.
3. Test preparation by Deutsche Telekom: document analysis of the development- and test documentation, source code analysis, etc.
4. Test execution at the development location of the supplier. Presentation of the analysis results, discussion of the results with the developers.
5. Evaluation with checklists.

3.4 Evaluation of the test results

The test execution is evaluated in test reports. The results are summarised in checklists for each subsystem and each life cycle phase. The next step is to prepare quality reports for each system and on the highest level quality comparison reports for our management.

3.5 Results and experiences

During our tests we became acquainted with many different development models and philosophies.

Often the development of NKÜ 2000 was the first large software project of our suppliers' transmission business units. Mostly, it was a Europe-wide development. There were co-ordination problems in the beginning. The software and development process quality differed from one location to another.

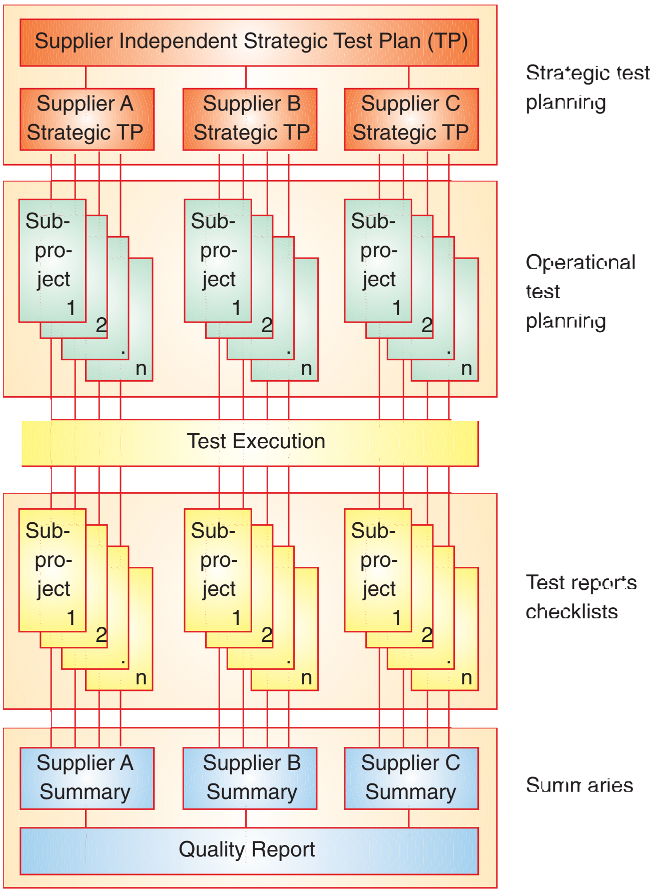


Figure 3 Betatest NKÜ 2000

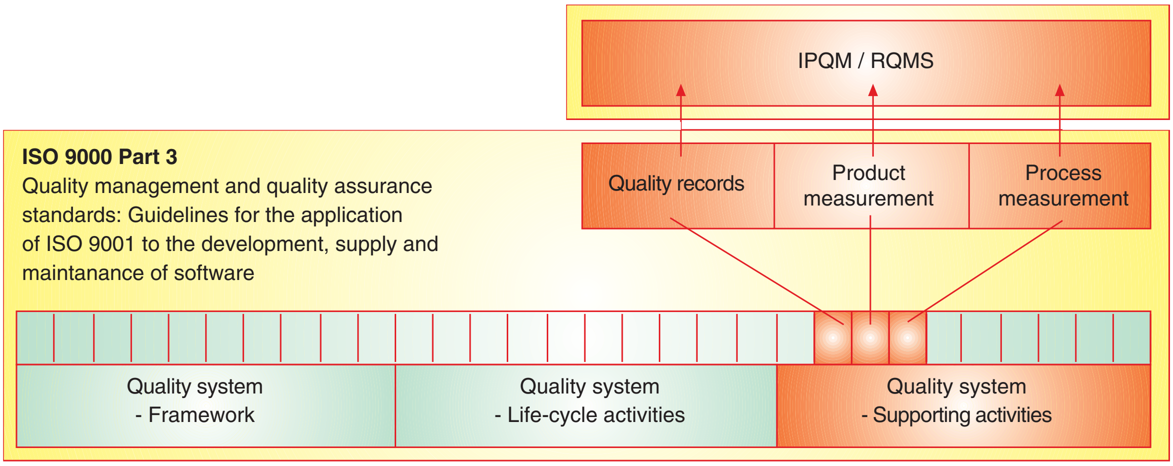

IPQM / RQMS
ISO 9000 Part 3
Quality management and quality assurance standards: Guidelines for the application of ISO 9001 to the development, supply and maintanance of software
Quality records
Product measurement
Process measurement
Quality system - Framework
Quality system - Life-cycle activities
Quality system - Supporting activities

Figure 4 E-IPQM & E-RQMS are based on ISO 9000 Part 3 / ISO 9001

Object oriented development as required in Deutsche Telekom's Technical Requirements was seen. DXC 4/1 – software was often the first object oriented project of the development departments. Of course, there were some problems.

Usually, the supplier supported us well during the running tests. We had very constructive feedback to our comments and ideas. There were very intensive discussions which have been helpful for future tests and which have had influence on further developing our test methods.

4 EIRUS

The following gives a short introduction into EIRUS, the relations between E-IPQM/E-RQMS and ISO 9001, and our work within EIRUS.

For a complete introduction into EIRUS, see Rolf Johansson's article in this issue [3]. He explains the benefit of uniform metrics for suppliers and PNOs (Public Network Operators). The following shows the point of view of a PNO, implementing EIRUS together with the suppliers in development, operation, and maintenance of transmission systems DXC 4/1.

4.1 Short introduction

EURESCOM, the European Institute for Research and Strategic Studies in Telecommunications, dealt with the "Introduction of Uniform Quality Measurements for Telecommunications" from 1993 – 1995 in the project P307 "Reliability Engineering" Task 4 [4].

The result of this project was a recommendation for introduction of the Bellcore Generic Requirements IPQM [5] and RQMS [6] in a modified form for Europe. The European versions are named E-IPQM and E-RQMS. The first modification was an introduction strategy for implementation of the metrics and measurements.

E-IPQM aims at the introduction of a high level development process with internal feedback in the supplier development process. This feedback should be used for process improvement. A lot of metrics are proposed for this feedback, they are a measure for development process quality. These metrics can be reported to PNO according to bilateral agreement. E-IPQM runs during the development process of the product.

E-RQMS describes measurements, which are a measure for product quality. E-RQMS starts with a system test and covers the system in operation.

The user group EIRUS of the European PNOs was founded to implement the practical steps and adapt E-IPQM and E-RQMS together with suppliers.

The results of the implementation should be an input for EURESCOM project P619 Task 4, which is currently running. It should be the platform for further development of E-IPQM and E-RQMS.

In the beginning the user group was dominated by experts for switching systems, although in the Bellcore documents there is no limitation for the kind of systems. In RQMS there are special parts for a lot of systems including switching, transmission and operating.

4.2 Basis

All PNOs have their own contracts with their suppliers. In these contracts the used test methods of the PNOs are described. Possible data, which the PNOs require from the development, maintenance and operation processes, are also included.

EIRUS does not change existing contracts. For further systems E-IPQM and E-RQMS should be implemented in the contracts, as far as the PNOs require them and the suppliers are prepared to deliver them. The implementation phases are learning phases for both PNOs and suppliers.

An important question for the suppliers is the amount of additional work. To answer this question a glance at ISO 9001 [7] follows in the next chapter.

4.3 E-IPQM, E-RQMS and ISO 9000/9001

Most of the suppliers work in accordance with ISO 9001. ISO 9000 Part 3 describes guidelines for the application of ISO 9001 for development, supply and maintenance of software [8]. One chapter of ISO 9000 Part 3 describes the supporting activities of the quality system. There is a demand for the following activities:

- Quality records
- Product measurement
- Process measurement.

It is only stated that these points should be carried out, but there is no description how this is to be achieved. So the suppliers have to provide quality records, product and process measurements for their own process and product improvement.

E-IPQM and E-RQMS are detailed descriptions of how to do this. They are one possibility to fulfil the ISO 9000/9001 so they are based on these standards (see Figure 4).

For the introduction of E-IPQM and E-RQMS and the implementation into current projects EIRUS requires no introduction of new methods. At the first step EIRUS want to use existing quality records, product and process measurements. For new projects EIRUS expects further evolution of these points within suppliers development process in direction of uniformity to E-IPQM and E-RQMS. This could enormously reduce the effort of the supplier to deliver special metrics for individual PNOs.

4.4 Quality assurance by Betatest NKÜ 2000 and reasons for EIRUS

We supported the user group from the beginning and forced the introduction into transmission systems.

Our reasons for implementing E-IPQM and E-RQMS are:

First reason: Implementation of uniform metrics within the organisation of one supplier.

- During our tests we noticed a lot of different methods for development, internal project management, project reporting and quality assurance. We found out differences in the development process quality within the same supplier. Some of the different development locations had taken their own process model into the project. There was a need for co-ordination and often it was not very easy for the project leader. We registered a requirement for process improvement.

Second reason: Increasing the efficiency of our test method.

- In order to make our work more efficient, we need better triggers to start our test activities. The E-IPQM and E-RQMS system could deliver additional triggers. This would save work for suppliers and Deutsche Telekom. Our test method would get a continuous share without continuous test sessions at the supplier location. (See: 6. Usage of the received metrics.)

4.5 Cooperation with suppliers

EIRUS started its work together with the suppliers in 1995. Today, most of the suppliers support EIRUS in the chosen projects. They wish to play an active role in the implementation and development of E-IPQM and E-RQMS.

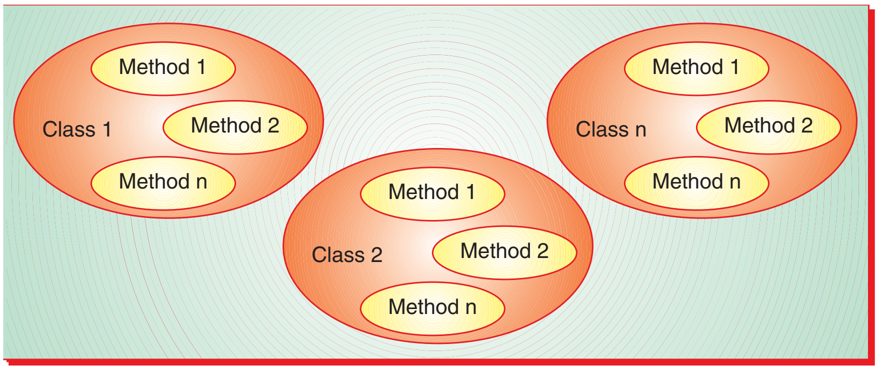


Figure 5 Structure of a object oriented system

Table 2 Estimation units for object oriented systems

Level			Example
1	Method level	Inside one method	LOC, McCabe
2	Class level	Inside one class	Number of methods Parents class Child classes
3	System level	Inside the whole system	Number of classes Depth of the inheritance tree Inheritance relationships Using relationships (Coupling, Cohesion)

We noticed that they are making an effort to uniform their project management, project reporting and quality assurance system in general. We think this process is influenced and encouraged by the customer requirement for a uniform metric and measurement system.

Often, the experts of the different projects involved in EIRUS within the same supplier (e.g. transmission and switching) are co-operating for the introduction of a uniform metric system. That has an effect on business units and projects, which have not been involved in EIRUS up to now.

5 Experience with E-IPQM up to now

During the implementation phase 0 we tried to implement the metrics Software *Size and Variance, Milestone Monitoring* and *Test Tracking*. The phase 0 E-RQMS metrics were not applicable for our transmission systems. *Outage and Downtime* is a special metric for switching systems, the adequate metric for transmission systems, *Performance and Availability* has been chosen for phase 1. The metric *Patches* is not applicable. There is no patching in the DXC 4/1 because of different software structure. The following is our experience with the E-IPQM metrics.

5.1 Software Size and Variance

There are many difficulties with *Software Size and Variance*. The metric requires a software estimation process with an estimation unit. The estimation units of the original Bellcore document such as *Lines of Code* (LOC) are not enough for object oriented development. In the area of Q3 interfaces and object oriented information models we have a lot of object oriented development.

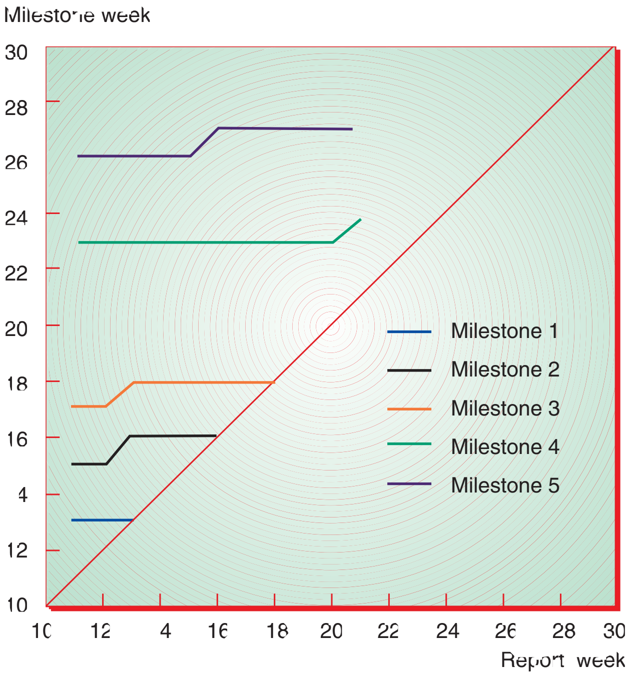


Figure 6 Milestone Trend Analysis

Therefore, there is a need for further development of E-IPQM. EIRUS is starting this work with a small working group together with three suppliers. Figure 5 shows the structure of an object oriented system, and Table 2 contains possible estimation units.

5.2 Milestone Monitoring

Milestone Monitoring is in common use in software development. It is necessary for the project management to have the process under control. We noticed many good ideas for this metric from the suppliers. One of them is the usage of *Milestone Trend Analysis* or *Milestone Delay* (see Figure 6) .

This diagram allows fast evaluation of the relevant milestones. We get monthly updates of the diagram. The whole history is included in the current chart. Every delay is shown as a line upwards. Milestones without delay are represented with lines parallel to the x-axis.

5.3 Test Tracking

Test Tracking is a well-known method for reporting the test process. We often used it during our *Betatest NKÜ 2000*. It was easy for our suppliers to deliver this metric.

6 Usage of the received metrics

The introduction of E-IPQM and E-RQMS not only needs implementation work on the supplier side. A supplier's question is: What will a PNO do with the received data?

EIRUS has realised that a formal assessment as described in IPQM and RQMS by green, yellow and red light, depending on a calculated index, is not useful. A problem can have many different reasons. To assess the effect of a problem, the supplier and PNO must have additional information. Knowing the reason for the problem they can discuss improvement plans and further activities.

The received data should give us a continuous view of the current development. In this way our test method could acquire a better continuity.

Data without background information about the development process does not seem useful to us. Therefore, we have agreements with our suppliers to provide us with additional information.

We think all information together gives us a comprehensive view.

The received data should be used to answer the following questions:

- How well are development process and test running?
- Are there any delays?
- How good is the development process quality?
- Are additional steps or improvement plans necessary?
- How effective are development and testing?
- What are the current problems?
- How good are maintenance and fault management of the supplier?
- How good is the expected product quality?

During the tests of the first releases we got a good impression of the software development DXC 4/1. Therefore, we were able to reduce expenditure for the next versions.

Now we try to fix the trigger events for our test execution.

Examples of trigger events are:

- Problems with previous version
- Own recommendations for changes in the development process during previous tests
- Change of development- or test tools
- Necessary regression tests after a long time without tests.

E-IPQM can deliver additional trigger events for test duration. As far as possible the received metrics can be continuously evaluated with influence on our tests. This fact could be more important with a growing number of reported metrics.

For a complete integration of E-IPQM into our test method we need further experience with more metrics, supplied over a longer time.

Summary

Up to now there have been good results with the introduction of E-IPQM and E-RQMS for different transmission projects between suppliers and PNOs.

With the background of quality assurance tests by *Betatest NKÜ 2000* we think E-IPQM is a good way to improve process quality. The implementation of the metrics will help supplier's project management to get real information about the development. The implementation should result in better product quality, measurable with E-RQMS metrics.

In the long run, implementation of E-IPQM and E-RQMS on the interface between supplier and PNO is more economical for both of them. Now we are in the very positive position of being able to collect experience jointly.

E-IPQM delivers additional trigger events for our test method. The implementation of E-IPQM into our test method is not yet finished. Therefore, we need further experience.

The results of the EIRUS implementation phase 0 are optimistic, the process EIRUS is running. The first experience shows a need for further development of E-IPQM and E-RQMS.

(Article received October 1996.)

Abbreviations

AU	Administrative Unit
C	Container
DXC	Digital Cross Connect System
DXC 4/1	DXC 4 (STM-1, E4) / 1 (E1)
E-IPQM	European In-Process Quality Metrics
E-RQMS	European Reliability and Quality Metrics in Telecommunications
EIRUS	European IPQM and RQMS Users
Eurescom	European Institute for Research and Strategic Studies in Telecommunications
IPQM	In-Process Quality Metrics
ISO	International Standardisation Organisation
LOC	Lines of Code
NKÜ	Transmission Network Node (in German: Netzknoten der Übertragungstechnik)
OAMT	Operation, Administration and Maintenance Terminal
PNO	Public Network Operator
RQMS	Reliability and Quality Metrics in Telecommunications
STM-1	Synchronous Transport Module 1 (155 Mbit/s)
SYNET	Project SYnchronous NETwork of Deutsche Telekom
TU	Tributary Unit
TUG	Tributary Unit Group
TZ	Technology Centre (in German: Technologiezentrum)
VC	Virtual Container

References

1 EN ISO 8402. *Quality management and quality assurance : vocabulary.* August 1995.

2 ISO/IEC 12207. *Information technology : software life cycle processes.* August 1995.

3 Johansson, R. EIRUS. *Telektronikk,* this issue.

4 Eurescom P307. Reliability engineering. Abstract of *Towards Uniform Quality Measurements in Europe : E-IPQM and E-RQMS, version 1.0.* February 1995.

5 Bellcore Generic Requirements GR-1315-CORE. *In-process quality metrics, issue 1.* September 1995.

6 Bellcore Generic Requirements GR-929-CORE. *Reliability and quality measurement system, issue 1.* December 1994.

7 ISO 9001. *Quality systems : model for quality assurance in design/development, production, installation and servicing.* August 1994.

8 ISO 9000. Quality management and quality assurance standards, part 3. *Guidelines for the application of ISO 9001 to the development, supply and maintenance of software.* June 1991.

Glass-BOX – more than a transparent black box

BY CHRIS ALDENHUIJSEN

Within PTT Telecom, DNO is the department that is responsible for the development of telecommunications services. DNO has renewed itself and its methods, to be able to cope with an increasing number of services to be developed and with the shorter times available to market. The new processes are designed to enable PTT Telecom to deliver the services on time, with the right quality. Glass-BOX® is the name given to the process that defines the interaction between DNO and its external suppliers as the key of the vendor management policy.

Glass-BOX provides guidelines, requirements and tools like Project Management Plan, Project Quality Plan and Project test strategy, and processes like the Test and Acceptance process and System Integration, together with all the roles and responsibilities involved.

Background: services through co-operation between PTT Telecom and suppliers

PTT Telecom is the operator of the public fixed telecommunications network in the Netherlands. For mobile services, it has been working in a competitive market since 1995, and for the services delivered with the fixed network competition will start mid-1997 by deregulation of the current monopoly.

Like other Public Network Operators, PTT Telecom in recent years has been adding to the services that it has traditionally been providing with its fixed telecommunications network, new services like ISDN, 06-services – similar to the USA 800- and 900 services – and Call Forwarding. This process is expected to go on in the years to come, at a considerably increased pace. For the provision of these services, new functionality is necessary in the network.

DNO – for Diensten en Netwerk Ontwikkeling (Services and Network Development) – is the department of PTT Telecom that is responsible for the technical realisation of telecommunication services and infrastructure renewal, and thereby for the availability of functions at the right moment with the right quality. Its main activities consist of specification, project management, quality management and testing.

PTT Telecom does not develop or produce any equipment, it confines itself to supplying services. In supplying its services, PTT Telecom depends entirely on products from its suppliers. PTT Telecom has two important suppliers of switching systems and a large number of suppliers for transport, management and operations support systems.

PTT Telecom develops its telecommunication services in close contact with the suppliers of the systems involved. In that process, DNO writes the specifications for the functions in the systems and the suppliers are responsible for the development, production and verification of the software and hardware necessary for the functions.

The time-to-market, quality and costs involved in the creation and operation of services are to a large extent determined by the performance of the suppliers. The speed of introduction of services depends, among other factors, on the time the supplier needs for development. Development of a new release for switching software can take up to two years. During that process, DNO follows the development of the systems at the supplier's by monitoring the progress of the project and the supplier's control over it and, in the case of delays, it keeps an eye on the efforts for recovery.

Failures in the services can be caused by faults in the software of switching systems. The sheer size and complexity of the software in switching systems makes any hope that these systems can be without faults illusory. The supplier ensures that proper quality assurance measures are taken, and that adequate tests are executed to detect faults in the product in an early stage; DNO has the right to verify this by surveillance and audit activities.

The services nearly always depend on the correct co-operation between the products from different suppliers. The activities involved in the creation of this interworking are together called System Integration. They are generally executed, as a separate project, by an external party in the role of System Integrator. Usually, this role is confided to the supplier of one of the systems involved in the interworking; in that case this supplier plays two separate roles in the process, that of the party that is responsible for the operation of the system parts embedded in its own (switching or support) system, and that of System Integrator. Unexpected problems can appear during this System Integration with products of different suppliers, and also for System Integration quality assurance measures and test procedures are defined.

Finally, DNO is also responsible for Acceptance Testing of the existing services as part of releases, and of the new services as supplied by the Systems Integrators.

Working in projects

In order to be able to cope with the increased number of services to be developed, DNO has in recent years renewed itself and its methods. For this innovation, DNO has executed several quality programs, directed at its internal processes as well as at the interfaces with its environment.

For the internal operations of DNO, the following principles are accepted:

- A culture within the organisation of pragmatic attitudes between business partners
  - Formal client–supplier relationships between engineering and procurement departments and the service oriented stake holders with the product
  - Formal commitment from the start of client parties, as opposed to projects being started under 'technology push'
  - Client oriented attitudes in engineering and procurement departments
  - Correct co-operation between teams within the project
  - Internal calculation of costs and benefits in conformance with market practices
- Good knowledge among all people involved, of project practices and of project management
- Good project support from information systems and from administrative services

® *Glass-BOX is a Registered Trade Mark of PTT Telecom B.V.*

- The presence in the business processes of sufficient preventive elements, as opposed to corrective elements and procedures:
  - Effective reviewing of plans, rather than correcting products after faults have resulted in failed tests or in failures occurring in the field
  - Effective testing of all aspects of the product in the earliest possible stage
- Clearly defined business processes, procedures and responsibilities
  - The possibility for early estimation of total project costs, and for controlling the actual costs during the process.

To reach this situation, changes have been introduced within DNO by implementing a *systems model*. According to this model, DNO is seen as a production unit, with new and improved services as its major products, among consultancy, impact analysis and overall project management. The costs of the unit are met by selling the services. Along these lines, productivity norms and tariffs per hour are defined for each function, and measurable targets for the results of project managers as well as for hierarchical managers.

The most important interfaces of DNO are those with its clients within PTT Telecom, like Marketing and Sales departments and Operational departments, and with the suppliers of telecommunications equipment. For the latter of these interfaces, a set of procedures, guidelines and requirements that define the interaction between DNO and its external suppliers was created, under the name Glass-BOX.

Description of Glass-BOX

In the change process directed at defining project oriented procedures within DNO, it was found that a project oriented approach within DNO could only be successful if it was matched with a similar approach at the suppliers' side of the process as part of the chain to create an improved or new service. Also, it was seen as necessary that the coupling of both processes was assured by visibility of the processes to the other party. This is why DNO created and introduced Glass-BOX, the process which defines the interaction between DNO and its external suppliers.

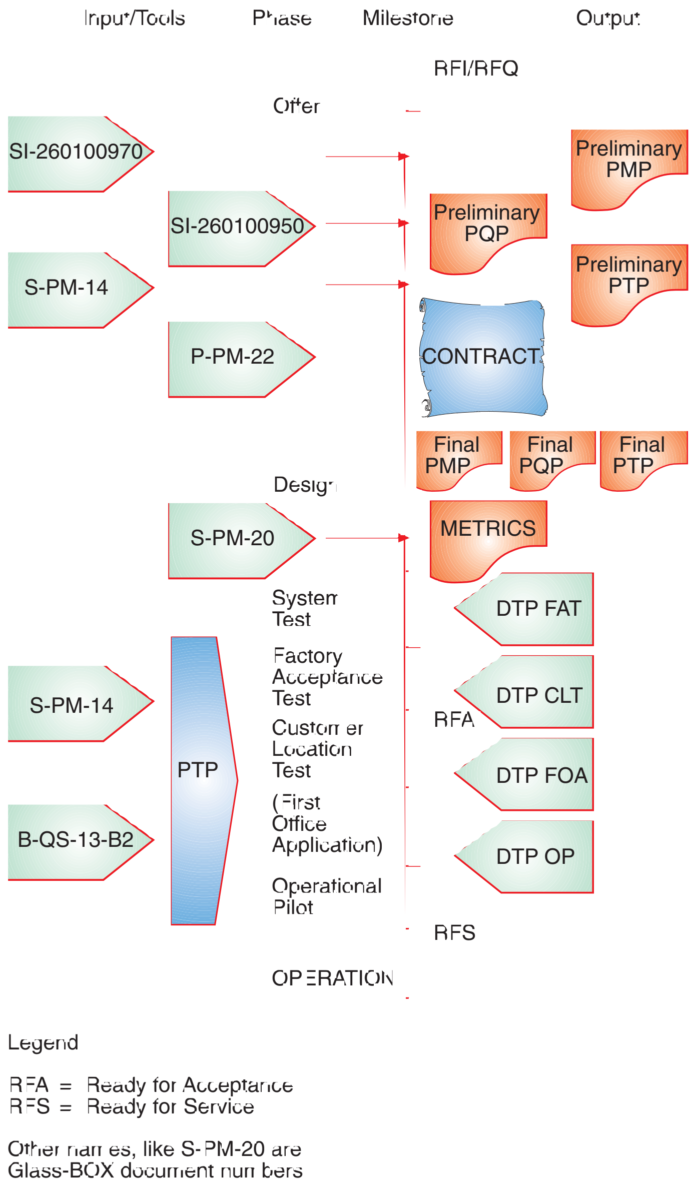


Figure 1 The Glass-BOX model uses preventive instruments for quality assurance and progress of software development

The purpose of Glass-BOX is to assure that good project plans are defined, and to give PTT Telecom an insight into the supplier's plans and in the project progress and the supplier's control over the process. Glass-BOX was primarily focused on software development, but it is in principle applicable to all (telecommunication) projects.

The Glass-BOX philosophy is based on applying *process management* principles, at DNO's own side of the processes as well as at the supplier's side, visibly and verifiably for DNO. For this, the supplier must have explicit plans for the various parts of the process. These plans must be:

- In a format and with a minimal content as defined by PTT Telecom
- Reviewed by PTT Telecom at predefined points in the project life-cycle

- Followed with regular reporting on the actual implementation, project progress and control and the deviations with regard to the plans, in a pre-defined format

Glass-BOX defines the development process for PTT Telecom services, and within this process the responsibilities and the interactions between PTT Telecom and the suppliers. Quality assurance is incorporated in the process, with the intent that it will work to prevent faults rather than to correct them. Special attention is given to defining the Testing and Acceptance process, also with the objective to have the various aspects of the system tested as early in the process as feasible.

In Glass-BOX, the following process management instruments are defined:

1. Project Management, in the Project Management Plan (PMP)
2. Project Quality Management, in the Project Quality Plan (PQP)
3. Project Test Management, in the Project Test Plan (PTP) and refined in various Detail Test Plans (DTPs).

In principle, the PQP describes in more detail the quality aspects that are treated summarily in the PMP, and in their turn the testing activities that are part of Quality Management activities in the PQP are worked out in the PTP and in the DTPs.

The development process for services starts with the contract phase during which PTT Telecom and the suppliers reach agreement on the functionality of the software in the system elements involved. This phase is followed by the design phase, in which the supplier develops the software. Then follows the Test and Acceptance phase.

The figure shows the new process with the Project Management Plan, the Project Quality Plan and the Project Test Plan in the contract phase. The process metrics are shown, and the various plans for the Test and Acceptance process.

The supplier is responsible for the PMP, PQP and PTP. It is essential that the supplier writes the plans in close co-ordination with PTT Telecom, the documents are also reviewed by PTT Telecom.

In the following sections, these plans are described in some more detail.

Project Management Plan

The PMP is the leading project document which defines the technical and management processes that are necessary for reaching the project goals. The PMP contains a description of the project goals, deliverables, organisation, dependencies and responsibilities, milestone definition and milestone planning over time, reporting mechanisms and risk analysis.

The plan must be written to conform with specification of PTT Telecom. Normally, preliminary versions of PMP, PQP and PTP are produced in the contract phase, with final editions available soon after the contract has been signed.

Project Quality Plan

The PQP defines the quality assurance measures taken for the project. It lists all aspects of the quality assurance measures, and all responsibilities down to the level of specific functional roles played by the people in the project.

The PQP is based on the supplier's own 'regular' quality assurance system, tailored to the needs of the project. Specific requirements for a project can be more stringent or allow more freedom than is prescribed in the generic system; the requirements stated explicitly in the PQP always prevail over the generic requirements.

Every PQP must at least contain the list of Quality Gates including entry/exit criteria, quality assurance measures with regard to subcontractors, plans for audits, reviews and surveillance, and quality reporting to PTT Telecom including in-process quality metrics.

PTT Telecom requirements for the supplier's quality assurance system are based on the ISO 9001/900-3/9004-2 standards.

Project Test Plan and Detail Test Plans

In the past, PTT Telecom had more and more been in the situation that the company itself ws executing many of the tests that were needed to get a product that was acceptable for introduction in the network.

The new strategy for Acceptance Testing is based on the principle that the supplier must prove that the product has the right quality. In addition, the test process must ensure that faults are detected as early in the process as possible, because correcting errors is cheaper earlier in the life cycle.

While the final responsibility for acceptance remains with PTT Telecom, the emphasis of PTT Telecom's involvement is on ensuring that adequate tests are conducted according to well defined plans.

The test objectives and planning of all phases are defined by the supplier in a Project Test Plan. The PTP also gives a time schedule for the production of the Detail Test Plans for each of the phases. In the DTPs, detail of test specifications and planning for test execution are defined.

Development Process Metrics

During the development process, the supplier produces periodical reports to keep PTT Telecom informed of the progress made. In these reports, an important place is taken by quantitative measurements that express percentages of goals reached, numbers of weeks delayed, etc.

Within Glass-BOX, the specification of these reports, and the items on which reporting should take place, is a live document. For a start, a subset was chosen of the measurements as found in the then current version of IPQM (In Process Quality Metrics, see *References* at the end of the article), the measurements proposed by Bellcore for monitoring software development processes.

Meanwhile, PTT Telecom has decided to use the set of IPQM measurements as adopted by EIRUS, the European user group for IPQM and RQMS metrics (see *References*). Naturally, the relevant Glass-BOX document will reflect this and other developments, and in due time it will take into account the measurements recommended by EIRUS.

The Test and Acceptance Process

This Test and Acceptance process is divided into four phases:

- FAT – the Factory Acceptance Test. In this phase, the supplier proves as far as possible that the product conforms to the functional requirements.
- CLT – the Customer Location Test. In the CLT, the proper interworking of the product with its environment is tested.
- FOA – the First Office Application. In the FOA, the validity of the procedures with regard to the operation of the product is tested.
- OP – the Operational Pilot. In this phase, the conversion of existing operational switches to the new release is tested and operational performance (system availability) is measured.

For each of these phases, Entry- and Exit-criteria are specified in the Project Test Plan, stating the conditions for transition from each phase to the next.

FAT

The Factory Acceptance Test is conducted at the supplier's premises and under his responsibility regarding execution. The schedule for test execution is communicated well before to PTT Telecom, and PTT Telecom is entitled to witness test execution. In the case that a test fails to pass, the supplier and the PTT Telecom project leader together decide on the classification of the importance of the fault (Fatal, Major, Minor).

The products are tested as far as is possible without the product being connected to systems that the supplier cannot be expected to have at his disposal, such as systems from other suppliers, PTT Telecom's management support systems, etc.

In the FAT, all aspects of the product's behaviour are taken into account, the functions that are newly introduced in the new release as well as the functions that were already present in the prior release.

The tests that will be executed are described in the Detail Test Plan of this phase (DPT-FAT), that is written by the supplier and reviewed by PTT Telecom before start of the FAT.

CLT

In the Customer Location Test, aspects of the product's behaviour are tested which call for direct connection to systems that the supplier does not have available, such as systems from other suppliers. The tests are conducted under the responsibility of the PTT Telecom project leader. The schedule for test execution is established in agreement with the supplier. The supplier project team is entitled to witness the test execution; as in the FAT, fault classification is done in agreement between both project leaders.

The site for the CLT is a special test environment, set up according to specifications of PTT Telecom to be as representative as possible for the operational conditions; it is equipped with facilities for access to current releases of all the systems that the product must work with. Contrary to the suggestion given by its name, it is not essential that the premises for the CLT are operated or owned by PTT Telecom. It could equally well be a specialised testing laboratory run by a third party, e.g. a different operator with whom PTT Telecom collaborates in testing.

The CLT starts with the installation by the supplier and restricted functional testing. In the CLT for new releases of a product, only those functions of the product are tested which were also present in the prior release. For testing the new functions, a separate process will be executed with new versions of the Operations Support systems, etc., as far as necessary for the use of the new functions. Normally, this is done under the procedures for System Integration (see *Background*), and in that case the separate phase is called SI-CLT.

The tests in the CLT are described in the Detail Test Plan of this phase (DPT-CLT), which has been written by PTT Telecom and reviewed by the supplier.

FOA

In the First Office Application, the new software is tested on a switch with live traffic. In principle, for this test a new switch is chosen which is ready to be taken into service but which has previously handled no traffic. Like in the CLT, during the FOA the new functions in the release are not used, the only services tested are those that are operational in the network.

In the FOA, the new product is loaded for the first time into a switch which is situated in a normal operational environment, and which is run by operational people as opposed to testing specialists. First, the new operator instructions and the manuals are tested, as are the changed operational procedures like those for setting the operational parameters of the switch.

When all goes well, the switch is entered into the network, and live traffic from selected users is gradually directed to the FOA switch, which will now co-operate with the operation support systems and with the other switches. These other switches can be from different suppliers or from the same supplier, in which case they will be running the old release.

The FOA is only applicable to new releases, this phase is not used for new functions in the software which are normally handled according to the System Integration procedures.

Again, the tests which are executed are described in the Detail Test Plan of this phase (DPT-FOA), written by PTT Telecom and reviewed by the supplier.

OP

When the regular working of the new release is tested in the FOA, one important aspect remains to be examined which could cause much trouble during roll out of the new release. This aspect is the conversion of switches with live traffic to the new release, and it its covered in the Operational Pilot. In the OP, the testing is concentrated on the procedures and functions that must ensure the correct conversion of all kinds of data involved in traffic handling, like billing records, settings of switch parameters, settings of parameters for user services (e.g. Call Forward), etc.

For new functions, normally under System Integration procedures, the appropriate phase is called the SI-OP.

The tests in the OP are described in the Detail Test Plan (DPT-OP), written by PTT Telecom and reviewed by the supplier.

After successful conclusion of the Operational Pilot, the acceptance process is finished and the product is accepted. New releases are ready for general roll out, the SI functions are ready for commercial introduction.

From idea to Registered Trade Mark

Glass-BOX started as one of the results of a change process directed at defining a project oriented way of working within PTT Telecom's Services Development department. In this process, the interfaces to the suppliers became involved, leading to the set of requirements and procedures now known as Glass-BOX.

Glass-BOX requires investment in human resources. It takes time to discuss, write and review PMPs and other documents, and it is not always easy to reserve resources for these tasks, which in the beginning are seen as overhead rather than as normal, efficient ways of conducting projects.

Introducing Glass-BOX took about two years. The importance of reserving time for the new, additional project activities was recognised more and more. For some employees who had difficulties in adapting to the new procedures, extra training and coaching was needed to make them see the advantages of the new ways of conducting business. Others saw the more formal approach as a sign of distrust, between partners that had been doing business for many years. It was an important factor that the suppliers saw that the counterparts at PTT Telecom themselves used the project methods and procedures. Experiences were shared between both parties, and together they used their creativity to adapt their ways of working to the Glass-BOX requirements.

Glass-BOX has changed the ways DNO works. Now, project leaders at PTT Telecom, as well as at the suppliers', see Glass-BOX instruments as useful tools for successfully conducting their development projects.

Nowadays, co-operation between telecom operators is spreading throughout the world, and Europe is no exception. Forms of co-operation can be seen in many fields, and one of its forms consists in groups of operators developing common attitudes towards suppliers and considering common testing of products that have much in common. In such a group, in which PTT Telecom participates, Glass-BOX is accepted as the basis for the operator–supplier interfaces. At least within this group, the label "Compliance to Glass-BOX" has recently gained some significance. To be sure that this label in the future will continue to have the meaning that PTT Telecom wishes it to have, the name Glass-BOX has been legally registered as a Trade Mark.

PTT Telecom considers it to be in its interest to contribute to promotion of modern, quality conscious development methods in the telecommunications industry. It does this by co-operation with other companies, e.g. in EIRUS. Also, it hopes that sharing its experiences with Glass-BOX, which it has developed for its own use, will serve the same purpose.

References

The Bellcore document defining IPQM is: IPQM, In-Process Quality Metrics. Bellcore. Generic Requirements GR-1315. Issue 1, September 1995. This document contains detailed measurement definitions and calculation rules for the measurements that are recommended in E-IPQM.

EIRUS is the E-IPQM and E-RQMS User Group. E-IPQM and E-RQMS are quality measurement systems based on Bellcore's IPQM (see above) and RQMS. RQMS is a system which measures quality of the performance of switching systems in Operation, and of the Maintenance process. More information on EIRUS and contact addresses can be found on the EIRUS homepage via:

http://www.eurescom.de/public/newpub.htm

For more information on Glass-BOX, please contact:

Harry Hendrich
PTT Telecom N DNO SUP
P.O. Box 30150
2500 GD Den Haag
The Netherlands

Tel: +31 70 343 9804
Fax: +31 70 343 9849

E-mail: /C=NL/ADMD=400NET/
PRMD=PTT Telecom/S=Hendrich/
I=HJFM/

or the author:

Chris Aldenhuijsen
KPN Research
P.O. Box 421
2260 AK Leidschendam
The Netherlands

Tel: +31 70 332 5332
Fax: +31 70 332 6477

E-mail: c.w.h.m.aldenhuijsen
@research.kpn.com

Acknowledgement

Parts of this article, including its title, are based on material that was written for use within PTT Telecom, by:

ir J. Snoeren
PA Consulting Group
Beneluxlaan 905
3526 KK Utrecht NL

and by:

drs F.H. ter Haar,
formerly at KPN Research,
now in PTT Telecom Mobile Services
Department.

Review of international activities on software process improvement

BY ØYSTEIN SKOGSTAD

1 Introduction

This paper is based on a working document in Eurescom project 619; "Customer – Supplier technical interfaces". In this project the whole concept of interworking between the supplier and the customer is studied. The topic of the paper is to give a brief impression of telecommunications related international activities of relevance for people working in areas like systematic quality improvement, software process improvement, system specification and acquisition and the like.

Before we embark on the subject, some definitions might be considered useful. These are not formal definition, but rather informal descriptions of some vital terms used later on.

1.1 Process improvement

With process improvement we think of all types of improvements in the work processes embedded in the total operation of an organisation. Systematic process improvement is a key factor for future success both for customers and suppliers. Achieved or planned process improvement is a major reason for the changes in the technical interfaces between a customer and a supplier.

1.2 Contents of a requirement specification

On the technical side, the basic documentation regulating the interface between a customer and a supplier is the requirement specification. A requirement is a statement that defines

- The functional properties of a product. With "product" in this context one should think of telecommunication network elements, such as switching systems, transport systems, terminals and the like. A product is constituted of various technologies, such as mechanics, electronics (hardware) and software.
- The non-functional properties of the product. Important non-functional properties may be reliability, maintainability and other quality factors.
- The usage aspects of the product. Most information technology systems have some form of user interface that should be carefully specified. One example may be the layout of the switching system operator – machine interface, and the definition of working procedures, or *work views* (for the operators of the product).
- The interfaces between the system and its surroundings (e.g. other systems).

1.3 Metrics

Measurements are needed to achieve quality improvement in a structured and cost effective way. One has to find out the quality performance of today of the essential products and processes, set goals for improvement, check whether the goals are reached, and finally set out new goals for further improvement.

There are three basic types of metrics:

- Product metrics that are used for evaluating some properties of the product. An example of these may be downtime performance for a switching system.
- Capability metrics are the metrics used to evaluate some properties of a supplier's capabilities. An example of such metrics may be the stability and maturity of the supplier's production process.
- Process metrics are the metrics used to evaluate some properties of a specific process. An example may be delays related to agreed milestones for a delivery.

2 Background EURESCOM projects

Process improvement, quality assurance activities and related metrics have been studied in several Eurescom projects.

2.1 P227 Software Quality Assurance

The P227 project (Software Quality Assurance) recommends a number of Quality Assurance Activities which can be used to enable the customers to establish a quality management system to assure and enhance the quality of the delivered software for telecommunication systems. These activities are to be performed by the customers, by the software suppliers, or by both.

The main objectives of P227 were

- To establish a common set of procedures and indicators for quality assurance of processes and products during all phases of the software life cycle
- To define methods to measure software quality indicators
- To set up a framework for the technical procurement process with special focus on software.

The results from P227 are documented in the following volumes:

Deliverable 1:

• Part I	Customers Assessment Documentation
• Part II	Suppliers Assessment Documentation
• Part III	General Survey of Software Quality Assurance

Deliverable 2: Recommendation for Quality Assurance of Telecommunication Software.

Deliverable 3:

• Part A	Recommendation for Quality Assurance of Telecommunication Software
• Part B	Technologies for Quality Assurance of Telecommunication Software:
- Vol. 1	Guidelines for Software measurement
- Vol. 2	Process Quality
- Vol. 3	Reliability
- Vol. 4	Maintainability
- Vol. 5	Metrics
- Vol. 6	Annex – Technical reports and survey reports

Deliverable 4: Buy IT; Recommended Practices for Procurement of Telecommunication Software.

2.2 P307 Reliability Engineering

The P307 project (Reliability Engineering) in its Task 4 gives recommendations for utilisation of adapted Bellcore metrics on a European basis. It lays the foundation for the formation of a European user forum for the proposed metrics (EIRUS). The deliverable contains the complete definitions of the E-IPQM and E-RQMS metrics.

E-IPQM provides insight into the process of software development as performed by the supplier, and early indications of problems.

E-RQMS provides insight into the operation and maintenance activities performed by the customer in co-operation with the supplier.

2.3 P619 Technical interfaces between PNO and supplier

In Eurescom project P619, the interface between the customers (PNOs) and their suppliers are under consideration. Important aspects of this interface are the requirement specification, supplier evaluation and follow up, and reliability and quality metrics.

3 European directives

Throughout Europe, and for different types of applications, there is a clear trend towards harmonisation of system requirements and acquisition processes.

For the interfaces between customers and suppliers, the most prominent European Directive is 93/38: Co-ordinating the procurement procedures of entities operating in the water, energy, transport and telecommunications sectors.

The Directive includes

- Details for technical specifications
- Details for qualification procedures.

The EEC Directive 93/38 focuses on the procedures of the procurement which are related to the establishment of an open European market. It is applicable if the product or service to be acquired is to be used for concession or enforced activities in the field of telecommunications.

EEC Directive 93/38 covers just a small part of the product life-cycle. It only describes the first steps of the product life-cycle:

- Technical specification (article 18)
- Call for tender (article 21)
- Supplier qualification
- Product selection.

4 EUROMETHOD

EUROMETHOD is an initiative by a European consortium, to provide guidelines for information system procurement across organisational boundaries.

4.1 Overview

The technical scope of EUROMETHOD is how to deal with Information Systems (IS). The definition of IS is fairly broad. Telecom software is not directly addressed.

The various parts of EUROMETHOD are structured in Figure 4.1.

The *Customer Guide* addresses the following topics:

- IS adaptations
- Preparation of Call for Tender Transactions
- Use of EUROMETHOD in the Production Process
- Use of EUROMETHOD in the Completion Process.

The Customer Guide may be of value to contract officers, procurement personnel and project managers. There are several useful hints and suggestions for the procurement process which should be known by everybody engaged in such work.

The *Supplier Guide* covers the same topics as the Customer Guide as seen from the supplier side.

The *Delivery Planning Guide* covers the following topics:

- Delivery Planning
- How to Assess a Problem Situation
- How to Define a Strategy
- How to Define Decision Points.

The *Method Bridging Guide* covers the bringing together of various techniques of commonly known Information System Engineering Methods. Object oriented methodologies are not particularly mentioned, neither are telecommunications oriented methods.

In the IS scope of the Guides and Models presented, the focus is on a general framework for procurement actions.

5 EPHOS

EPHOS is short for European Handbook for Open Systems. EPHOS is an initiative launched in 1989 by the EU Member States' public procurement representatives and the Commission of the European Communities. It is aimed at assisting procurers on the acquisition of open systems.

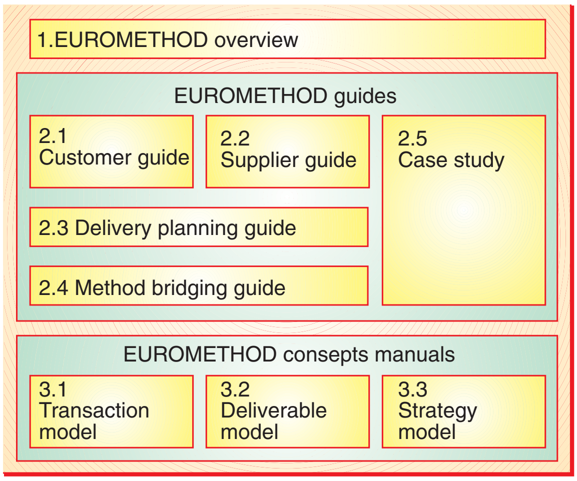


Figure 4.1 Overview of EUROMETHOD

5.1 EPHOS Handbook

The scope of the EPHOS Handbook is for Open Systems. This covers many aspects of information handling activity. Work continues on expanding the coverage of EPHOS to match the full open systems requirements. The topics contained in the current (spring 1995) version of the EPHOS Handbook have been selected by the Member State administrations. The majority of these topics cover aspects of communication within Open Systems.

The first EPHOS Handbook was published in 1992. This Handbook covered

- X.25
- Message Handling Services (MHS)
- File Transfer, Access and Management (FTAM).

The current issue of the Handbook also contains guidance on

- Document Formats
- Electronic Data Interchange (EDI)
- Virtual Terminal (VT)
- Character Sets
- File Transfer, Access and Management (FTAM)
- Message Handling Services (MHS)
- Directory Services
- Local Area Networks (LAN)
- LAN/WAN Inter-working
- Cabling
- OSI Management
- Testing.

Work is going on to further extend EPHOS with modules covering

- Data base enquiry
- Transaction Processing
- X.25 (maintenance)
- ISDN
- Metropolitan Area Networks (MAN)
- Operating System Services
- Security in business requirements
- OSI Management.

The new module on X.25 will replace the previous one on this topic. The module on OSI Management will extend the previous one. Future versions of EPHOS will cover other important aspects of Open Systems in response to the priorities assigned by the Member States.

5.2 The modules

The guidance on the selection of various open system types is divided into modules. Each module covers a given functionality within open systems, for example Document Formats. Each module consists of two parts:

- Part I: Basic Information and Procurement Advice
- Part II: Supplementary Information.

Part I: Basic Information and Procurement Advice

Part I is the core part of each module and is written in terms suitable for public procurers and planners of information technology and telecommunications systems and services. The target reader is a non-technical person who needs to specify system and service requirements, but may do so infrequently. The person knows how to write a requirement specification in terms of *what* is wanted in functional terms. Furthermore, the person has some basic knowledge of information technology.

The handbook gives guidance on

- The choice between alternative solutions
- Applicability of standards
- Availability of conformance and inter-operability testing.

This information is generally presented to the reader in the following way:

1 Introduction to the module

- 1.1 What functionality is offered by the module
- 1.2 When to use the module
- 1.3 Basic terms and concepts, that are specific for the module
- 1.4 How to choose the right solution

2 Standards, Options, and associated Procurement Clauses

- 2.1 Relevant profiles
- 2.2 Additional common requirements of EPHOS
- 2.3 Possible further user requirements
- 2.4 Problems and warnings

3 Conformance and inter-operability services

Part I of each module contains information considered of vital importance to the procurer. Part I can refer the reader to Part II which contains more detailed guidance, explanations, and tutorial information.

Part II: Supplementary Information

Part II gives explanations and justifications for the advices in Part I. Some further technical details and tutorial-like descriptions are provided on selected topics. Part II does not include any additional procurement advice not already given in Part I. The structure of Part II is equivalent to the structure of Part I. In addition, a chapter on Integration and Migration may be included. The typical reader of Part II is a more technically competent person seeking advice on behalf of the procurer or a technically skilled procurer looking for additional information.

6 ETSI

6.1 Overview

The youngest European standards making body, recognised by the European Council of Ministers by Council Directive 83/189, ETSI, was set up in 1988 to set standards for Europe in telecommunications and, in co-operation with the European Broadcasting Union (EBU) and CEN/CENELEC respectively, the related fields of broadcasting and office information technology. ETSI was born of a recognition that a pan-European telecommunication infrastructure with full inter-operability was the only basis on which a European market for communications equipment and services could thrive. Its main task is to accelerate the process of technical harmonisation.

The telecommunications market has changed considerably since ETSI was established, and ETSI has changed with it. In 1988 most countries in the European Union had monopoly network operation and service provision via their Public Network Operators. Now telecommunications are being liberalised throughout Europe, and competition is reducing the emphasis on transport related telecommunications to a commodity type product and value added services. This in turn is applying business pressures to the traditional telecommunications actors.

ETSI is becoming increasingly adept at responding to the market environment and anticipating new developments, and inevitably it has developed a global role, contributing to the standards making process world-wide.

ETSI is a forum (the only one in the telecommunications arena) which brings together all the key players such as network operators, service providers, manufacturers, administrations, users, and the research community. Only within ETSI there is a single forum which can take an overview of the market requirements for standardisation based on a genuine consensus. Furthermore, open systems are vital to the building of the infrastructure of the information society, and openness is a key feature of ETSIs standardisation activity.

ETSI has already achieved much in the challenge to create a harmonised telecommunications system, for example in ISDN, broadband, mobile communication and satellite communication – just four of the numerous areas in which significant progress has been made.

6.2 Publications

ETSI has submitted publications in various areas. There is next to nothing directly concerned with process improve-

Table 1 ETSI publications

ETS 300 411 : 1995-05	DE/TM-02215	Title: Transmission and Multiplexing (TM); Performance monitoring information model for the Network Element (NE) view	Scope: Agree an information model which describes in object-oriented terms a model for management of synchronous digital networks
I-ETS 300 416 : 1995-07	DI/TM-02105	Title: Transmission and Multiplexing (TM); Availability performance of path elements of international digital paths	Scope: Study and specify the availability aspects of the digital transmission infrastructure as seen from the transmission network access points
Draft prI-ETS 300 465 : 1995-01	DI/NA-042129	Title: Broadband Integrated Services Digital Network (B-ISDN); Availability and retainability performance for B-ISDN semi-permanent connections	Scope: Specification of performance parameters for availability and retainability of semi-permanent connection types
ETR 003 : 1994-10 Edition 2	RTR/NA-042102	Title: Network Aspects (NA); General aspects of Quality of Service (QoS) and Network Performance (NP)	Scope: General principles for QoS and NP in digital networks, including ISDN (revision of ETR 003)
ETR 011 : 1990-08	DTR/NA-042203	Title: Network Aspects (NA); The relationship between network component performance and the overall network performance	
ETR 155 : 1995-02	DTR/NA-052206	Title: Asynchronous Transfer Mode (ATM); Operation Administration and Maintenance (OAM) functions and parameters for assessing performance parameters	Scope: Analysis of performance parameters for B-ISDN, taking into account OAM functions and parameters
ETR 118 : 1993-12	DTR/SPS-03009	Title: Broadband Integrated Services Digital Network (B-ISDN); Switching, exchange and cross-connect functions and performance requirements	Scope: Switching and exchange performance requirements for B-ISDN systems
TCRTR 008 : 1993-08	DTR/NA-043203	Title: Network Aspects (NA); Network architecture, operation & maintenance principles and performance. Telecommunications Management Network (TMN) Vocabulary of terms	Scope: Document 43201 revised and expanded taking into account recent work & comments received during the NA vote on the original TCR-TR

ment. Very little is published related to metrics, or measurements related to processes. However, ETSI has some publications related to “performance”. This topic is related to Quality of Service, as well as to network element reliability. The ETSI publications shown in Table 1 may thus be of some interest for process improvement type of work.

7 ITU

7.1 General

Founded in Paris in 1865 as the International Telegraph Union, the International Telecommunication Union took its present name in 1934 and became a specialised agency of the United Nations in 1947.

The ITU is an inter-governmental organisation, within which the public and private sectors co-operate for the development of telecommunications. The ITU adopts international regulations and treaties governing all terrestrial and space usage of the frequency spectrum as well as the use of the geostationary satellite orbit, within which countries adopt their national legislation. It also develops recommendations to facilitate the interconnection of telecommunication systems on a world-wide scale regardless of the type of technology used. Spearheading telecommunications development on a world scale, the ITU fosters the development of telecommunications in developing countries. It achieves this by establishing medium-term development policies and strategies in consultation with other partners in the sector. It provides specialised technical assistance in the areas of telecommunication policies, the choice and transfer of technologies, management, financing of investment projects and mobilisation of resources, the installation and maintenance of networks, the management of human resources as well as research and development.

In essence, the ITU’s mission covers the following domains:

- A technical domain: to promote the development and efficient operation of telecommunication facilities in order to improve the efficiency of telecommunication services, their usefulness, and their general availability to the public.
- A development domain: to promote and offer technical assistance to developing countries in the field of telecommunications, to promote the mobilisation of the human and financial resources needed to develop telecommunications, and to promote the extension of the benefits of new telecommunications technologies to people everywhere;
- A policy domain: to promote, at the international level, the adoption of a broader approach to the issues of telecommunications in the global information economy and society.

7.2 Recommendation Z.410

With respect to process improvement, special attention should be given to the new ITU-T initiative on life cycle models, prepared by Study group SG X.

The present draft recommendation provides guidelines for applying the ISO/IEC 12207 standard to telecommunication systems. The ISO/IEC 12207 general framework for creation and management software is tailored and adapted to meet the specific needs for telecommunication products. It lists quality assurance elements, suggests what must be done and implemented in each life-cycle phase and which quality assurance activities shall be performed and who is responsible. These topics should be considered by a quality manager through the whole software life-cycle for the establishment of a quality assurance system.

The draft recommendation Z.410: Quality in Telecommunication Software Life-Cycle Processes (draft edition 17/4/96) considers the following topics:

- Quality Model
- Quality aspects in primary processes. The primary processes considered are initiation, development, operation, maintenance and retirement.
- Quality aspects in supporting processes. The supporting processes considered are documentation, configuration management, quality management and problem resolution.
- Quality aspects in organisational processes. The organisational processes considered are management, infrastructure, improvement, and training.

8 ISO and IEC

8.1 The organisations

The International Organisation for Standardisation (ISO) is a world-wide federation of national standards bodies from some 100 countries, one from each country.

ISO is a non-governmental organisation established in 1947. The mission of ISO is to promote the development of standardisation and related activities in the world with a view to facilitating the international exchange of goods and services, and to develop co-operation in the spheres of intellectual, scientific, technological, and economic activity.

ISO’s work results in international agreements which are published as International Standards.

ISO collaborates very closely with its partner, the International Electrotechnical Commission; IEC. An agreement reached in 1976 defines responsibilities: the IEC covers the field of electrical and electronic engineering, all other subject areas being attributed to ISO. When necessary, attribution of responsibility for work programmes to ISO or IEC is made by mutual agreement. In specific cases of mutual interest, joint technical bodies or working groups are set up. Common working procedures ensure efficient co-ordination and the widest possible global application.

ISO and IEC are not part of the United Nations, but have many technical liaisons with the specialised UN agencies. Several are actively involved in international standardisation, such as the International Telecommunication Union, the World Health Organisation, the Food and Agriculture Organisation, the International Atomic Energy Agency, etc.

ISO maintains close working relations with regional groups of standards bodies. In practice, the members of such regional groups are also members of ISO and the principle is generally accepted that ISO standards are taken as the basis for whatever standards are required to meet the particular needs of a given geographical region.

ISO and CEN (European Committee for Standardisation), for example, have defined procedures for the development of standards that will be acceptable both as

European Standards and as International Standards. The benefits for industry are wide-reaching. With the motto, “Do it once, do it right, do it internationally”, being practised, industry does not need to be involved in both European and international fora in the same areas.

8.2 Quality Assurance Standards

A basic prerequisite for process improvement is a defined process. The most prominent method of defining processes nowadays is as a part of a quality assurance system. Thus, the quality assurance standards, serve as a “ground floor” for the other process related initiatives.

The following is a list of the most important quality assurance standards submitted by ISO.

- *ISO 8402 – Quality vocabulary.* It provides formal definitions of the most important terms in the ISO 9000 family. Two terms are of special interest:
    - Quality Management: activities carried out by organisations to satisfy the quality-related expectations of customers and others.
    - Quality Assurance: activities carried out by organisations to provide to external parties (e.g. customers, regulatory bodies, shareholders) or to internal parties (e.g. management), confidence that the organisations will consistently meet the requirements for quality.
- *ISO 9000-1 Quality management and quality assurance standards – guidelines for selection and use.* This is the road map for the ISO 9000 standards; it provides information about the appropriate model for external quality assurance to adopt.

The following standards are Supporting standards, i.e. standards which are meant to be Guidelines for the use of the other standards:

- *ISO 9000-2: Quality management and quality assurance standards – guidelines for the application of ISO 9001, 9002 and 9003.*
- *ISO 9000-3: Quality management and quality assurance standards – guidelines for the application of ISO 9001 to the development, supply and maintenance of software.* The ISO 9003-3 document is under revision. A new Committee Draft (CD) may be obtained from the national standards organisation.

When a quality system according to the ISO standards is implemented, quality system audits should be performed as part of the work of keeping the implementation in line with the standards, and according to the needs of the company. Some ISO standards give regulations of the management of quality system audits. These standards provide tools and techniques which can be used to help in evaluating the application and progress of the Quality Management principles:

- *ISO 11001-1: Guidelines for auditing quality systems – auditing.*
- *ISO 11001-2: Guidelines for auditing quality systems – qualification criteria for quality systems auditors.*
- *ISO 11001-3: Guidelines for auditing quality systems – management of audit programmes.*

The mostly used ISO quality assurance standards are the Quality Assurance requirements standard:

- *ISO 9001: Quality systems – model for quality assurance in design/development, production, installation and servicing.* This standard refers to activities an organisation carries out to provide to external/internal parties confidence that it will consistently meet the requirements for quality. ISO 9001 is of relevance for organisations carrying out design, production, installation and servicing; the other Quality Assurance standards are applicable to organisations involved in production and installation (ISO 9002) and in final inspection (ISO 9003).

The following standards are Quality system guidance standards, i.e. standards giving advice on various aspect of quality assurance in a company:

- *ISO 9004: Quality management and quality system elements – guidelines.* It describes how to develop quality systems which apply Quality Management principles.
- *ISO 9004-2: Quality management and quality system elements – guidelines for services.* It describes how to develop quality systems which apply Quality Management principles to organisations supplying services (e.g. telecommunication services, maintenance, distribution).
- *ISO 9004-4: Quality management and quality system elements – guidelines for quality improvement.* This standard is a set of management guidelines for implementing continuos quality improvement within an organisation. The ways of adopting such guidelines depend upon factors related to the nature of the organisations.

8.3 Software Engineering Standards

In ISO/IEC, Joint Technical Committee number 1 (JTC1 Study Committee 7; SC7) has work under way related to process improvement and metrics for a process and product. It is cited here mainly for reference purposes. The most promising works of this committee are found in:

8.3.1 ISO/IEC 14528 Software Product Evaluation

This standard is planned for the following parts

- Part 1 General overview (a CD[1] exists)
- Part 2 Planning and management (a CD exists)
- Part 3 Process for Developers (a CD exists)
- Part 4 Process for Acquirers (a WD[2] exists)
- Part 5 Process for Evaluators (a CD exists)
- Part 6 Evaluation modules (a CD exists)

8.3.2 ISO 9126 Software Product Quality

The current version is issued in 1991. A revision is planned but its issue date is not set. This standard is planned for the following parts:

- Software quality characteristics and metrics – Part 1: Quality characteristics
- Software quality characteristics and metrics – Part 2: External metrics

[1] *CD: Committee Draft.*

[2] *WD: Working Draft.*

- Software quality characteristics and metrics – Part 3: Internal metrics.

ISO 9126 and ISO/IEC 14528 are planned to function as a whole; the former taking metrics into account whilst the latter examines the evaluation aspects.

8.3.3 ISO/IEC 12207 Software Life-cycle Processes

Issued 1995, and planned for revised issue in year 2000. In ISO 12207, the following processes are incorporated:

- Acquisition process (included in this is the system requirements specification)
- Supply process
- Development process
- Operation process
- Maintenance process.

In addition, there are requirements on the supporting processes; documentation, configuration management, quality assurance, verification, validation, joint review, audit, and problem solution. There are also requirements for the organisational processes; management, infrastructure, improvement, and training

Work is underway, with supplementary documents of ISO 12207 (WD for the time being)

- 12220-2, Life-Cycle Process
  – Software Configuration Management
- 12220-3, Life Cycle Process
  – Project Management
- 12220-4, Life Cycle Process
  – Quality Assurance
- 12220-5, Life Cycle Process
  – Verification & Validation
- 12220-6, Life Cycle Process
  – Formal Review and Audits.

8.3.4 ISO/IEC 12119 Information technology – Software packages – Quality requirements and testing

This standard was issued in 1994 and gives the requirements for “off the shelf” products. Topics covered include requirements on product description, user documentation and instructions for testing. The standard contains a useful bibliographical annex which gives an overview of basic standards in the areas of product conformance assessments and software technology.

9 SPICE

The SPICE project (Software Process Improvement and Capability dEtermination) is an international collaborative effort to develop a standard for software process assessment.

The purpose of SPICE is to provide a framework for the assessment of software processes.

The work done by the SPICE project is now launched as an international standard (technical report) by ISO/IEC JTC1 SC7 WG10.

The results of the project are:

- A “best-practice” model of software engineering and management
- A measurement framework for assessment against the best practice model
- A set of indicators to be used in the assessments (these indicators are on a “ISO 9000-3” level, rather than on the detailed metrics level)
- Training, experience and qualification requirements for assessors
- Guidance on how to use the results of process assessment for process improvement
- Guidance on how to use the results of process assessment for the purpose of capability determination.

These results are issued as a technical report (in 9 parts) by ISO/IEC JTC1.

The parts issued are:

- Software Process Assessment
  – Part 1: Concept and Introductory Guide
- Software Process Assessment
  – Part 2: A reference model for processes and process capability
- Software Process Assessment
  – Part 3: Conducting an assessment
- Software Process Assessment
  – Part 4: Guide to conducting assessments
- Software Process Assessment
  – Part 5: Construction, selection and use of assessment instruments and tools
- Software Process Assessment
  – Part 6: Qualification and training of assessors
- Software Process Assessment
  – Part 7: Guide for use in process improvement
- Software Process Assessment
  – Part 8: Guide for use in determining supplier process capability
- Software Process Assessment
  – Part 9: Vocabulary.

The potential usage of this material is as a base document for

- Supplier assessment
- Process improvement. This theme is important because today, many problems are encountered due to the fact that customers lack necessary competence to understand improvement work at their suppliers.

Information on current status and maturity of the SPICE products can be obtained through the reports on trials that have been executed. The general judgement seems to be still some hesitation towards the cost-effectiveness of SPICE. More information can be obtained through the Software Process Newsletter.

10 TRILLIUM

The Trillium Model[3] consists of the following chapters;

- Model Overview. The model is characterised by
  - A telecommunications orientation
  - Provides a customer focus
  - Provides a product perspective
  - Covers ISO, Bellcore[4], Malcolm Baldrige, IEEE and IEC standards
  - Includes technological maturity
- Implementation Guidelines
- Model Description. This defines the structure of the model and illustrates its relationships to industry standards.

[3] *Model for Telecom Product Development & Support Process Capability. ©Bell Canada, 1994 (Release 3).*

[4] *Bellcore TR-TWT-00179 and TR-TWT-001315 (early versions of IPQM and RQMS).*

- Essential Information About Trillium Practices
- Trillium Capability Areas, Road maps and Practices. A Road map is a set of related practices that focus on an organisational area or need, or a specific element within the product development process. This part of the Trillium Model defines practices in the following areas:
  - Organisational process capability
  - Human Resources Development and Management
  - Process (including process improvement)
  - Management
  - Quality System
  - Development Practices
  - Development Environment
  - Customer Support.

In total, 508 practices are identified. These practices are mainly in the form of references to practices defined in the underlying documentation. The value added by Trillium is therefore

- A profile of related practices for a special capability area
- Identifications of practices "deemed practical" (and indirectly, information about usage of the particular practice at least by Bell Canada).

Trillium is claimed by Bell Canada to be used to assess the product development and support capability of prospective and existing suppliers of telecommunications or information technology-based products. Further releases of the model may include some of the following: hardware development, manufacturing and service capability.

11 Conclusions

For businesses involved with telecommunications software development, knowledge of the standardisation initiatives in the area of software development is important. These initiatives will form the basis both for

- Requirements specifications, that are to be foreseen in the future.
- Supplier evaluation activities. Such activities will be of more systematic use in the future, in connection with following up of the suppliers. Supplier evaluation will also take its place in connection with qualification of suppliers before the contracts are placed.

For public network operators, knowledge of the standardisation initiatives is important to understand the way quality improvement should take place. It will also help in understanding the suppliers' way of thinking.

Acknowledgement

As this paper is based upon a study done in EURESCOM Project P619, it is a pleasure to acknowledge the contributions from colleagues in that project, for making valuable suggestions on the topics discussed. Especially an acknowledgement to colleagues from CSELT – Patrizia Bondi, Gianni Benso, Roberto de Bonis and Laura Marchisio – who contributed with valuable information on several of the topics discussed.

The Error Propagation Phenomenon – an introduction

BY BJARNE E. HELVIK

As a result of cooperation between the units of a distributed system, an error in one unit may cause errors in, and subsequent failures of, other physical separate units. Similarly, the same logical fault present in the copies of software in many units may be triggered by a condition spreading through the system under specific circumstances. This phenomenon is called error propagation and has a great impact on the dependability of distributed control and management systems, like the ones found in telecommunication. Nearly coincident failures in a number of network nodes, and at worst instability and severe network outages, may occur.

The prime objective of this paper is to give a brief introduction for the non specialist. As a background, a brief introduction to software reliability is given, with emphasis on the fault → error → failure process.

1 Introduction

This paper deals with error propagation. This phenomenon has great impact on the dependability of distributed computing systems. Because of the physical separation of the system units, there is a tendency to regard the failure process of the different units as independent. This is not generally true. As a result of cooperation between units, an error in one unit may cause an error in another unit, i.e. error propagation. See Figure 1 for an illustration. An error is, in this context, a deviation of the internal state of a unit from the correct one.

Our studies were triggered back in the middle of the 1980s by the change in switching system architecture towards physical distribution and at the same time a stronger logical linkage between the various physical separate control units. This trend in design of telecommunication control and management systems towards distribution of the functions on a large number of processors was first seen in the control part of switching systems as indicated in Figure 1. Cf. for instance System 12 (Alcatel), 5ESS (AT&T) and LINEA UT (Italtel), [1, 2 and 4]. Both distribution of functions among different processors and load sharing between processors which handle the same functions, are used. The effect of error propagation in this kind of system is demonstrated by clustering analysis in [3]. There is also a trend in telecommunication networks towards a closer logical coupling between the nodes of the network. This trend started with the more powerful signalling systems, for instance SS7 [5]. Hence, the current generation of telecommunication networks/systems is more susceptible to the effects of error propagation. This was dramatically exemplified on January 15, 1990 when the so-called *AT&T's 9-hour glitch* occurred [6]. During this glitch, the AT&T network could handle only half the normal volume of long distance and international traffic. The problem was caused by a software fault which caused a node failure if a new call message arrived immediately after a node was brought back to service after an error recovery. A minor failure activated this fault and the situation spread rapidly throughout the entire network. Since then several severe incidents caused by failures of signalling system number 7 has occurred, and this kind of failures has been thoroughly studied, see for instance [7]. By the foreseen introduction of a common distributed computing platform, cf. [8, 9], on which the service provision, control, management and administrative tasks of the network will be handled, telecommunications become more susceptible and vulnerable to error propagation. Propagating errors may severely reduce the dependability of the telecommunication infrastructure and in the worst case cause major outages.

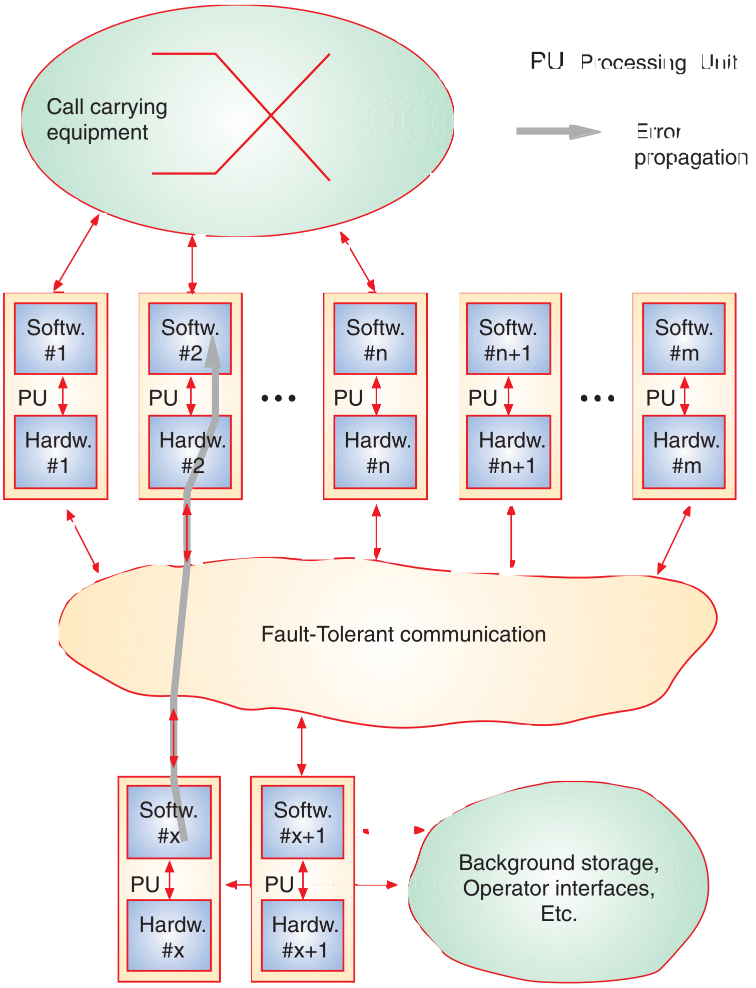


Figure 1 Sketch of a typical switching control system architecture with error propagation indicated

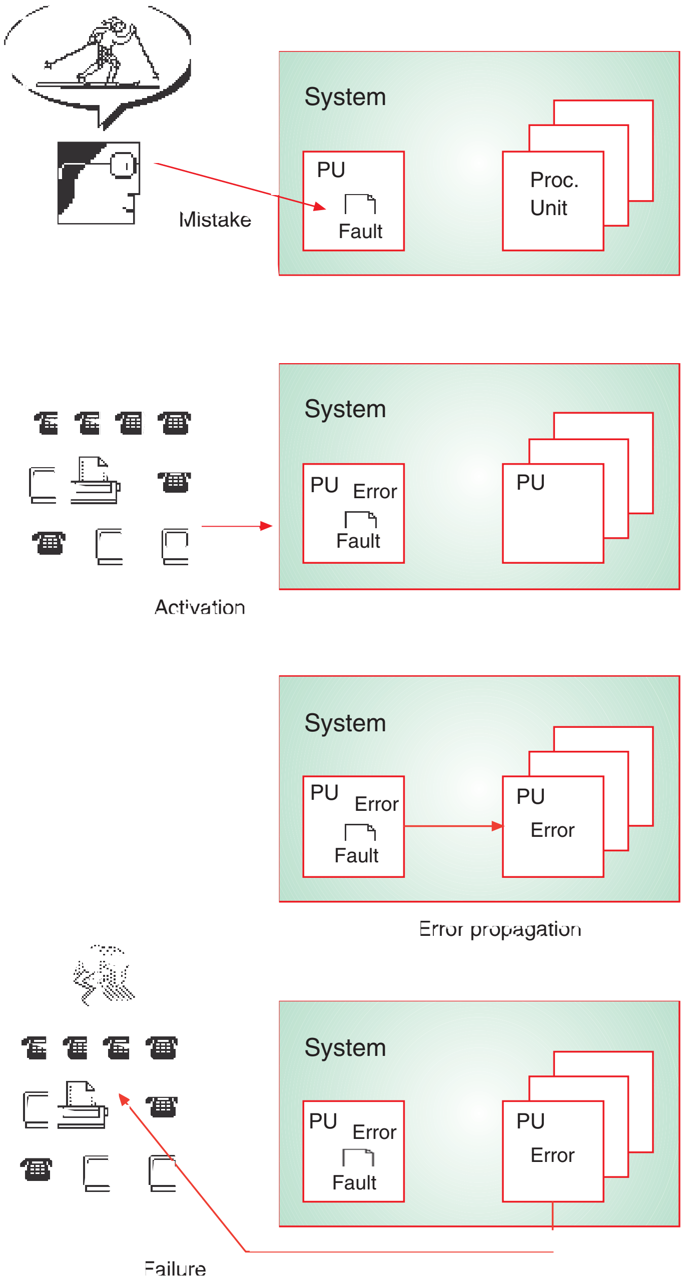


Figure 2 Sequence of events from logical mistake in system development to system failure

The research activity related to error propagation in distributed systems has mainly been directed towards the design of checkpoint – rollback recovery schemes, especially in distributed data bases, where the data consistency requirement is very high. In these systems it is important to be able to trace the error propagation path. Up till recently, most of the modelling effort has concentrated on this aspect. These approaches are very pessimistic since they assume that an error present in a process always will propagate to the cooperating processes by messages sent. However, it should be kept in mind that the technique of checkpointing and rollback recovery is, in most cases, not applicable to communication control and management systems due to their real time requirements, irreversibility of a large number of the events in the system and their size/complexity.

In spite of its importance, modelling and evaluation of the propagation process has received moderate attention [10, 11]. Outside the area of recovery block based software fault tolerance, there has, to the author's knowledge, not been published any systematic efforts towards design methodologies which cope with error propagation. However, the error propagation is taken into account in many designs. For instance, it is common to limit the error propagation by performing a check of the plausibility of incoming messages. More systematic means for reducing the error propagation between modules are incorporated in ESS 5 [12]. Approaches to reduce the error propagation are reported for the MARS system [13]. Outside the area of software fault tolerance based on recovery blocks and distributed data bases, there are, however, few results available, at least in the open research community, dealing with:

- The error propagation phenomenology
- Modelling and evaluation
- System design methodologies trying to reduce the propagation of errors.

The prime objective of the paper is to give an introduction to the area for the non specialist. In the next two sections, a brief introduction to software reliability and error propagation is given, before some concluding remarks are made.

2 Software failures

The theory of software reliability stems from the early 1970s, is still evolving, but is not yet a mature engineering discipline. The fact that software may cause unit and system failures, similar to failures caused by hardware faults, is still strange to some, since software does not deteriorate and physically fail in the same way as hardware. The software induced failures are often not considered, (for instance, the dependability specification of a system is often restricted to its hardware) or considered as inevitable. This section is therefore included to give a brief introduction to some basic concepts. For a state-of-the-art introduction to software reliability, see [14].

2.1 From mistake to failure

The sequence of events causing a system to fail because of a software fault is illustrated in Figure 2. We regard a "general" system made up of a set of semi autonomous processing units, PUs. The first event of the chain is a *mistake* during specification, design, coding or configuration of the actual system[1]. As a result of this mistake, a *fault* is embedded in the code or the configuration data of the system. It is common knowledge that faults are sought avoided by use of development methodologies and tools, quality assurance, testing, etc. Experience has shown, however, that in larger systems, faults will still be embedded when the system is put into operation. Often it is neither cost efficient nor feasible to try to correct these faults.

The next event takes place when one of these remaining faults are activated. The activation takes place when the faulty construct in the (software) system is exercised by the load[2] put on the system. The activation usually takes place when a rare and unforeseen combination of external input and internal state occurs. The activation of a fault results in an *error,* which is a deviation of the internal state of the process (processing unit) from the correct one. For instance, the internal mapping of the status of some external equipment is set to a wrong value. The behaviour of the processing unit does not (yet) deviate from the correct one.

During the cooperation between processes (PUs) the error may be propagated. For instance, the wrong status of the external equipment, mentioned above, is transferred to another processing unit. Error propagation is discussed further in Section 3.

Sooner or later one of the errors in the system is likely to cause a *failure* of one or more PUs. A failure is a deviation of a unit's or a system's behaviour from its specified behaviour[3]. Let us say that the wrong status mentioned above is used by the receiving processing unit and this causes this unit and the external equipment to halt. Propagation of errors within the system before they cause a PU or system failure is discussed in Section 3. Further discussion and definitions of concepts and terms may be found in [15] and [16]. (Note that in this paper [15] is used.)

Similarly to the activation of faults, the error propagation and the process where errors emerge as failures depend on the load of the system.

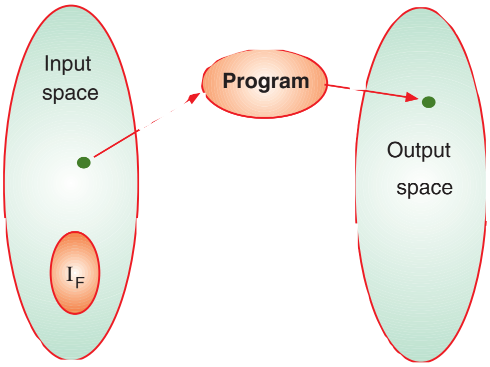


Figure 3 Simple Input Output model of software failure

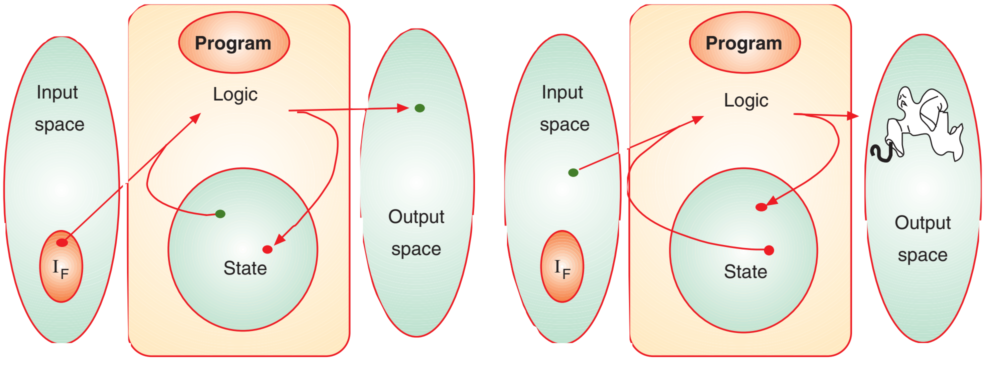
2.2 The influence of the environment

The importance of the load offered to the system is evident from the previous section. The stochastic behaviour of the load causes a stochastic software failure pattern, even though the consequences of a given fault are deterministic for a given sequence of inputs to the system. This may be illustrated by the "Input Output model" shown in Figure 3. The software is regarded as 'a program' which transforms values from an input space into results in an output space. Some input values will unveil logical faults in the program and give an incorrect or no result. In the model, these input values form the subspace I_F. The load offered to the system will determine the input values, and when this load has a stochastic behaviour the failure process will also be stochastic.

The model of Figure 3 is too simplistic for communication software. It is not only the current input, but also the internal state of the system that contributes to the activation of a fault. Hence, also the input previous to the activation, which forced the system into the current internal state, is relevant. This leads to the extension of the "Input Output model" shown in Figure 4. In this model the 'program' is defined by its logic and internal state. Corresponding to the discussion in Section 2.1, there will first be an activation of the logical fault which results in an error in the internal state. Later, this error may be propagated. The propagation may be regarded in two ways: a) when the 'program' models one process/PU, the output space is a part of the input space of other programs, or b) when the 'program' models the entire software system, the corruption of the internal state space becomes worse. As a consequence of the erroneous internal state space and an input (from inside or outside I_F) a failure (incorrect or missing result to the environment) may occur.

It is convenient to regard the following two properties of the load separately:

Load mix / user profile. This property reflects the traffic and service mix offered the system by its users (subscribers) including the user behaviour, the composition of operation and maintenance activities, etc. We may regard all the inputs to the system as coming from an input space. The load



a) Fault activation

b) Error propagation and failure

Figure 4 Input Output model of failure of continuously running software with internal state

[1] *Keeping strictly to the nomenclature of [15], this mistake is a failure of the development and installation process. The term mistake is used to avoid confusion with failure of the system.*

[2] *Load should be understood in a broad sense, also including operation and maintenance activities.*

[3] *With this definition, error propagation is an undetected (not recognized) failure of the source PU.*

mix / user profile may be regarded as the distribution of the choice of inputs from various parts of this space.

Load intensity / traffic. This property is the frequency with which inputs are offered according to the load mix / user profile, for instance the number of call attempts per hour.

The failure rate varies with the load intensity [17, 18 and 19]. It may also vary from one installation, e.g. an SPC exchange, to another for the same load intensity and for similar system configurations because of differences in traffic mix and operation. Similarly, large shifts in the failure rate (fault detection rate) is found between various test phases of a system.

2.3 Consequences

The consequences of faults and errors with respect to the system, its modules or its functions, depend on the design of the system. There are three main design philosophies:

Fault intolerant systems are unable to handle faults and errors during operation. Dependability of these systems is sought achieved solely by means trying to prevent faults, e.g. conservative design, derating and quality assurance. If a logical fault is left in the system or a physical (hardware) fault occurs, the system will fail some time after this fault is activated.

Fault tolerant systems are on the other hand designed to detect and handle errors during operation. Hence, a fault in the system will have none or a limited effect on its operation. Fault tolerance of permanent physical faults also includes location and isolation of the fault to a module, initiation of repair and reinitialization and inclusion of the module into the system after the repair.

Fail safe systems[4] are designed to enter a safe and stable state as soon as an error is detected, without handling the error. This principle is applied in safety related systems, like nuclear power plants, where it is more important to avoid accidents than to keep the system operational. Fail safe operation is uncommon in telecommunication systems.

Various design philosophies may be applied to the different parts of the system. For instance, in SPC systems the part of the system serving only a single subscriber is usually fault intolerant, while the rest, serving eight or more subscribers, tolerates hardware faults. Similarly, some functions of the system may be fault tolerant or fail safe, while others are not. Telecommunication systems have several intact defect criteria which relate to the effect of the failure, e.g. the number of subscribers or the offered traffic affected, the services or facilities affected and the duration of the failure.

Telecommunication systems used for public switching are designed to tolerate hardware faults which may have a significant influence on the system performance. Techniques for making the systems tolerant to logical (software) faults, like recovery blocks and N version programming [20, 21], are not used, for cost, efficiency and maintainability reasons, and because these techniques are still immature. Some robustness towards errors caused by logical faults are, however, built into the systems. For instance, audit routines, isolation against propagating faults and minimum reinitialization after a process or unit failure. By minimum reinitialization is meant that only the parts of the system suspected to contain an error are reinitialized. The objective is that the correction of an error will affect only a subset of the traffic and the functions performed. These techniques combined with load sharing may limit the consequences and provide some tolerance towards logical faults.

3 Error propagation

3.1 Error propagation mechanisms

The error propagation mechanisms themselves are rather poorly understood. Some investigations are done of the propagation of errors caused by hardware. These investigations are based on experiments with hardware fault injectors on the circuit level, see for instance [22, 23 and 24]. To the author's knowledge no systematic investigation is published for errors stemming from software faults. An injection experiment simulating software induced errors has been done [25]. In this experiment, however, the error to failure process was accelerated.

The mechanisms which cause propagation of errors resulting from software faults may depend heavily on the system design. For instance, in a system made up of semi-autonomous processes which communicate by message passing (typically found in telecommunication control systems), the mechanisms may be different from those in a system with a strictly hierarchical organization. Examples of types of error propagation from one process (PU) to another are:

- Forwarding, where incorrect data or sequence information are sent from one process to another
- Responses, where a process gives no or a wrong response to a request from another process
- Addressing, a message is sent to a wrong destination
- Timing, because of an error, a process transmits/forwards a message so late that the process which should have received the message makes an erroneous action
- Cooperation, the protocol between two processes does not work properly (e.g. resulting in a deadlock)
- Resource handling, where a common resource is not properly initialized, reserved before use or released after use.

A variant of the sequence of events described in Section 2.1 may also occur. Replica of identical code may be placed in a number of different PUs. Hence, if there is a logical fault in this code, it may be activated in a number of PUs almost simultaneously if the same "trigger" condition arises. This may be the case if the trigger condition propagates as a consequence of an initial event. The AT&T 9-hour glitch [6] mentioned in the introduction is an example of this kind of propagation.

3.2 Error life cycle

As a basis for the discussion, the error propagation is illustrated in Figure 5 as a cause consequence diagram. This figure enhances Figure 2. As pointed out previously, some faults are most likely still embedded in the code and/or data of a system when it is put into operation. These faults may be activated and cause errors in the internal state of a processing unit. The error will affect the transient data in the unit. It may also affect the code and permanent data, if this inform-

[4] *These systems are sometimes called fail stop systems.*

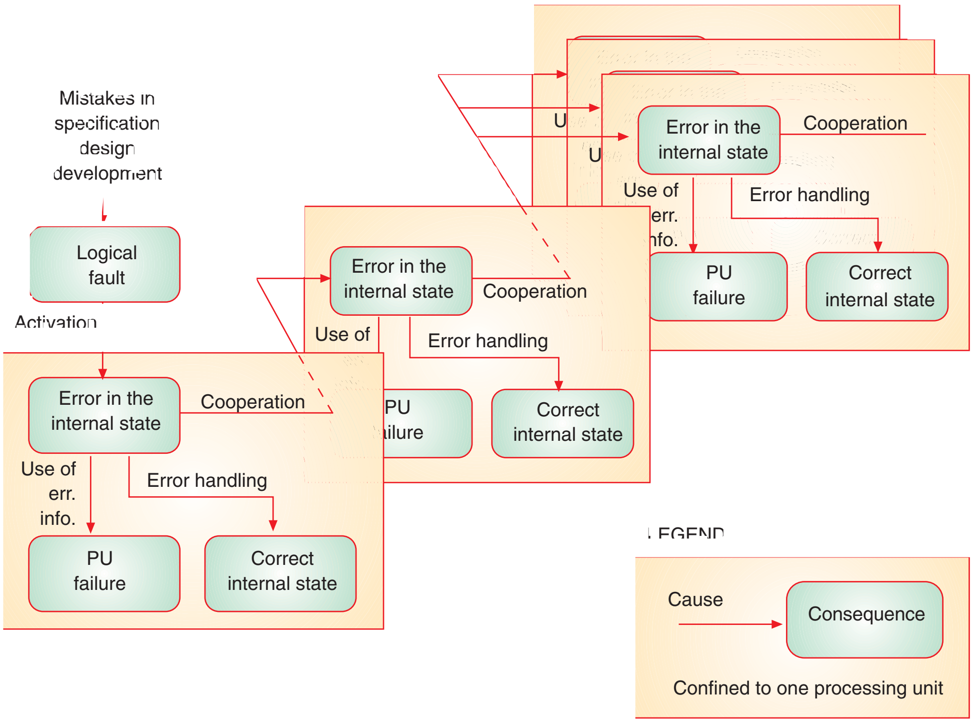


Figure 5 Cause consequence diagram of error propagation in a distributed system

ation is unprotected. When the type of error is discussed, we have found it convenient to split the process data into the following categories [26]:

- *Control data,* which contains information about the further sequence of the processing. For instance, the state of an EFSM (Extended Finite State Machine) is set to a wrong value, or, in a conventionally structured program, a flag is set to wrong value.
- *Resource data,* which contains information about internal resources in the system, e.g. the status of a message buffer, or resources in its environment, e.g. the number of free channels in a direction.
- *Coordinating data,* which contains information about timing and synchronization in the system. For instance a wrong internal clock or wrong timing information about external devices.
- *Other data,* e.g. results from calculations.

One error may affect many data items, and the items affected may be a combination of the types listed above.

Subsequent use of erroneous information may cause a failure of the PU. For instance, it may enter an infinite loop, halt or try to violate (hardware) protection mechanisms of the unit. (This failure will initiate an action trying to remedy the error, typically a restart or a reload of the unit.)

Alternatively, the error may be detected and corrected without disturbing the operation of the system, or be corrected with an acceptable disturbance. For instance, audit routines may detect an improperly linked list. The “correctness” of this list may either be restored by using redundant information to rebuild the list, or by re-initiating (parts of) the list and accepting the loss of some transient data and the corresponding calls in the set-up phase. The error may also be rectified without hampering the operation of the system by pure luck. For instance, a pointer with an incorrect value may be updated and set to a correct value without being used first.

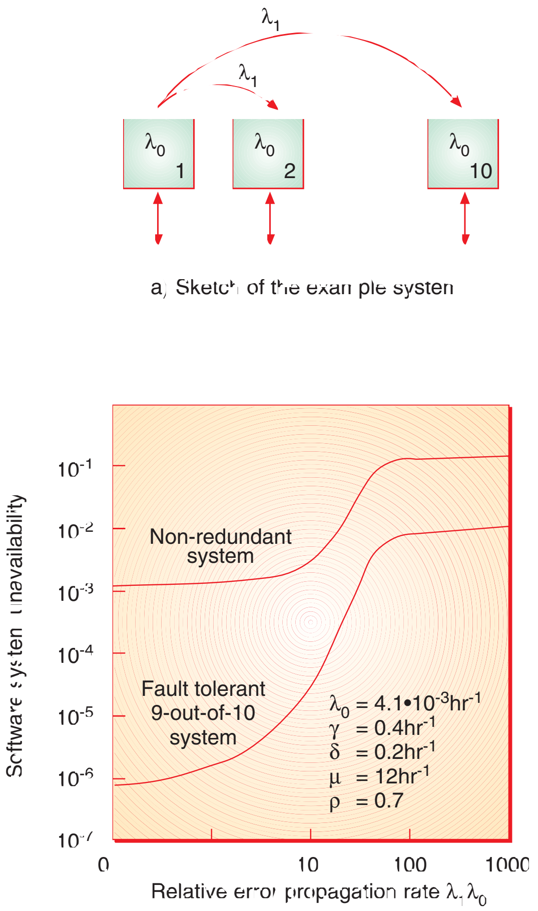
A third possibility is that the error propagates to another unit before it causes a failure or is corrected in the first unit. See Figures 2 and 5. Error propagation takes place during cooperation between processes of the two units. The mechanisms are outlined in Section 3.1. The error may propagate from one unit to many other units. In the same way, an imported error may propagate further, as sketched in Figure 5. Two main factors influencing the propagation are identified:

- The intensity in the cooperation between the processes. This factor is expected to be proportional to the load / offered traffic to the unit.
- The logical linkage between pairs of processes, i.e. how much the internal state of one process depends on the internal state of the other process.

The error latency is the time from a data item is set to an erroneous value until a failure is recognized or the error is corrected. The error latency is very important when we regard error propagation. The longer the latency, the larger is the probability of error propagation. The latency will also influence the failure pattern of the system and influence its service characteristics. There is no firm knowledge about the latency of errors caused by software faults, but there are indications that the error latency is in the range of half an hour to some hours [19] and even days [3]. See also Figure 7. The error latency is also likely to depend on the load of the system and the type of system. It is interesting to note that similar latency time constants are seen for random changes of memory bits [27].

3.3 Influence on system performance

It is important to be aware of the consequences of error propagation on system performance. An increase in the PU and system failure rate and a decrease in the availability, are obvious. This is illustrated in Figure 6, which shows results from an example system. A homogeneous system of ten identical units is regarded. The internal fault activation rate of each unit is denoted λ_0. As presented in Section 2.1, a fault activation takes place when a logical fault causes an error in the internal state of the unit. The rate that this error, when present, propagates with to another unit, is denoted λ_1. See the box “Modelling of error propagation” for further details. Figure 6.b shows the unavailability of the system as a function of the relative error propagation, i.e. λ_1 / λ_0. The *upper curve* shows the unavailability when all units are required for the system to be operative. It is seen that the effect of the error propagation is small as long as the propagation rate is smaller or similar to the activation rate. When the propagation rate becomes ten times or more the



a) Sketch of the example system

b) System unavailability as a function of relative error propagation

Figure 6 Effect of error propagation on unavailability of an example system [10]. (For a presentation of model parameters see box "Modelling of error propagation".)

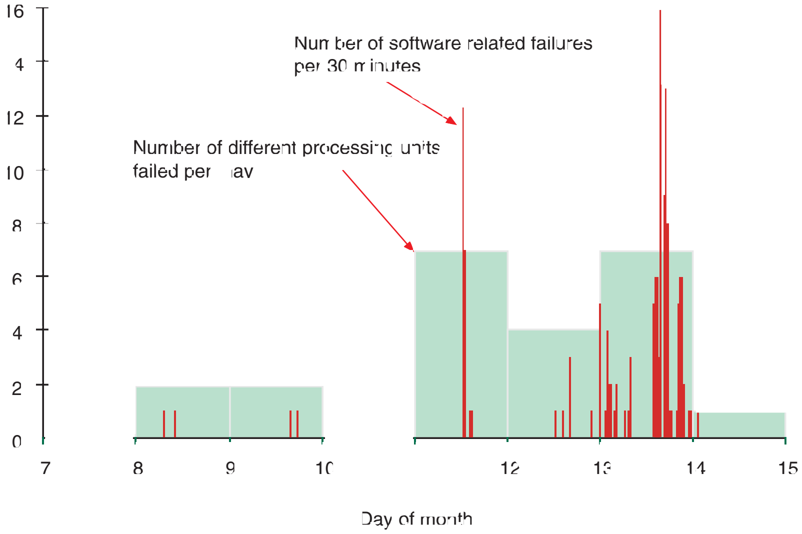


Figure 7 Excerpt of the failure log of a distributed communication control system

activation rate, the unavailability becomes rather poor. The *lower curve* shows the unavailability when the system tolerates the failure of one unit, i.e. a nine out of ten configuration. It is seen that in this case, the system is much more sensitive even to a low error propagation and the impact of an increasing propagation rate is larger. For a discussion of this effect see [10].

Note, however, that the failure pattern is no longer Poissonian, i.e. failures do not occur independently of each other. Hence, the influence on system performance of a given failure rate differs from what we are accustomed to. The most important effect on the PU (and system) failure pattern are:

- Failures tend to occur in severe bursts, i.e. if one failure is experienced, the next is likely to occur rather soon.
- Many PUs are involved in a failure burst. If one processing unit (starts to) fail, there is an increasing probability of failure of the cooperating units.

The higher the error propagation, the larger is the burstiness of the failure pattern and the more units are failing during a burst. The burstiness is also influenced by the error handling abilities of the PUs. Figure 7 shows a sample from the software failure pattern of a large distributed communication control system. It is seen that during the first four days we have a normal operational mode with few failures. During the 11th through the 13th day severe bursts of failures occur, affecting several processing units. See also [3] for an analysis of the effect of error propagation on system behaviour.

The burstiness of the failure pattern has made studies of the transient/dynamic properties of systems more important. The distribution of the time between failures and the interval availability, where the interval is short compared to a year, e.g. a day or a week, gives extra information about the burstiness. This information is important when the service and operational characteristics of the system are to be valuated. See the information box on "Modelling of error propagation" for a brief discussion on tools for evaluation of systems with error propagation with respect to these measures.

Note also that due to error propagation, distributed systems may exhibit a bistable (unstable) dependability behaviour. That is, the system has a *good* operational mode where it never or seldom fails because of logical faults, and a *poor* operational mode where it fails frequently. A system is potentially unstable if the error propagation rate may become higher than the error handling rate. For the use of clustering analysis to identify the operational modes of a system see [3].

In telecommunication systems, the penalty of a failure is often considered to increase more than proportional with the consequences in terms of affected subscribers or traffic lost. For instance, if N subscribers are affected, the penalty is N^{α} where $\alpha > 1$. Hence, clustered failures are more harming than the same number of "random" faults, since more subscribers/traffic tends to be affected at (nearly) the same time, i.e. the risk associated with the system is higher.

4 Concluding remarks

In this paper the error propagation phenomenon is introduced. The presentation may be summarized in the cartoon-like Figure 9. Error propagation is identified as the source of a number of important unsolved problems concerning the dependability of telecommunication systems and distributed computing systems in general. The discussion of this paper leads to the following remarks con-

Modelling of error propagation

Both system level modelling and process level modelling should be dealt with. System level modelling is concerned with the development of realistic dependability models, how to best characterize the dependability properties of a system and tools for system evaluation.

- A system level modelling procedure for telecommunication control systems has been developed. The aspects of this model dealing with error propagation is later reported in [10]. The basic PU model is shown in Figure 8, where an excited state is introduced to model the presence of an error. The propagation of errors is modelled by a linear relationship

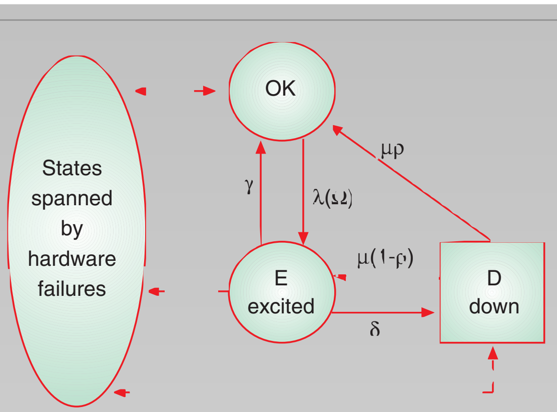
  $$\lambda(\underline{\Omega}) = \lambda_0 + \sum_{j>0} \# E_j \cdot \lambda_j$$

  where λ_0 is the internal fault activation rate of the PU, $\# E_j$ is the number of excited PUs in group j, and λ_j is the error propagation rate from group j.

- Due to the bursty failure pattern caused by error propagation, cf. Figure 6, the dynamic/transient behaviour of the system has to be determined and studied. For instance, the conditional distribution of time between failures instead of the MTBF, and the distribution of the interval availability instead of the steady state availability. The dependability characteristics should be related to a number of service levels. These levels are given by the number of subscribers, the traffic and/or the services affected by a failure.

- Tools are needed for dependability modelling and evaluation. A modelling methodology specialized for the needs of distributed telecommunication control systems has been developed. The methodology is based on local state diagrams with interdependencies, as indicated in Figure 8, and with synchronization primitives to include events causing simultaneous transition in two or more local diagrams. Introduction of various scheduling mechanisms for simpler modelling of resource contention is also considered. These diagrams are denoted SLSD (Synchronized Local State Diagrams). To be able to evaluate large and complex systems, experimental tools are developed for:

  - Building and handling models of large systems. The modelling framework is based on a hierarchical structuring combined with an object oriented approach.
  - Automatic generation of transition matrices from the local diagrams.
  - Direct simulation.



$\lambda(\Omega)$ Occurrence rate of errors in the internal state of the PU, including the effect of error propagation.

Ω State of the entire system

γ The rate by which errors in the internal state of the PU are removed or "dies out" without interruption of processing

δ The manifestation rate of an error in the internal state of the PU

μ The average delay from an error in the internal state of the PU is detected (software failure) until useful computation is resumed

ρ The relative number of software failures, which is remedied by a complete restoration of the internal state (E.g a reload)

$1-\rho)$ The relative number of software failures, which is remedied without a complete restoration of the internal state

Figure 8 Subset of the Markov model of one processing unit (PU), showing the state space spanned by software errors and failures

These tools are reported in [28]. The most promising technique for analysis of large systems is a direct simulation of the SLSD. To obtain sufficiently accurate estimates, with a reasonable CPU time, importance sampling must be included [29].

At the process level, the error propagation mechanisms, cf. Section 3.1, the error handling and the failure processes are modelled to determine the error propagation rates λ_j, the failure rates δ, and the error handling parameters γ, μ and p indicated in Figure 8. The objective is to provide improved input to the system level evaluation and to be able to evaluate various system designs and design methodologies with respect to their error detection and handling properties, as well as their error propagation and latency characteristics. Some preliminary models are obtained but not validated.

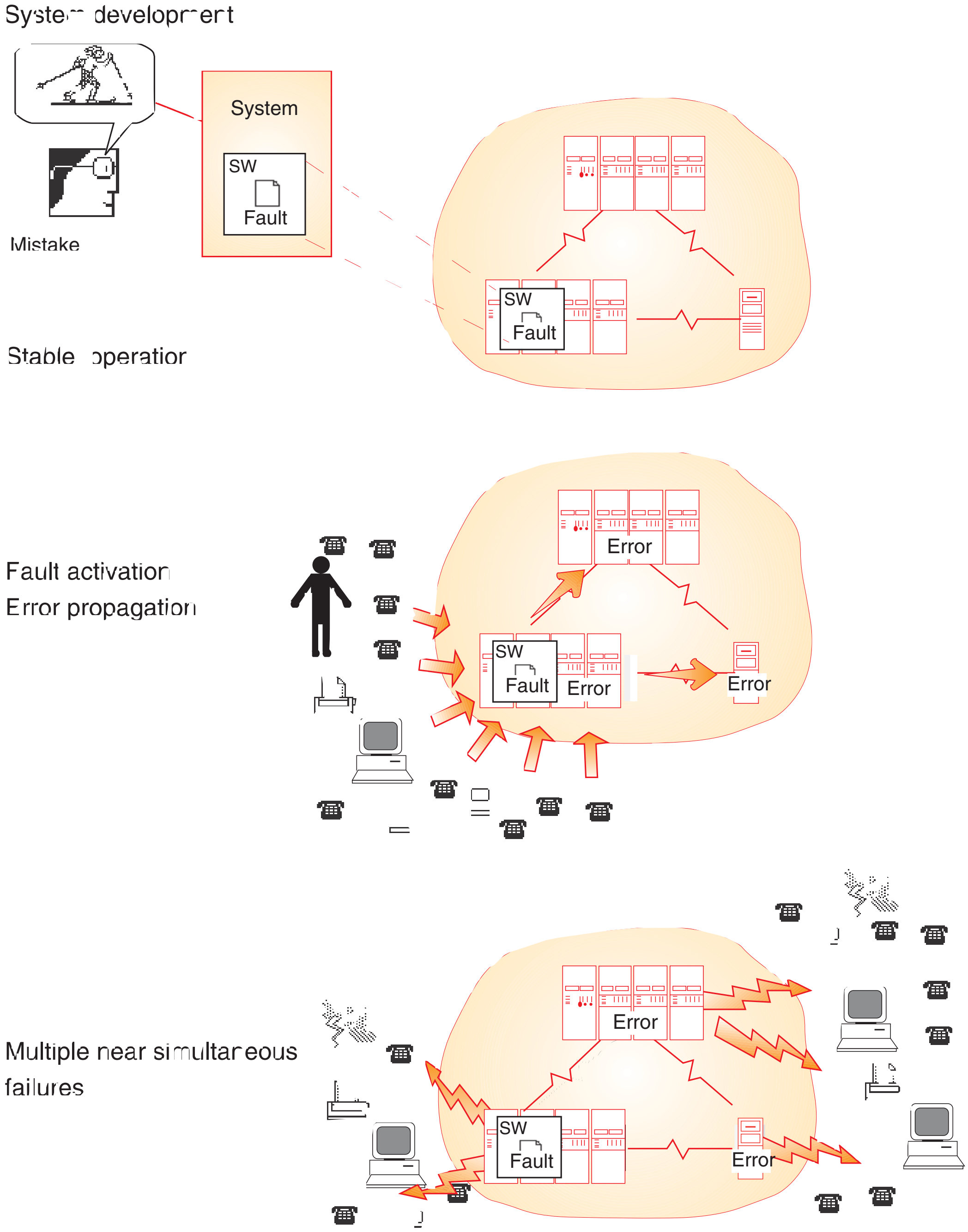


Figure 9 Error propagation in telecommunication systems

cerning telecommunication systems and networks:

- The effect of error propagation should be accounted for during specification, investigations (evaluation of candidates) and acceptance of new systems and architectures.
- Use of short term / dynamic properties in system evaluation and monitoring, e.g. the interval availability over a day, week and month, is necessary to account for the bursty failure pattern.
- When a new system is considered, its error isolation properties and robustness towards error propagation should be investigated.
- Error propagation tends to cause a more bursty failure pattern. Attention should be paid to the risk caused by error propagation, since the failure-effect increases by the greater simultaneity of unit failures.

There is a trend within telecommunication system architecture towards introduction of a common distributed computing platform for provision of services, network control and management as well as the handling of some administrative tasks. This introduces a far tighter logical coupling between various network elements and functions than in the current systems/networks. Telecommunications systems will therefore become more susceptible to error propagation and the consequences may be severe.

References

1 ITT. System 1240 digital exchange. *Electrical Communication,* 59, (1/2), 1985.

2 Smith, W B, and Andrews, F T Jr. No. 5 ESS : overview. In: *International Switching Symposium (ISS),* Montreal, Canada, Sept. 1981.

3 Helvik, B E, Gylterud, S A. Identification of operational modes of distributed systems by cluster analysis. *Telektronikk,* 93, (1), 1997.

4 Briccoli, A et al. Performance design of a distributed switching system. In: *Proc. 12th International Teletraffic Conference (ITC-12), Torino, Italy,* 1–8 June 1988, vol 2, p. 2.1A.2.1 –10.

5 CCITT. *Blue Book : Signalling System No. 7,* vol VI. Geneva, ITU, 1988. (Fascicle VI.7, .8 and .9, Recommendations Q.700 –.716, .721 – .766, .771 – .795.)

6 Fitzgerald, K. Vulnerability exposed in AT&T's 9-hour glitch. *The Institute,* 14, (3), 1990, p. 1 and 6.

7 Schütte, W (ed.). *Methods for fault handling and prevention in signalling system No. 7,* EURESCOM P307 (Reliability engineering). Deliverable D 1, Oct 1994.

8 Barr, W J, Boyd, T, Inoue, Y. The TINA initiative. *IEEE Communication Magazine,* 31, (3), 1993, 70–77.

9 Appledorn, M, Kung, R, Saracco, R. TMN + IN = TINA. *IEEE Communication Magazine,* 31, (3), 1993, 78–85.

10 Helvik, B E. Modelling the influence of unreliable software in distributed computing systems. In: *Proc. 18th International Symposium on Fault-Tolerant Computing,* 27–30 June 1988, 136–141.

11 Shin, K G, Lin, T-H. Modelling and measurement of error propagation in a multimodule computing system. *IEEE trans. on Computers,* C-37, (9), 1988, 1053–1066.

12 Bellino, A. *Introduction to the 5 ESS switch,* 23 June 1989. Technical visit at AT&T Labs, Naperville.

13 Damm, A. Fault-tolerant distributed real-time systems : MARS. In: *Fault-tolerant computing systems.* F. Belli, F, Gerke, W (eds.). Springer-Verlag, September 1987, 362–372.

14 Lyu, M R. (ed.). *Handbook of software reliability engineering.* McGraw-Hill/IEEE Comp. Soc. Press, 1996.

15 Laprie, J-C (ed.). Dependability : basic concepts and associated terminology. *Dependable computing and Fault Tolerant Systems, vol 5,* Springer, 1992.

16 ITU-T. *Terms and definitions related to quality of service and network performance including dependability,* Recommendation E.800, August 1994.

17 Iyer, R K, Rossetti, D J, Hsueh, M C. Measurement and modeling of computer reliability as affected by system activity. *ACM Trans. on Computer Systems,* August 1986, 214–237.

18 Bredrup, E et al. The activity-dependent failure intensity of SPC systems : some empirical results. *IEEE Journal on Selected Areas in Communication,* SAC-4, (7), 1986, 1052–1059.

19 Helvik, B E, Swensen, A R. Modelling of clustering effects in point processes. An application to failures in SPC-systems. *Scand. Journ. of Statist.,* 14, 1987, 57–66.

20 Randell, B. System structure for software fault tolerance. *IEEE Trans. on Software Engineering,* 1, (2), 1975, 220–232.

21 Avizienis, A. The N-version approach to fault-tolerant software. *IEEE Trans. on Software Engineering,* SE-11, (12), 1985, 1491–1501.

22 Segall, Z and al. FIAT : fault injection based automated testing environment. In: *Proc. 18th International Symposium on Fault-Tolerant Computing,* 27–30 June 1988, 102–107.

23 Arlat, J, Crouzet, Y, Laprie, J-C. Fault injection for dependability validation of fault-tolerant computing systems. In: *Proc. 19th International Symposium on Fault-Tolerant Computing,* 21–23 June 1989, 348–355.

24 Goswami, K K, Iyer, R K. Simulation of software behaviour under hardware faults. In: *Proc. 23rd International Symposium on Fault-Tolerant Computing,* 1993, 218 – 227.

25 Chillarege, R, Bowen, N S. Understanding large system failures : a fault injection experiment. In: *Proc. 19th International Symposium on Fault-Tolerant Computing,* 21–23 June 1989, 356–363.

26 Stålhane, T, Myrstad, T. *Analysis of errors in telecommunication systems.* Trondheim, ELAB, 1990. (ELAB-RUNIT report STF40 A90097.) (In Norwegian.)

27 Chillarege, R, Iyer, R K. Fault latency in the memory : an experimental study on VAX 11/780. In: *Proc. 16th International Symposium on Fault-Tolerant Computing,* 1–4 July 1986, 258–263.

28 Heegaard, P E, Helvik, B E, Gotaas, E. Synchronized local state diagrams (SLSD) : description of a dependability modelling method under development. *Safety of Computer and Control Systems (SAFECOMP 91),* 30 Oct – 1 Nov 1991, 83–87.

29 Heegaard, P E, Helvik, B E. Dependability simulation of systems modelled by local state diagrams : application of importance sampling. *Nordic Seminar on Dependable Computing Systems 1992 (NSDCS'92),* 19–21 August 1992.

Identification of operational modes of distributed systems by cluster analysis

BY BJARNE E. HELVIK AND SVEN ARNE GYLTERUD

The units of a distributed system tend to fail collectively, and the failure process is far more bursty than a Poisson process. This is due to the logical coupling between units and incomplete error recovery. Observations of systems indicate that they may have different operational modes, e.g. "good" modes where only few sporadic failures occur and "bad" modes where many units fail repeatedly within a short time interval. The latter modes severely reduce the service offered. This article discusses how such modes may be identified by cluster analysis from the operational log of a system. Furthermore, it is illustrated how the mode-changing behaviour of the system may be described by a state model. The principles are illustrated, and some technicalities concerning the cluster analysis are settled by analysing 19 months of the operational log of a large distributed system. Characteristic operational modes are found. The corresponding state model is tested with respect to Markov properties.

1 Introduction

In the telecommunication network, there is an ongoing introduction of distributed control systems. Examples of such systems in today's network are the signalling system number 7 (SS7) and the control of switches. There is a tendency towards letting a single system control an increasingly larger part of the network functionality, for instance, in the Intelligent Network (IN), e.g. [21] and [17] and for the Telecommunication Management Networks (TMN), e.g. [31]. The dependability and stability of these systems are of the utmost importance.

When these properties of a distributed system are assessed, independence between failures of the different units are often assumed. This may be true for hardware failures, However, the different units are tightly logically coupled, may have parts of the software in common, and may simultaneously be exposed to similar operational conditions. Both theoretical work and experience have shown that this may cause errors to propagate between units, causing correlated failures and instability ([13], [9], [35]). This phenomenon, called error propagation, has great consequences for the dependability of distributed control systems.

This article addresses the dependencies between the units of a distributed control system and its objective is to develop a methodology where, from operational data collected from the system,

- An identification of typical operational modes of the system is enabled. In some of these, units fail nearly simultaneously and the systems have an insufficient performance,
- The behaviour where units seem to fail collectively are recognized,
- The further development of an undesired system condition / operational mode is predicted, e.g. what is the probability of the current erratic behaviour of unit A for spreading to other units of the system, and what is the probability for this behaviour to cease.

The methodology is based on a pre-processing and scaling of the operational data before a cluster analysis is carried out to identify typical operational modes. Next, a state diagram type of the model of the system is established. These issues are dealt with in Section 2. Example results from the application of the methodology on 19 months of operational data from a large distributed control system is presented in Section 3, before the article is concluded in Section 4.

Before starting the development of the methodology, it should be related to other approaches for extracting information from operational data. They may be divided into two classes. In the first class, an analysis of series of failure related events and an identification of event patterns are carried out ([3], [16], [18], [23], [24], [37], [38], [39]). The objective is usually prediction of coming system failures, e.g. disk crashes. This approach regards the system as a single entity, while our objective is to identify interrelations between system units. In the second class, the rule based approach, filtering of information and identification of the cause of the failure(s) from a large number of alarms is the objective ([1], [4], [11], [10], [30], [32], [42]). It is based on already available (expert) knowledge of the system and its behaviour. The objective of this article is more towards establishing such knowledge.

2 Methodology

The developed method is based on post-processing of operational data with focus on identifying operational modes of the investigated system. A detailed presentation is found in [12].

2.1 Outline

The presentation of the methodology is divided into four parts, starting with a presentation of *operational data* to identify typical kinds and contents of operational data, and to indicate the type of information to be extracted, i.e. operational modes of the system. *Identification of periods with similar dependability behaviour* is the main methodological objective. This is reached by transformation of the data and application of cluster analysis. Techniques for *validation, presentation and post-processing of the results* are demonstrated. An example of use of the results by calculating *transitions between operational modes* illustrates the last part of the method.

2.2 Operational data

Most of today's telecommunication equipment produces powerful logs of operational data. A common feature/problem with these logs is the enormous amount of data with well hidden information.

- What kind of operational data exist?
- What information is contained in operational data?

Typical contents of a log-message (in the dependability context) may be date and time, network address of the unit producing the message, fault type and error class diagnosed by the unit itself, and recovery action taken. These log-messages have a unique format for the unit producing the message. In a distributed system, single units may produce individual logs, and procedures for collecting, converting and merging of these to one common format have to be carried out.

The complexity of this process is a limiting factor to pertain consistent log-messages, number of nodes to collect data from and reliability of the collected operational data. The introduction of a TMN based operation and maintenance system may remove this obstacle.

2.2.1 Operational modes

The information hidden in the operational data may be stated as *operational* *modes*. Operational modes are recognized by periods where the *simultaneous* failure pattern of the system units has similar characteristics. Hence,

- Is it feasible to find dependent or similar behaviour among different units producing individual failure logs?
- Is it feasible to identify periodic or rare failure modes in a system?

To illustrate this, a log with a varying number of log-messages within the same time intervals and an example of a corresponding identification of operational modes is shown in Figure 1.

It is also important to note that some failure patterns are *extreme* patterns that rarely appear more than once. One such pattern may not be characterized as a *mode,* but several such patterns may be collected into one *extreme* mode.

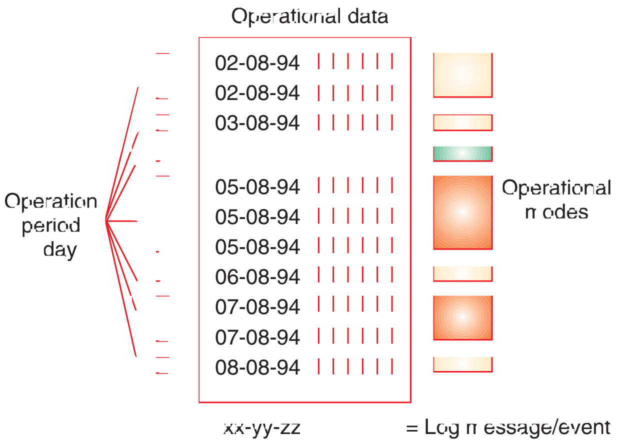


Figure 1 Operational data

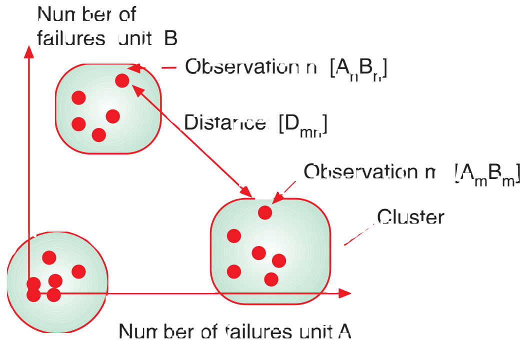

Number of failures unit B
Observation n [$A_n B_n$]
Distance [D_{mn}]
Observation m [$A_m B_m$]
Cluster
Number of failures unit A

Figure 2 Illustration of clusters in a system with two units, i.e. a two dimensional space

2.3 Identification of periods with similar dependability behaviour

The main part of the methodology is concerned with a method for identification of periods with similar dependability behaviour by use of cluster analysis. According to the principles of cluster analysis, three main steps are carried out, 1) pre-processing of the operational data, 2) application of cluster techniques to perform the actual clustering, and 3) validation of the cluster results.

As a general rule, alternative procedures should be investigated for each of these steps. This has been carried out for real data sets [12]. However, in this article, we will focus on the chosen solutions and the rationales for these.

2.3.1 Cluster analysis

Cluster analysis is a well-known technique for classification of data, see for instance [28] and [19]. The technique is mainly used within topics related to biology, chemistry and pattern recognition.

We will describe cluster analysis in the following steps, also with a reference to Figure 2.

- **Data matrix**
  The raw data to be analysed is processed and transformed into a data matrix where each row represents an *object* and each column represents *one characteristic* of this object. In our case, an object is an *observation* of the system during one operation period, say a specified number of hours. The characteristics of the object may for instance be the number of failures for given system units during that operation period. Each object (row) will now represent one point in the n-dimensional space spanned by the characteristics.
- **Distance**
  The representation and calculation of the distance between the objects is the most important step to make clusters which reflect the purpose of the analysis. The complete distance measure consists of *scaling* and *calculation,* [36]. To be able to establish physically motivated clusters, the *distance* between the objects must be represented in a numerical way which reflects the physical "distance" between the objects. In our case the distance represents the difference in system behaviour between observation periods of the system. Distance calculation ends up with a distance matrix representing the distance between all objects (observations).
- **Cluster algorithms**
  A huge number of different algorithms are available to perform the actual *clustering* of the data, see e.g. [28], [19], [34], and [22]. An optimal algorithm is one which in the most optimal way identifies clusters, given the actual set of input data. Referring to Figure 2, this means points closely located. The algorithms, operating on the distance matrix, use different clustering criteria to build the clusters. In our case, the algorithms shall identify closely *located* observations of the system.
- **Cluster validation**
  However, it remains to be validated whether the identified clusters represent *real* partitions of the data set. Fur-

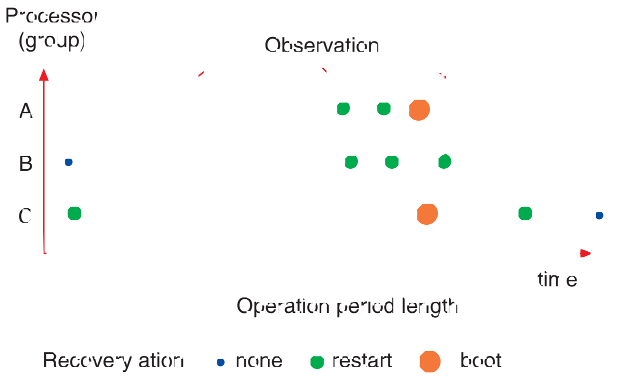


Figure 3 Illustration of events observed in a distributed system

	No of failures			Recovery action		
	A	B	C	None	Restart	Boot
Observation	3		2	5		0
Observation 2	5	3	6	7	5	2
Observation 3	3	0	2	4		0

Figure 4 Example of data matrix

thermore, it must be determined which clustering criterion is best suited for the actual application. The validation process during a cluster analysis is very important, due to the nature of the problem [8]. Keywords regarding validation are global/internal validation of clusters, ref. [5], [19], [36], and [33], simulation of cluster algorithms, ref. [26] and [29], result correlation [8], and cluster count tests, ref. [28] and [27].

2.3.2 Pre-processing to data matrix

The relevant registrations in the operational log may be viewed as a (multi-dimensional) series of failure related events as shown in Figure 3. Since the failure effect is not usually included in the log, the recovery action taken may be used to indicate the severity of the event.

These registrations must be mapped into a *data matrix* which is the fundament for cluster analysis, see Figure 4. One *observation* of the system is the accumulated registrations during one operation period of the system. This observation is represented by one row in the data matrix. The columns in the data matrix represent registrations in the log (failures) from single, or groups of, units of the system. System units may for instance be different control elements or groups of processors handling the same functions.

In a distributed system, the number of different system units may be large. Depending on the purpose of the analysis and to get a data set with reasonable complexity (computing time, interpretability), several criteria for grouping or selections of measurable system hints may be used.

The objectives of the case studies, to be dealt with in Section 3, were to investigate a functional partitioning of the system and to test the methodology in a more detailed analysis of specific parts of the system. Related to Figure 4, this means that A, B, and C were defined in two ways:

- Processors were grouped according to function in the system. The accumulated number of failure events for the processor groups are entered into the data matrix.
- Individual processors (within a subset of the groups) identified by their network addresses.

One functional processor group or one network address is represented by one column in the data matrix, see Figure 4. The number of columns may be increased to reflect the level of severity of each failure event. Several options were considered for what to include and how to represent this into the data matrix [12]. The choice was to focus on the recovery action following the failure events, e.g. None, Restart and Boot. As a trade-off, the total number of the different recovery actions were represented by columns in the data matrix as severity indicators, see Figure 4.

Performing an analysis with objectives other than ours, e.g. studies of the effect on “geographical” or topological factors, processor type and operating system, etc., the grouping may be done accordingly. Furthermore, the columns of the matrix representing groups and individual entities may be expanded with several self-diagnosed fault types. The degree of freedom is large and not restricted to the choices of our analysis.

2.3.3 Time intervals for observations

The choice of time interval for each observation, i.e. determining the length of the operation period, see Figure 3, is related to the need for detection of extreme behaviour or the need for localizing common operational modes of the system. The time interval must be a trade-off between being sufficiently long to observe the typical behaviour within a mode and being so short that shifts between operational modes are not smoothed. The interval should be in the same order of magnitude, but less than the typical operational mode sojourn time.

A time interval of 8 hours was chosen empirically, based on the results from tests on the data sets, i.e.

- Visual inspection of plots of the number of registrations for a number of different time intervals using the real data, localizing characteristic shifts between the plots e.g. in Figure 8,
- Performing cluster analysis for a number of different time intervals and check the physical interpretability and the Markov properties.

This, together with experience gained from other work on traffic load and load related dependability behaviour of telecommunications system ([40], [2], [15]), resulted in the choice of an 8 hour interval. This was done even though the 8 hour interval had poor performance in the Markov tests, see Sections 2.5 and 3.2.4. This may be explained by the periodicity of the data caused by the same periodicity of the load. This is illustrated in Figure 5. Note the seemingly rather long memory of the system with respect to the failure process. This figure also indicates that improved results in the cluster analysis may be obtained if we select every third 8 hour period.

2.3.4 Scaling and distance measure

The objective of applying a total distance measure (scaling and distance calculation) is to reproduce, in a numerical way, the physically motivated interpretation of the distance between observations. For instance, the following issues must be accounted for:

- The most critical step (in a failure context) is the step between zero and one failure, i.e. the step from a fault-free system to a faulty system. Whether a processor fails 284 or 285 times during an eight hour period is of little significance.
- Some groups/units may produce a lot more registrations than others, e.g. depending on their functional role. However, the registrations of all units have to be comparable in a numeric way without any units *dominating* the observations.

To fulfil the requirements of the total distance measure, a new scaling method was developed. This scaling method is based on the probability distribution of the data, see Figure 6. The following steps are performed for each column i of the data matrix.

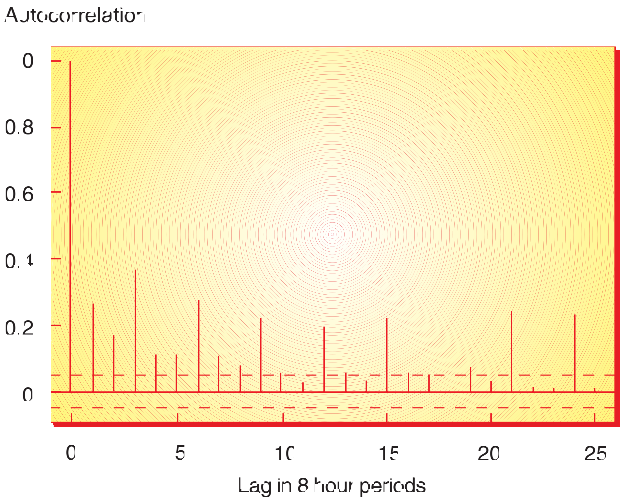


Figure 5 The auto correlation between the total number of registrations during 8 hour intervals

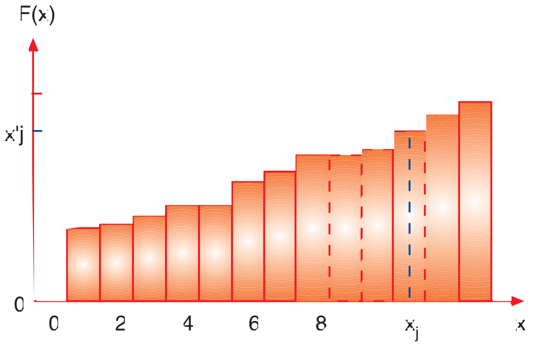


Figure 6 Empirical marginal probability distribution

- The marginal cumulative probability distribution function (CDF) $F_i(x) = \text{Prob}\{X_i < x\}$, of the data related to column i of the data matrix is estimated.
- The scaled value $x'_j = F_i(x_j)$, for each value x_j in column i, is used in the clustering analysis.

The main characteristic of the developed scaling method is the way the level of severity is reflected by *isolating* the fault-free registrations and smoothing the numeric difference between the level of each registration x_j. $F_i(x)$ was estimated empirically from the data. A Poisson distribution with estimated mean were also investigated, with a poorer outcome.

A Euclidean distance measure, as in Figure 2, was chosen for the actual calculation of the distance. The Manhattan distance measure was also tested but only minor differences were seen in the results. The Euclidean measure is easy to implement and is implemented in most cluster software.

It should be mentioned that a number of alternative distance measures, e.g. Euclidean, Manhattan and Canberra, and probability measures and scaling meth-

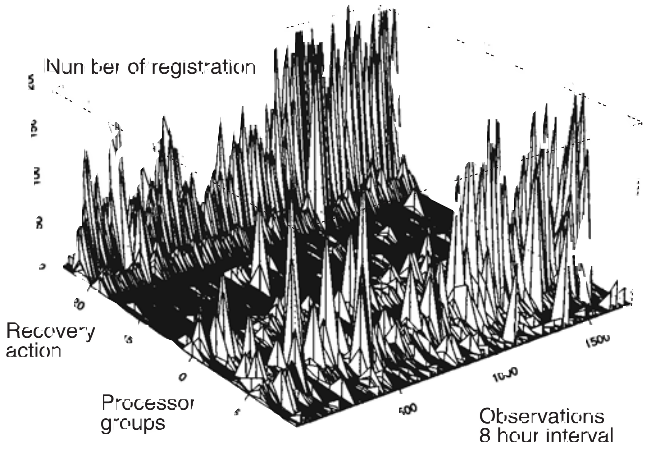


Figure 7 Example of data matrix

ods, e.g. logarithmic and z-transform (see for instance [36], [7], and [25]) were investigated. The result from these were compared with each other, with non-scaled data and with the above described scaling. Empirical evaluation criteria were used by interpreting the defined clusters and by a developed 3D plotting technique of the distance matrix. The main conclusion from this work is that a correct total distance measure is critical to reflect the physical information of the operational data in the cluster results.

2.3.5 Cluster algorithms

The theory of cluster analysis contains an enormous number of algorithms to perform the actual clustering of the data. Several algorithms were considered and tested. The key issue is to choose an algorithm with no side effects in the clustering related to the characteristics of the data.

Based on theoretical considerations and to some extent the available software (the SAS system, see [20]), four algorithms were chosen for further testing; Average Link, Complete Link, Ward's method and K-means, see [28] and [19]. The conclusion was that Ward's and K-means identified groups of observations defining special *modes* in the system, while algorithms like Complete Link and Average Link were well suited to identify single special *observations.* The evaluation criteria for choice of algorithm were the physical interpretability of the results, the correlation between algorithms and result characteristics, and the computing resources required.

The tests on the real data showed that the results are dependent on the cluster algorithm used. For instance, the Complete Link and the Average Link algorithms produced a few large clusters of observations with quite different characteristics, while the remaining clusters were small containing observations with very specific characteristics. Ward's method and K-means (the preferred algorithms in our case) produced clusters of more similar size with rather well defined operational modes.

2.4 Validation, presentation and post-processing of the results

As pointed out in Section 2.3.1, cluster validation is necessary to verify the obtained results. Several methods are available. Validation was performed by

- Comparing the clustering results produced by different algorithms (K-means and Ward's)
- Calculating the correlation between the registrations within the same cluster.

Several ways of presenting the cluster results were used:

- Kiviat plots (BCW) which use a star to represent the variables of a cluster, the length of the edges corresponds to the mean value of the variable, see Figure 10,
- Calculation of characteristic values of the variables of the registrations within each cluster, e.g. mean values, minimum and/or maximum values and probability distribution. The latter for comparison with the total probability distribution of each variable in the data set,
- 3D plots with the registrations sequenced in clusters along the x-axis, the columns in the data matrix along the y-axis and the registered values at the z-axis, e.g. in Figure 9. Also plots of the cluster sequences, i.e. the transitions between clusters, were made.

These cluster presentations were made to give a good characterization of the clusters, and to achieve a well-founded physical interpretation of the operational modes represented by the clusters. The operational modes were given a description which reflects their physical interpretation.

The main post-processing was related to calculation of transition probabilities for the transitions between clusters, and to tests of the Markov properties.

2.5 Transitions between operational modes

The identified operational modes, represented by clusters, may be interpreted to represent different *states* of the system. By this definition, state transition diagrams for the system may be built. Transition probabilities/frequencies are obtained from the cluster results by calculating the transition frequencies between the clusters. The intention of the state diagram is to give a view of a likely further development of an operational mode.

The Markov properties of the resulting models were investigated by Chi-Square type of tests. Two different techniques were used:

- Modes/states in isolation, by testing for independence between entry and exit transitions, see [41]
- Entire diagram [6].

If Markov properties are found, characteristics like recurrence times and sojourn times within sets of (failure) modes, may easily be calculated. These characteristic values give possibilities to predict the behaviour of the system, for instance, to initiate maintenance operations.

3 Case studies

Besides testing of different alternatives for the main steps of the cluster method, two extensive case studies have been carried out:

- **Case 1** analyses the system from a functional point of view. Processors are grouped according to the function(s) they perform in the system and represent the measurable entities.
- **Case 2** analyses individual processors (identified by their network addresses) within one or a few functional groups. This analysis gives a detailed physical view of operational modes within these subsystems.

We will here focus on results from Case 1.

3.1 Operational data

The data set used was extracted from the operational log of a public exchange, see Table 1.

For Case 1, the operational data was processed to a data matrix, with characteristics shown in Table 2. A 3D plot of the complete data matrix of Case 1 is shown in Figure 7.

3.2 Example results

In this section, examples of typical results generated from the use of the cluster method at the real data sets will be presented.

Table 1 Main characteristics of the system and operational data used for the trial investigation

Observation period	19 months
No. of registrations	67700 (only SW related reg.)
Type of system	Switching control system - Function sharing, proc. groups - Load sharing in groups - Mixed “type of redundancy”
No. of processors	277
No. of proc. groups	20 (size: 1 – 124)

Table 2 Characteristics of data matrix

Time interval	8 hours
No. of observations	1737
No. of columns	23
Column 1–20	Number of SW related failure events in distinct functional groups of processors
Column 21–23	Total number of events with recovery level None, Restart and Boot
Scaling method	Empirical marginal probability distribution
Distance measure	Euclidean

3.2.1 Pre-processing

An initial examination of the operational data is important to get a *feeling* of characteristics of the data in order to guide the choice of analysing methods.

The total number of failures per time interval was used to investigate the general behaviour of the system. These are shown in Figure 8. Fluctuations in the total number of failures are seen, including periods when the level of failures is generally high. The failure process does not seem to be stationary.

Auto correlation for the total number of failures showed a dependent behaviour for each third registration, cf. Figure 5. This is interesting because the 8 hour interval represents a logical partitioning of the day. The same characteristics were found for some control element types, while the auto correlation decreases fast to zero for others.

3.2.2 Operational modes of the system

Ward’s algorithm in the SAS system [20] was used to perform the clustering. The Pseudo-F test in the cluster software indicated 10 clusters as the optimal partition. A 3D plot of the data matrix sequenced by clusters as shown in Figure 9. This is to visualize the clustering and to compare with the unclustered data of Figure 7.

The mean number of processor group failures within each cluster was calculated and presented in a table and in a star plot shown in Figure 10.

Referring to the above, some main characteristics of the clusters may be found:

- Cluster 1 represents a fault-free mode.
- Clusters 4, 8, 9, and 10 represent modes where several functional processor groups have frequent failures. They differ in the total number of failures and in which processor groups that are the main contributors.
- Cluster 4 represents the most serious failure mode, since it has high values for registrations with recovery levels Restart and Boot.
- Cluster 10 represents a failure mode where a large number of different processors in the system fail frequently.
- Clusters 2, 3, 5, 6, and 7 are characterized by contributions from a limited number of processor groups, while the

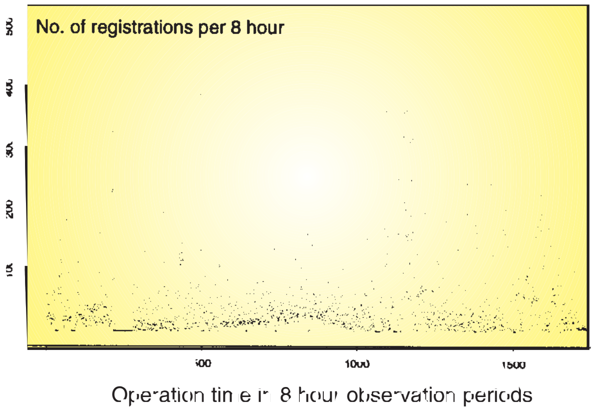


Figure 8 Total number of failure events

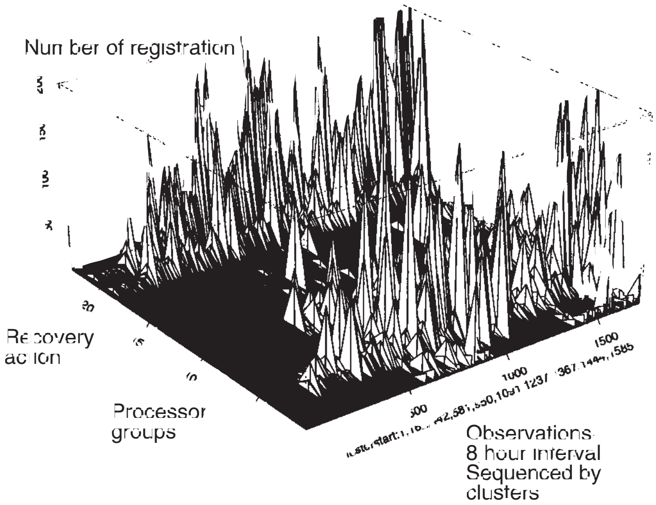


Figure 9 Example of clustering

other types are not faulty. The difference is contributions from different subsets of the groups.

A much more detailed analysis of each cluster, investigating the functional relations between the groups in a cluster, were performed in [12].

3.2.3 Cluster validation

Results of several algorithms were compared with the chosen result. In general, the results showed a high degree of correlation; besides that, Ward's algorithm further partitioned big clusters built by other algorithms.

The correlation between registrations within the same cluster was in general significantly higher than the correlation between random registrations.

3.2.4 State models and Markov testing

As pointed out in Section 2.5, the operational modes of the system, which are recognized as clusters, may be regarded as states. The frequencies of the transitions between these are obtained by parsing the cluster sequence. The result is shown in Figure 11, where the following is seen:

- The largest transition probability (0.61) from cluster 1 is the probability of returning back to itself. Hence, the fault-free mode seems to be the most stable.
- Note also that the failure modes 8 (0.55), 7 (0.37), 3 (0.31), 2 (0.30), and 4 (0.30) are rather stable.
- In general, the transition probability back to the same state is the largest for all clusters. This may indicate a shorter sojourn time than the interval of 8 hours. However, this is not the case in the more severe mode 10, where transitions are mainly done to mode 4, which is even worse.
- Most states have a high transition probability into mode 4, which is the most severe fault state of the system. The transition probabilities into failure modes 2, 10, and 9 are also generally large.
- The probabilities of returning to the fault-free mode 1 are small from all states: on average 4.5 %.
- During investigation of Markov properties of the system, cycles of behaviour may be detected.

The Markov properties of the "state/mode transition model" of the system were tested as outlined in Section 2.5. The tests showed that no Markov properties were present in the system's shift between modes.

3.2.5 Observation from other cases

The above presented results originate from one out of several examples. Some interesting results from the other ones are:

- Case 2, see the start of Section 3, generated results where the shifts between operational modes seem to be Markovian.
- Large dependencies between specific network addresses were found.
- In examples with a limited number of columns, the tendency was to find extremely well-defined clusters/operational modes.

3.3 Discussion and remarks

The main result of the use of the cluster methodology was that the method was well suited to identify *operational modes* of a distributed system.

- The modes localized have a well defined physical interpretation.
- Operational modes are well defined for the analysed systems, indicating a logical coupling between different processors of the system and functional groups of processors.
- Some modes are very sharply defined, in the sense that only a limited number of units are faulty in the defined modes.
- Other modes are more *fluffy,* in the sense that several units are faulty, but the rate of the failures is similar, and in that way well interpreted modes are defined.
- The severe fault states of the system are reflected by modes where nearly all system units fail repeatedly.

Besides the definition of operational modes of the analysed system, some other observations should be noted:

- Pre-processing of the data showed some cyclic behaviour of the data and fluctuations in the level of failure events. The auto correlation has a periodicity reflecting the time of the day. It is slowly decaying, which indicates long periods of related failures.
- The results give a base for further investigations of logical coupling within a large distributed system.
- A general conclusion with respect to the Markov properties of the results may not be given. This indicates a varying "memory" in the failure development, dependent on which aspects of the system that are going to be investigated, i.e. how the attributes (columns in the data matrix) are defined.

At the end of this section, it should be pointed out that the performed validation of the results is still insufficient for 100 % confidence of the results.

4 Conclusions and outlook

This article has presented a methodology for identification of operational modes of a system from its logged failure data. Used correctly, the cluster analysis based method is a suited means for analysing the dependability related behaviour of a distributed system. It enables an identification of dependencies between different system units and a prediction of the progression of its operational modes. It cannot, however, be concluded that this progression is Markovian.

The parameterization and the scaling function are tested on data from one system only. Hence, in order to assure the generality of the methodology, it must be tested on more systems. Another aspect which should be investigated, is to take into account the load/time of day dependency of the system. A further possibility is to include into the data, the operation and maintenance actions taken. By this, the consequences of the various actions in the different operational modes may be learned. This may be used to improve system operation.

Note that supplementary techniques may be used to study the behaviour of the system within an operational mode (cluster) for identification of the cause of an undesired mode.

It is foreseen that a model of the operational modes of a system may be obtained from and integrated into the alarm surveillance and performance monitoring of TMN based O&M systems. This will provide information, not available in today's O&M systems, about the current mode/state of the system, a rough estimate of its recurrence time and its likely next mode(s)/state(s), see Figure 12.

With the introduction of centralized management systems and based on the standardized frameworks of the Telecommunication Management Network (TMN), operational data collection and consistent log formats may be obtained in an easier way. Centralized management systems carry out the procedure of collecting data from several nodes or network elements. Standardized formats (e.g. alarm types) make it easier to analyse the behaviour of elements stemming from different system families and make comparisons. Data may be collected for a complete or partial network and analysed without extensive procedures for converting the data.

The operational data in this article is processed in a dependability context. Within TMN it is possible to produce logs not only for Fault Management, but also Performance Management, Configuration Management and Security Management. The proposed methodology may be adapted to cover analysis of logs from these management areas or event to detect dependencies between behaviour of events within several functional management areas.

Acknowledgement

This work is based on data collected and pre-processed by Bjørn Muggerud at Telenor Research & Development. The authors wish to thank him for his effort and co-operation. We also wish to thank Poul E. Heegaard for discussions during the work.

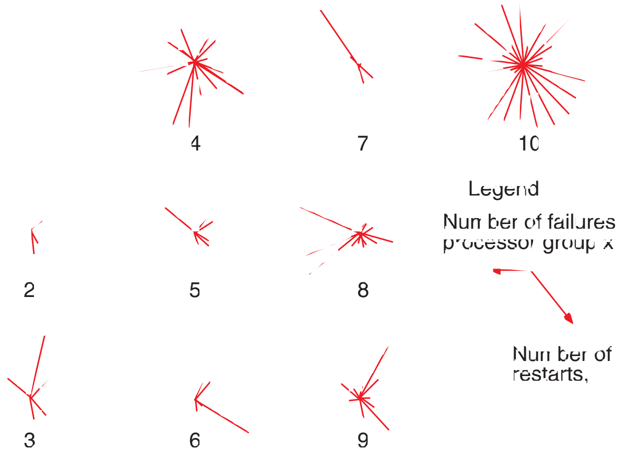


Figure 10 Star plot of cluster means. The length of each represents the mean number of failures within the corresponding processor group

References

1 Aakvik, G, Nordbø, I. *System for alarm interpretation : initial study.* Trondheim, SINTEF DELAB, 1992. (Technical Report STF40 F92076.) (In Norwegian.)

2 Bredrup, E et al. The activity-dependent failure intensity of SPC systems : some empirical results. *IEEE Journal on Selected Areas in Communication,* SAC-4, (7), 1052–1059, 1986.

3 Czeck, E W, Siewiorek, D P. Observations on the effect of fault manifestation as a function of workload. *IEEE Transactions on Computers,* 41, (5), 1992.

4 Dahbura, A T. Panel session : expert systems for diagnosis and diagnostics. In: *Proceedings 19th International Symposium on Fault-Tolerant Computing,* 21–23 June 1989, 422–425.

5 Dubes, R, Jain, A K. Exploratory data analysis. *Advances in computers,* 19, 113–208, 1980.

6 Osteyee, D. Testing Markov properties of time series. Time series analysis. *Proceedings of the International Conference, Houston, Texas,* August 1980.

7 Goodall, D W. A new similarity index based on probability. *Biometrics,* 22, 882–907, 1966.

8 Everitt, B S. Unresolved problems in cluster analysis. *Biometrics,* 35, 169–181, 1979.

9 Fitzgerald, K. Vulnerability exposed in AT&T’s 9-hour glitch. *The Institute,* 14, (3), 1, 1990. (A news supplement to IEEE Spectrum.)

10 Grogono, P et al. Evaluation of expert systems in telecommunications. In: *The World Congress on Expert Systems Proceedings,* 16–19 December 1991, 755–763.

11 Goyal et al. COMPASS : an expert system for telephone switch management. *Expert Systems (UK),* 2, (3), 112–126, 1985.

12 Gylterud, S A. *Identification of a model of system behaviour from operational data.* Trondheim, Norwegian Institute of Technology, 1992. (Diploma thesis.) (In Norwegian.)

13 Helvik, B E. Modelling the influence of unreliable software in distributed computing systems. In: *Proceedings 18th International Symposium on Fault-Tolerant Computing (FTCS-18),* 27–30 June 1988, 136–141.

14 Heegaard, P E, Helvik, B E. *Analysis of operational data and QoS based optimization in the future telecommu-*

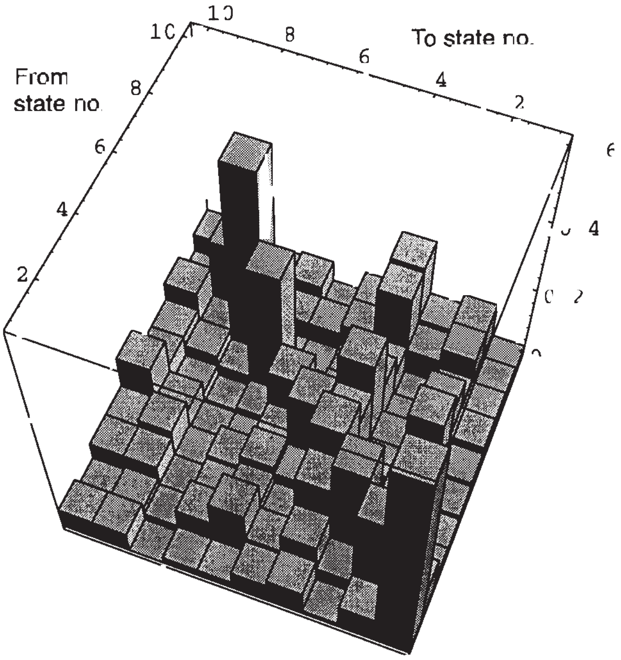


Figure 11 Plot of the state (/operational mode/cluster) transition probabilities from one period to the next

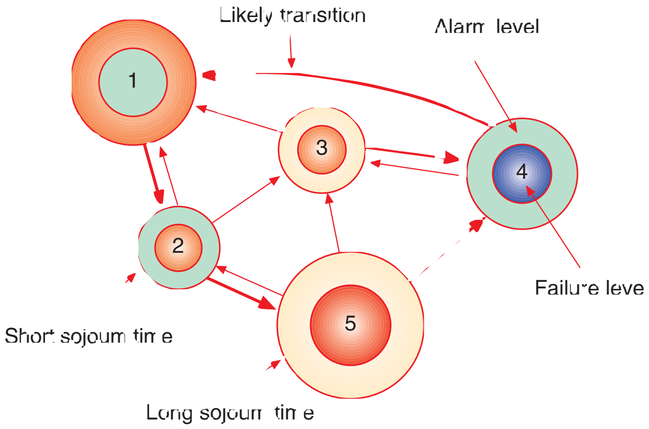


Figure 12 Sketch of a mode diagram for operation support. Thickness of arrows indicates likeliness of transition; size of the state its sojourn time. Colour (grey scale) indicates current consequence (failure level) and expected future consequence (alarm level). From [14]

nication network. Trondheim, SINTEF DELAB, March 1994. (Technical Report STF40 F94021.) (In Norwegian, restricted.)

15 Helvik, B E, Swensen, A R. Modelling of clustering effects in point processes : an application to failures in SPC-systems. *Scandinavian Journal of Statistics,* 14, 57–66, 1987.

16 Hansen, J P, Siewiorek, D P. Models for time coalescence in event logs. In: *Proceedings 22th International Symposium on Fault-Tolerant Computing,* 8–10 July 1992, 221–227.

17 Iwama, M, Kano, S (eds.). Special issue: toward the global intelligent network. *IEEE Communication Magazine,* 31, (3), 1993.

18 Iyer, R K, Young, L T, Iyer, P V K. Automatic recognition of intermittent failures : an experimental study of field data. *IEEE Transactions on Computers,* 39, (4), 1990.

19 Jain, A K, Dubes, R C. *Algorithms for clustering data.* Prentice Hall, 1988.

20 Jaffe, J A. *Mastering the SAS system.* New York, Van Nostrand Reinhold, 1989.

21 Løken, B, Espvik, O (eds.). Special issue on Intelligent Networks. *Telektronikk,* 88, (2), 1992.

22 Li, X. Parallel algorithms for hierarchical clustering and cluster validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 12, (11), 1990.

23 Lee, I, Iyer, R K, Tang, D. Error/failure analysis using event logs from fault tolerant systems. In: *Proceedings 21st International Symposium on Fault-Tolerant Computing,* 25–27 June 1991, 10–17.

24 Lin, T-T Y, Siewiorek, D P. Error log analysis : statistical modeling and heuristic trend analysis. *IEEE Transaction on Reliability,* 39, (4), 1990.

25 Lance, G N, Williams, W T. Mixed-data classificatory programs. *The Australian Computer Journal,* 1, 1967.

26 Milligan, G W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Working Paper Series.* Columbus, Ohio State University, College of Administrative Science, 1978.

27 Moureau, J V, Jain, A K. How many clusters? *IEEE International Conference on Computer Vision,* 634–636, 1987.

28 Massart, D L, Kaufman, L. *The interpretation of analytical chemical data by use of cluster analysis.* Wiley-Interscience, 1983.

29 Mojena, R. Hierarchical grouping methods and stopping rules : an evaluation. *Computer Journal,* 20, 359–363, 1975.

30 Prerau, D S, Papp, W L. The TARGET expert systems : trouble analysis and resolution for remote telecommunications equipment. In: *The World Congress on Expert Systems Proceedings,* 16–19 December 1991, 764–772.

31 Pyle, R H (ed.). Special issue on OSI network management systems. *IEEE Communication Magazine,* 31, (5), 1993.

32 Readdie, M et al. *MAES : an intelligent assistant for network operators.* ESA/ESRIN contract 8510/89/HGE-I, May 1991.

33 Ling, R F. On the theory and construction of k-clusters. *Computer Journal,* 15, 326–332, 1972.

34 Roubens, M. Fuzzy clustering algorithms and their cluster validity. *European Journal of operational research,* 10, 294–301, 1982.

35 Sekar, R C, Arthurs, E, Cameron, J. *Fault propagation in packet networks.* Morristown, NJ, Bellcore.

36 Sneath, P H A, Sokal, R R. *Numerical taxonomy.* San Francisco, Freeman, 1973.

37 Tang, D, Iyer, R K. Analysis and modeling of correlated failures in multicomputer systems. *IEEE Transactions on Computers,* 41, (5), 1992.

38 Tang, D, Iyer, R K, Subramani, S S. Failure analysis and modeling of a VAXcluster system. In: *Proceedings 20th International Symposium on Fault-Tolerant Computing,* 26–28 June 1990, 244–251.

39 Tsao, M M, Siewiorek, D P. Trend analysis on system error files. In: *Proceedings 13th International Symposium on Fault-Tolerant Computing,* 28–30 June 1983, 116–119.

40 Iversen, V B. *Computer- and teletraffic theory.* Den Private Ingeniørfond, 1985. (In Danish.)

41 Conover, W J. *Practical nonparametric statistics.* John Wiley, 1971.

42 Waghray, A, Magnusson, O S, Dziezic, R T. Expert network operations management system. In: *The World Congress on Expert Systems Proceedings,* 16–19 December 1991, 773–780.

Plant and people – the influence of human factors in telecommunications network performance

BY JOHN MELLIS

This paper reviews recent developments in understanding and modelling the effects of external influences, including weather and human factors, on the performance and reliability of telecommunication network physical infrastructure. The analysis focuses on traditional 'telco' copper cable access networks, but the implications for optical fibre networks are also shown to be significant. The result is a demonstrated need for improved standards for the performance of external network plant and cable. One response is BT's OTIAN® optical plant design specification, which is described and discussed.

1 Introduction

In the analysis of telecommunication network performance and reliability, the system is often narrowly defined in terms of the transmission layer: i.e. the multiplexing, transmission and switching hardware of the core network. In recent years, the reliability of transmission and switching systems has improved, while the relative importance and potential impact of other "layers" of the network has become more recognised. For example, the Operational Support Systems (OSS) software has grown, with the increasing complexity of switch processors and network management systems, so that in many cases the reliability of complex telecommunication networks is dominated by the contribution of the supervising software [1]. There has also been a growing awareness of the importance of operational realities and their crucial effects on network availability. Plant maintenance, fault diagnosis, and circuit planning and provision activities are all recognised to have a strong influence on the performance of the external access network in particular, but the nature and extent of these influences are not clearly understood in systematic terms, and quantitative analysis of the effect of external and human factors on network performance has been extremely limited.

This paper reviews recent work in analysing the contribution of external network infrastructure (plant and cable) to the overall reliability of the network. The first sections describe the approach used in a macroscopic analysis of fault rates in the BT UK external access network, where the contributing effects of external influences (e.g. weather) and human activity are included and quantified for the first time. The latter sections extend the consideration of these ideas to our emerging optical fibre access networks, and show that improved specifications and plant components are necessary if truly robust networks are to be designed and implemented.

The physical infrastructure layer is a good place to begin the task of modelling and understanding the influence of external factors on network performance. Comprehensive computer models of the physical network can be built using a relatively small set of infrastructure elements or objects; and as the ultimate service-independent network platform, the reliability of our physical plant and cable forms the basis of all other network quality of service measures.

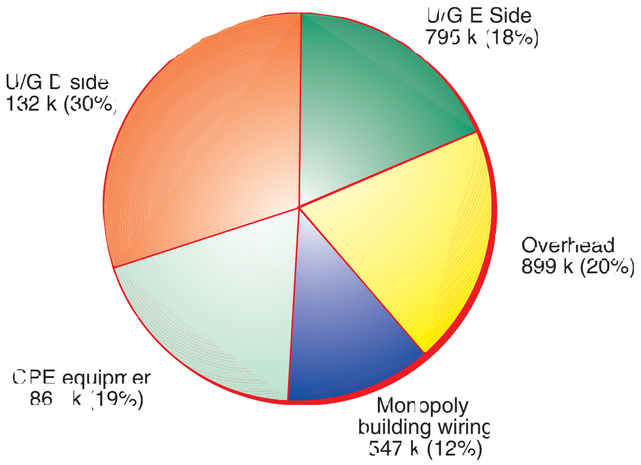


Figure 1 A typical distribution (1993) of annual fault volumes in the external and customer premises network

2 Traditional 'telco' copper access networks

BT's UK telephone network is based on an access network of a "traditional Telco" design, composed mainly of twisted-pair copper conductor cables, routed variously in underground ducts, overhead spans, or in some cases buried direct-in-ground. The flexibility to reconfigure this network is usually provided at a Main Distribution Frame (MDF, in the local exchange building), at a Principal Crossconnection Point (PCP, usually a street cabinet) and at the Distribution Point (DP, usually a pole supporting overhead dropwires) near the customer premises. Approximately 27 million working lines supply residential and business customers with a variety of PSTN and ISDN-based services, and the network is continuously reconfigured (for line provision), maintained, and repaired by a dedicated workforce of around 50,000 people. Around 4 million faults are recorded, diagnosed and cleared from this network annually. As shown in Figure 1, most of these occur in the external underground distribution network ("D-side") between the PCP and the customer, with the remainder spread roughly equally between overhead cables, "E-side" network (from exchange to PCP), Customer Premises Equipment (CPE), and so-called "monopoly wiring", those portions of in-building wiring which are the responsibility of the telecommunications company. The emphasis of this paper is on the external cable network, and so CPE and monopoly wiring faults are not considered in any further detail.

Usually, analyses of the reliability of this external network have taken a time-averaged view, where mean-time-between-failures (MTBF) levels have been deduced from long-term historic network fault rates [2,3] or a microscopic view, where component reliability targets have been set by deduction from the end-to-end availability and MTBF requirements for the network [4]. In practice widely fluctuating, highly time-dependent fault rates are more typical of

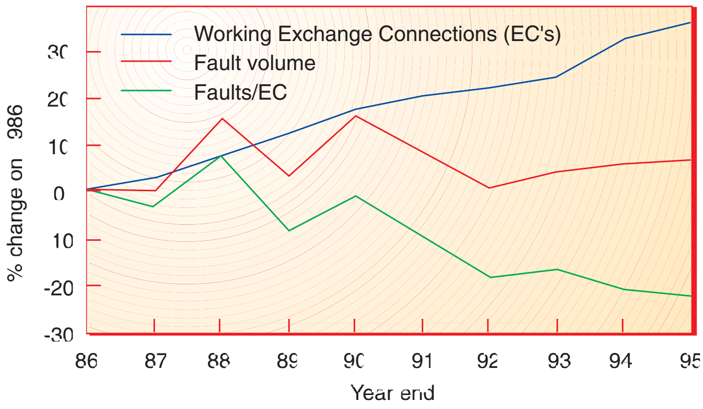


Figure 2 Trends and fluctuations in external network size, fault volume, and fault rate (faults per exchange connection)

telecommunication networks, and peaks in network fault rates can cause significant problems in network operational resource planning, and in the customers' perception of quality of service and network availability.

The fluctuations in network fault rate occur on a wide range of timescales, from hours to months. Figure 2 illustrates the fluctuations in even annual fault volume for the BT external access network (i.e. E-side, D-side and overhead) since 1986. A steady trend of improvement in fault rate can be discerned, as the network grows with the introduction of new plant and methods. However, the average fault volume (and maintenance burden) remains unchanged, with troublesome peaks in some years. Fluctuations on a monthly and weekly timescale are proportionally larger, and complicate the problem of resource scheduling and repair. The fluctuations are usually vaguely attributed to a number of variable causes such as the weather; network traffic levels; third party accidental or malicious damage; variable working practices; or the wear-out, or random failure, of specific types of external line plant. However, the systematic relationships between these postulated causes and actual recorded fault rates have been rarely investigated, and within most telecom network operating companies, wide differences of opinion are found regarding the relative importance of the fault rate drivers. The next section describes the formulation of a rigorous, but simple, mathematical model (FRAMEwork, for Fault Rate Analysis, Modelling and Extrapolation). We have used the model to clarify the relationships between fault rates and fault drivers, using a macroscopic or "black box" approach, emphasising the systematic relationships, not the detailed fault causes. The model predictions are compared with real network performance data, from specific access network areas, to quantify the relationships between fault rate and various assumed causes. These relate both to human network intervention and environmental effects, which have in turn been measured directly or indirectly. The relationships derived from this initial framework have been used to develop a generic model of the important fault generation processes, with the following objectives:

- To prioritise network fault drivers – for control or preventive action
- To predict future fault rates to assist in investment and resource management
- To assist policy formulation by cost/benefit analysis of maintenance policies
- To define new key performance indicators (KPIs) for comparing the effects of quality improvement initiatives.

3 A mathematical model for external network fault rates: FRAMEwork

3.1 Model structure

The basic structure of the generic fault rate model is illustrated in Figure 3, which shows the process flows and feedbacks associated with fault generation and repair. A number of assumed fault rate "generators" act upon the network infrastructure to produce incipient network faults at a rate which is dependent on the intensity of the generators, and the network's sensitivity to them. In general, both the generators and the sensitivities will be time-varying quantities. Three main classes of fault generator are considered: (a) environmental effects (e.g. the weather, fires, floods, and other sporadic interactions between plants or animals and the network infrastructure); (b) human factors including both the effects of third party damage (e.g. cable dig-up) and the effects of network intrusion by engineering work (e.g. network build, maintenance, and circuit

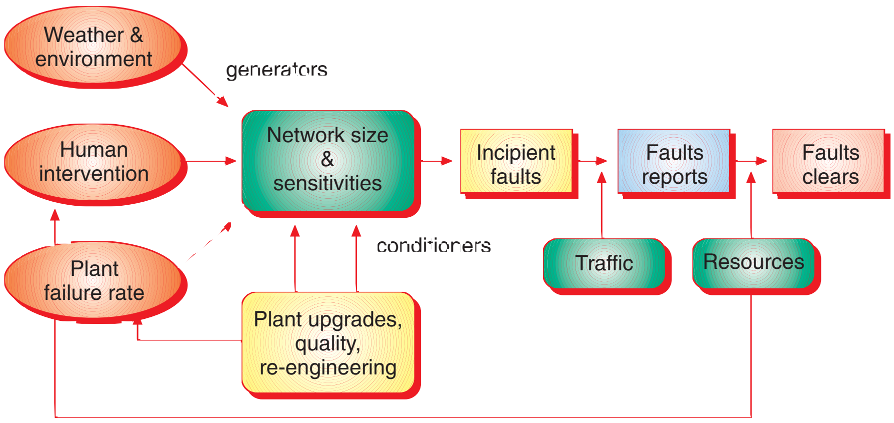


Figure 3 Structure of the generic FRAMEwork fault rate model showing how generators & conditioners act with network size & sensitivity to produce fault reports and clears

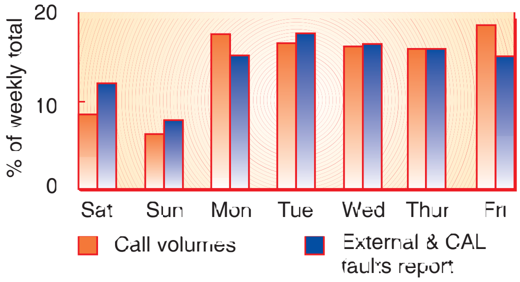


Figure 4 Correlation of daily external and customer apparatus fault reports with average daily network traffic levels

provision activity); (c) service-affecting equipment faults attributable to line-plant age, wear-out or random failure due to imperfect design. This last category is assumed in large networks to contribute a background "noise" level of faults which is relatively constant over time. All these generators produce incipient network faults which are converted to hard fault reports at a rate which depends on the particular fault detection mechanisms employed in the network. In general, these mechanisms (e.g. message failure reports, bit error bursts, customer fault complaints) will be dependent on the level of network traffic. Specifically, in the case of a network with no remote monitoring of terminal equipment, (such as a traditional public-switched telephone access network), the fault report rate will be totally dependent on traffic levels and customer fault reports. In fact, as shown in Figure 4, the correlation of daily traffic levels with daily average fault reports is a good one. Referring again to Figure 3, after detection and reporting, the network faults are then "cleared" by remedial action, resulting either in a fault repair or an inconclusive fault diagnosis report (e.g. fault not found). The intrusion of the diagnosis and repair activity must be considered; we have assumed that any manual intervention in the network is a possible fault generator, and this includes line provision, fault diagnosis and repair activity. Hence, a relationship between repair action and new faults must be defined in the model, either explicitly via a feedback loop, or implicitly by accounting for repair activity in the stack of fault generators.

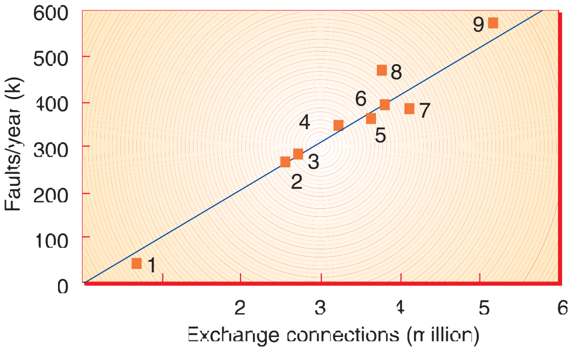


Figure 5 Scaling of zonal external access network fault volumes with zonal network size

The process just outlined describes the short-term life-cycle of faults from occurrence to repair, a cycle typically measured in minutes, hours or days. Within these timescales, levels of fault generators will fluctuate widely, as will network traffic loads. However, the sensitivities of the network to the various fault generators will vary more slowly with time, and over a smaller dynamic range. The network sensitivities will also be a function of both human and non-human factors, such as the intrusiveness of the fault diagnosis process; quality of repair and workforce training levels; condition, age and design of the network plant; and accessibility of cable ducts and enclosures. These factors are all encapsulated in a set of sensitivity parameters (one sensitivity for each fault generator) which vary slowly (weeks, months) under the influence of network "conditioners" in the form of quality improvement programmes, plant upgrades, process re-engineering initiatives, etc. The model therefore provides a rigorous way to compare the effects of fluctuations in the fault generators with longer-term trends in the underlying network fault sensitivities.

The simplest assumption of a linear relationship between faults generated and network sensitivities has been made, namely

$$F_i = G_i \cdot S_i \text{ and}$$

$$F = \Sigma_i (F_i) + B \qquad (1)$$

where F_i is the incipient fault rate due to fault generator *i,* with intensity G_i and network sensitivity S_i; B is the background fault rate due to random plant failures, and F is the total fault generation rate from all causes. More sophisticated relationships, including thresholds and non-linear expressions, could be used to express the dependency of fault rates on fault generators, and we have experimented with these. Although non-linear relationships can provide slightly improved fits between the model predictions and real data, this is at the expense of the clarity and usefulness of the model in providing simple network performance indicators.

In its simplest form the mathematical model is used to simulate monthly variations in fault rate and a number of simplifying assumptions become possible. The simplified model makes no distinction between the various sub-types of manpower activity (e.g. line provision and installation, repair, network construction, etc.) and allows for only one form of weather parameter, namely average monthly rainfall. Although the effect of wind speed and wind gust on overhead network damage is clear, we have found that storms correlate well with increased rainfall, and the average monthly rainfall in any locality is a convenient and reasonably accurate measure of weather severity in the UK. We have made the further simplifying assumption that at the monthly level, network traffic is relatively constant and that traffic variations can be neglected. Thus, the model in its simplest form becomes:

$$F = R \cdot S_r + M \cdot S_m + B \qquad (2)$$

where F is the monthly fault rate (faults per exchange connection) in the region under study; R is the recorded monthly rainfall (measured in mm); S_r is the local network sensitivity to rainfall (faults per exchange connection per mm); M is the recorded manpower activity in the area (man-hours); S_m is the network sensitivity to manpower activity (faults per exchange connection per man-hour); and B is the background rate of fault reports unconnected with the major fault drivers (faults per E.C.).

In all cases the normalisation of fault rate to network size (measured in number of working exchange connections) is important, because network size is the primary determinant of recorded fault volumes. Figure 5 illustrates this for the 9 operational zones of BT in the UK, whose sizes range from less than 1 million to more than 5 million working exchange lines. The scaling of fault volume to the size of the region is clear, and current network size is therefore a crucial parameter in the fault rate model.

The mathematical modelling problem has now been reduced to solving the simplified equation (2) for the sensitivity parameters *S* and the background fault rate *B*, given input data of the real fault rate *(F)* and the rainfall and manpower activity data for the corresponding time period. The effort required in the collection of weather and manpower data was considerable – even though much information was available in centralised databases, this required checking and normalisation for varying time periods (e.g. 4 week and 5 week months). U.K. Meteorological Office data was mainly used to provide rainfall information, but this was often compared to data gathered from rainfall monitors specially placed on the roofs of exchange buildings in the area under study.

The application of equation (2) to a set of monthly fault rate, rainfall and manpower statistics is essentially a "best fit" or regression analysis problem. Various techniques were trialled to achieve the best fit, as described in more detail elsewhere [5,6]. Overall the best balance of speed and simplicity was provided by the Microsoft Excel Solver facility which was used for most of the simulations described in the next section.

3.2 Model testing and initial application

In the preliminary application of the model, network fault performance data was collected from a particular exchange area (St. Albans) of around 40,000 exchange lines, and analysed, together with associated weather and workforce manpower activity data. The area under study was the subject of a quality improvement initiative [7] which allowed the comparison of network fault rates before, during and after the quality improvement exercise. Data collected included fault rate by network plant location (e.g. underground distribution network, overhead network, exchange-side network, etc.). The historical trend of the overall fault rate and its component parts was compared with corresponding data detailing manpower activity in the network, as allocated to various classes of work, including repair, build, line provision/installation, and safety improvements. The correlation between network fault rate and the input parameters describing manpower activity, weather and third party activity was analysed to identify a few principle fault rate drivers, and the simple mathematical model was used to simulate the apparent dependence of fault rate on these principle drivers. In this case a proprietary software dynamic modelling tool ("I-Think") was used to optimise the fit between the model output and actual historical fault rate by minimising the chi-squared sum function. The resulting comparison between actual and simulated fault rate behaviour is shown in Figure 6. Note that the starting point at period 1 shows a false zero in fault rate – the effect of the quality improvement programme was to approximately halve fault rate from its starting value to around 0.07 faults/line/year. However, during the period of most activity on network refurbishment, a transient increase in fault rate can be seen which is well fitted by the simulation model. The 3 discrete peaks which were observed in reality, but not well simulated by the model, relate to multiple fault events (e.g. cable damage) where several hundred fault reports were generated by each random event. For such a relatively small network area, random events of this scale will necessarily cause significant differences between real and modelled behaviour. The accuracy of the simulation is improved for larger network areas as shown in Figure 7 for a network region of around 1 million exchange lines. Here the simulated fault rate is shown broken into the 3 components related to (i) the background random plant failures including cable dig-up, (ii) the component proportional to the sum of all manpower activity, (iii) the component proportional to monthly rainfall in the region. Again the overall accuracy of the simulation is good and the larger size of the area simulated reduces the impact of individual random failures.

The relatively large fractions of the total fault rate which are correlated with weather and manpower activity are initially surprising, but broadly agree with much anecdotal evidence and the St. Albans test case. The largest peaks in fault rate in Figure 7 correspond to stormy months where overhead plant and cable damage was experienced, and where the FRAMEwork model's lack of a wind-speed-related input parameter causes some inaccuracy. Within these limitations, the structure and detail of the model are clearly sufficient to simulate the observed fault rate behaviour very well, and the model was used to investigate the variation in network fault behaviour across several network zones.

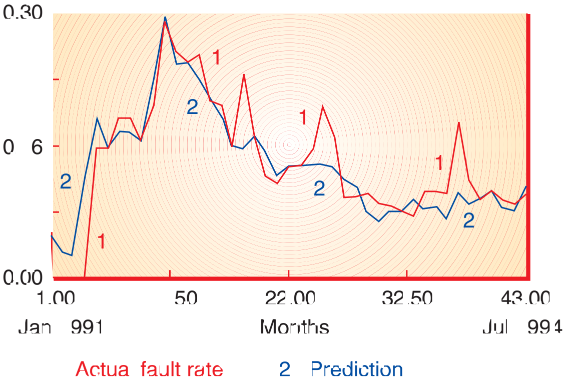


Figure 6 Comparison of actual and modelled fault rate (faults/line/year) for St. Albans exchange area over a 43 month period.

3.3 FRAMEwork: Detailed results

The model was applied to a succession of the other network zones with sizes of typically several million exchange lines, in order to determine whether a significant variation in the network sensitivities could be deduced. For larger zones even better agreement between actual and simulated fault rates could be obtained, as the effect of local spurious events was further averaged out. The results (e.g.

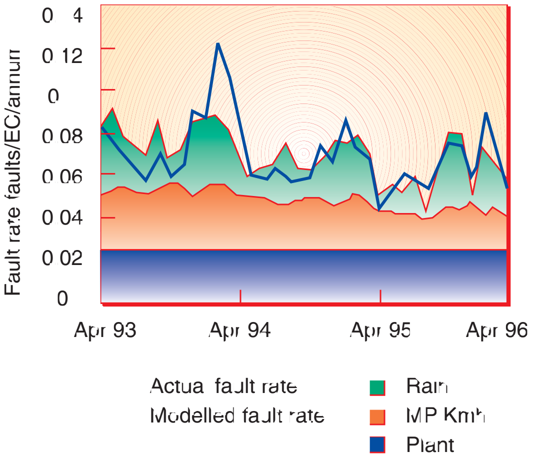


Figure 7 Modelled and actual time-dependent fault rate for Zone 1 over a 12 month period, showing the contributions from the model for the 3 principal fault categories

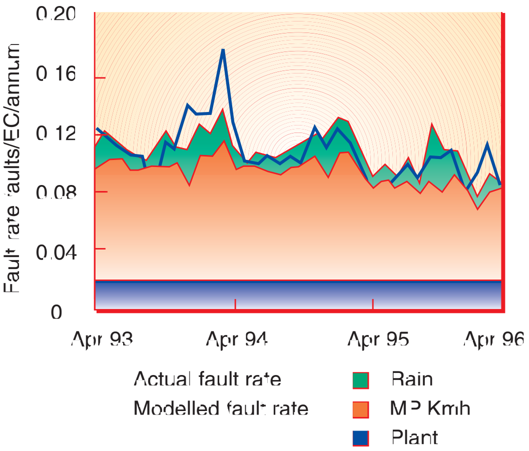


Figure 8 Actual vs. modelled time dependent fault-rates for Zone 2

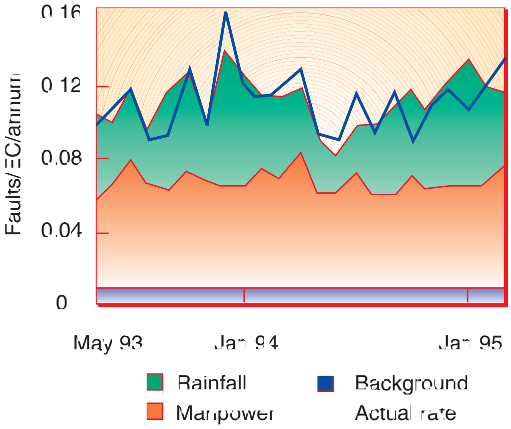


Figure 9 Modelled and actual fault rate for largest zone

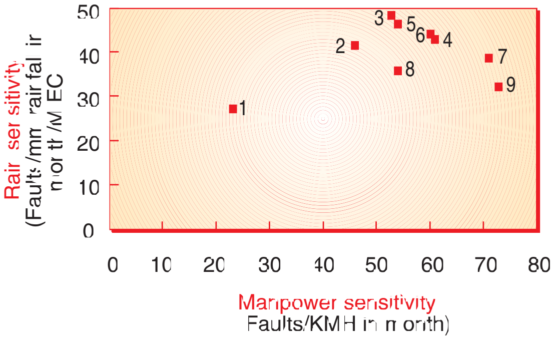


Figure 10 General range of deduced network fault rate sensitivity to monthly rainfall (mm) and manpower (kilo-man-hours) for several network zones

Figure 8) showed that a good modelled fit (typically +/- 5 %) to actual measured fault rates could always be achieved by adjusting the network sensitivities used, and that significant variations in apparent zonal network sensitivities indeed exist. The simulation for the largest zone is illustrated in Figure 9.

An informative way to reveal the differences between network zone sensitivities is illustrated in Figure 10, which shows the general range of networks sensitivities to weather and manpower effects, as deduced from all the zonal model solutions. The solutions for zonal sensitivity parameters lie in the general range of 20–50 faults/mm rainfall per million exchange connections, and 20–80 faults per kilo-man-hour of external network engineering activity. The deduced variations between zones may be highly significant; for example, revealing differences in network quality or condition, duct congestion levels, operational methods, and workforce productivity. The modelled network sensitivities offer a new set of key performance indicators for the network, and these can be used to assess the underlying performance in any region or time period.

Some caution must be applied in the interpretation of the results, however. The network sensitivity parameters have been normalised for the size of the network areas studied purely on a ‘per exchange-line’ basis, which neglects, for example, quite large variations between zones in the fraction of network plant which is sited overhead (where it is more vulnerable to the weather). A better way to normalise the zonal comparisons for areas of overhead network is therefore being devised.

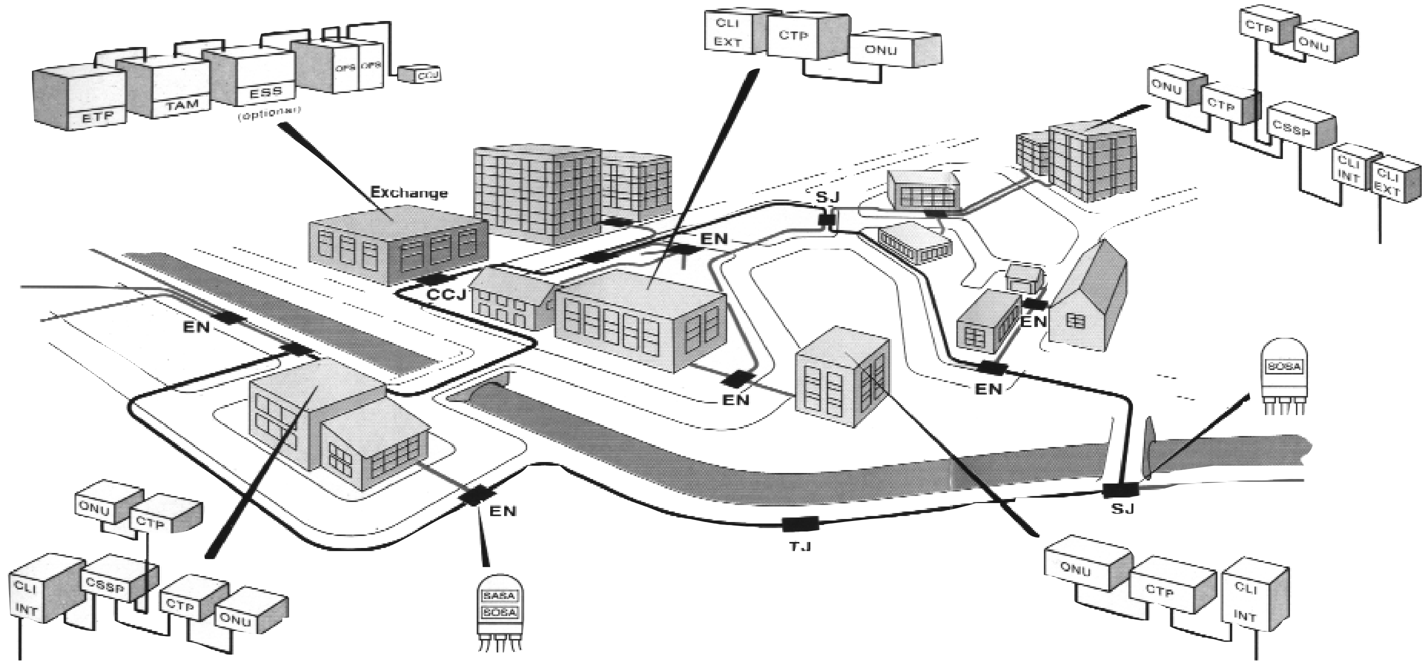
As a second cautionary note, the ‘faults/kilo-man-hour’ parameter is akin to the *inverse* of workforce productivity, and any attempts to reduce this parameter must be consistent with high achievement against other key performance indicators and overall productivity. A better measure of performance along this axis might relate the unwanted faults generated to the desirable output of the work being done (e.g. faults generated per repair, or faults generated per line installed).

Further investigation is required of the uncertainty bounds on the deduced network sensitivities, but many of the differences between zones are certainly significant. With improved uncertainty bounds, it will be possible to detect small underlying changes in network performance, in any given zone, and assess the impact of operations management initiatives and further quality improvements.

4 Digital and optical fibre networks

The mathematical model FRAMEwork has so far been applied only to the study of customer-reported fault rate behaviour in ‘traditional’ telecommunication access networks based on metallic cables, carrying mostly analogue telephony services. The large scale and density of these networks in most national telephone networks is immense, and this makes them clearly the most important influence on customer perceptions of service quality. However, it is interesting to apply the same thinking to the fault rate dynamics of more technically advanced networks, and to speculate about the influence of external factors on the reliability of networks such as those carrying digital private circuits, the Integrated Services Digital Network (ISDN), and the optical fibre access networks now being constructed in many countries. Although the technically advanced nature of these networks might initially be thought to confer greater reliability, further reflection shows this to be untrue for several reasons. Where the analogue copper telephony network relies on customer fault detection and reporting, the digital networks are more usually equipped with continuous automatic monitors and alarm systems, with a high efficiency in detecting incipient network faults. In many cases the transmission equipment has a high sensitivity to changes in the line transmission parameters, and transmission operating margins and alarm thresholds are low, so that the volume of alarm reports greatly outweighs the number of ‘real’ customer reported faults in an analogue network. In this case we could expect the rate of digital transmission alarm reports to be an even more sensitive function of the fault rate drivers, including the external influences of weather, third party intrusions and workforce activity.

In the case of optical fibre telecommunication networks, an even more fundamental consideration is introduced by the nature of the transmission medium. The



CLI INT – Customer Lead-In Internal
CLI EXT – Customer Lead-In External
CSSP – Customer Splicing and Splitting Point
CTP – Customer Termination Point
ONU – Optical Network Unit

SASA – Splitter Array Sub-Assembly
SOSA – Splice Only Sub-Assembly
ETP – Equipment Termination Point
TAM – Test Access Module
ESS – Exchange Splitter Shelf

OFS – Optical Flexibility Shelf
CCJ – Cable Chamber Joint
EN – External Nodes
SJ – Spur Joint
TJ – Track Joint

Figure 11 An optical access network comprising cables, track joints, spurs and customer termination points

conventional single-mode optical fibre deployed in telecoms networks is very sensitive to physical bending. Bend radii of 10 mm, or less, begin to transfer significant fractions of the optical power in the fibre core to the glass cladding of the fibre, where it is lost by radiation and absorption. This is an effect which when properly controlled can be put to good use (as in “clip-on” optical power meters) but if uncontrolled, can drastically degrade the transmission performance of the fibre. Around seven years ago, BT experienced the first significant examples of optical network transmission bit-error-rate faults generated by the handling of ‘live’ fibre, by engineering technicians in the course of their normal circuit repair and installation duties. These faults were induced in the ‘junction’ optical cables carrying 1310 nm wavelength traffic. Junction cables run between core network switches and local exchanges, where fibre jointing and re-routing activity is more common than in the core network, and where the density of cables and joint boxes is correspondingly greater. The fibres were being handled within splice trays in cable joints, to allow access to adjacent fibres for repair or installation, but significant transient optical power losses were being introduced even with careful handling [8]. It was realised that this problem would be greatly worsened in future dense optical access networks, for three reasons. First, the frequency of joint box entries (for customer connections and reconfigurations) is necessarily higher in the access network compared to the junction and core networks. Secondly, the access network fibre transmission systems are being required to operate with relatively small margins of optical power budget, to reduce costs, compared to core network links which often have power budget margins of 10 dB or higher. Thirdly, the sensitivity of optical fibre to bending greatly increases with increasing optical wavelength, and modern access systems are beginning to use the 3rd optical fibre transmission window at around 1550 nm.

This problem of unacceptable optical access network mechanical sensitivity could potentially be addressed by redesigning the conventional single-mode optical fibre, or using specialised optical fibre less susceptible to the bending effects. For reasons of simplicity and backwards-compatibility, an alternative approach was adopted, involving the specification of a new system of optical cable-plant equipment which would be specifically designed to manage the fibre in a controlled and accessible way.

5 OTIAN® – a new standard for optical network infrastructure performance

The development and performance of the new optical plant systems have been described in several published papers [9–12]. Some key principles were applied in defining the requirements for the new Optical Telecommunications Infrastructure for Access Networks (OTIAN®):

- Continuous management of optical fibre to preclude unwanted bending
- Controlled access (including test access) to all fibres in cable joints, independent from adjacent fibres
- Use of common hardware designs in all parts of the network (exchange, external, customer premises)
- Ability to house passive optical splitters, multiplexers, and filters
- Ergonomic and modular hardware to reduce installation costs compared to traditional optical plant.

All these objectives were successfully and rapidly achieved by a specification and development project undertaken collaboratively with manufacturing industry partners, and a number of suppliers now offer a full range of OTIAN®-compliant products, including BICC Cables, Pirelli Cables, and Raychem. The key requirement principles were met by designing modular splice-tray "stacks" which allow independent but bend-controlled access to all the fibres in the cable joint. The splice module design is identical irrespective of the module's location in the network, which as shown in Figure 11 could be in an external cable node (EN), cable track joint (TJ) or cable spur joint (SJ), a customer splice point (CSP), exchange flexibility shelf (OFF) or in the case of a passive optical network employing optical splitters, in a Splitter Array Sub-Assembly (SASA). The designs differ in detail between manufacturers, but in practice all have been found to provide the required immunity from the undesirable effects of fibre handling, and all provide non-intrusive test access for live fibre using the Clip-On® power meter principle. Installation of OTIAN® plant in new BT access networks began in 1994 and deployed volumes are increasing as its potential even in the junction and core networks is realised. The new plant system has been warmly received by customers and technicians. Installation times and labour costs are reduced compared to traditional 'uncontrolled fibre' plant designs and ergonomic yet compact cable joint and customer termination modules have been developed. The troublesome problem of transient optical losses induced by fibre handling which was experienced with the old plant systems has been removed [13]. Capital costs of the OTIAN® plant hardware have been falling as deployment volumes rise and development and tooling costs are recovered, so that the whole-life cost of OTIAN® access networks can be shown to be significantly lower than networks using the 'old' plant which still causes problems with congestion, fibre access and breakage, and reduced network availability. An additional advantage of OTIAN® plant has been that the improved fibre control allows a complete end-to-end specification of the transmission performance for optical access networks, including optical static and transient loss specification across all the fibre transmission windows. The OTIAN® requirements have become internationally established and are in the process of standardisation by ITU-T [14].

6 Conclusions

External factors, including weather, workforce activity, and third party intrusions can be seen to have a generally disregarded, but nevertheless crucial, influence on telecom network performance. We have analysed the likely scale of such influences on copper telephone access networks by building a general but simple mathematical model (FRAMEwork) which simulates the actual network fault rate dynamics with remarkable accuracy. To our knowledge this approach is the first attempt to quantify access network fault performance using a "black-box" or systems engineering approach, where the fault generation processes are not considered in detail, but only as mathematical relationships between fault causes and fault rates. This novel analysis and simulation has shown that a very accurate simulation of real access network fault trends can be generated using only a few assumed input fault drivers, and that human intervention is a significant determinant of fault rate fluctuations. The model offers the use of new key performance indicators which specify the network sensitivity to weather and a range of network engineering activities. The model is also a powerful aid to fault rate prediction, and enables a much more accurate assessment of the effect of quality improvement initiatives than could be achieved by simple observation of fault volumes or fault rate trends. With further development, the new network performance indicators could be highly useful internal comparators and international benchmarks. Future work will explore the possibility of improved model simulations by using e.g. non-linear and neural network simulation techniques.

Although the model has as yet been applied only to the plain old telephone network, evidence suggests that it is equally applicable to digital networks, where the sensitivity of the fault rate to the underlying drivers may be even higher because of increased fault detection rates and reduced operating margins. In the case of optical fibre telecom networks, the potential susceptibility to human factors is greater still, unless best practices are employed in the specification and design of optical cables and plant. One successful example is BT's OTIAN® requirement which has resulted in a new range of commercially-available plant hardware with vastly improved ergonomics, reduced installation times and whole-life costs, and elimination of transient losses due to the handling of optical fibre by engineering technicians.

Improved understanding derived from analytical models like FRAMEwork, or mechanical design advances such as OTIAN®, can help to reduce, but cannot eliminate, the influence of external factors on network performance. The importance of human and other external factors in the planning, delivery, and maintenance of telecommunication services will continue to be fundamental. If it is forgotten, there is a risk that the difference between planned and real network performance may widen, as the density and complexity of network hardware and software continue to grow.

Acknowledgements

This paper has drawn on the sustained excellent work of many people in BT and manufacturing industry, whom I thank for their co-operation and support. In particular I thank all my colleagues in the

Planning, Plant and Pilots development unit at BT Laboratories, but especially Matthew Thomas, Steve Chase, and Tony Gould for their development of the FRAMEwork model; and Nick Medlen, Paul Jessop, John Peacock and Dave Stockton for their work on OTIAN® development and evaluation. OTIAN and Clip-On are registered trademarks of British Telecommunications plc.

References

1 *Network reliability : a report to the nation.* Compendium of technical papers published by the U.S. National Engineering Consortium, Washington DC, June 1993.

2 Spencer, J. Getting a clear picture on video service reliability. *Proc. Eur. Conf. Optical Commun. and Networks (EFOC&N),* Brighton, UK, June 1995, 142–145.

3 Cochrane, P, Heatley, D J T. Reliability aspects of optical fibre systems and networks. *BT Technol. Journal,* 12, (2), 1994, 77–92.

4 Mellis, J et al. Performance and reliability requirements for passive components in local optical networks. *Proc. Eur. Conf. Optical Commun. and Networks (EFOC&N),* June, 1994, 112–115.

5 Thomas, M R et al. Fault rate analysis, modelling and estimation. *BT Technol. Journal,* 14, (2), 1996, 133–139.

6 Thomas, M R et al. Fault analysis, tools, methods and approach. *Proc. 9th Int. Conf. on Condition Monitoring and Diagnostic Eng. Management,* Sheffield, U.K., July 1996.

7 Prior, J, Chaplin, K. The St. Albans study. *BT Eng. Journal,* 14, 1995, 46–49.

8 Peacock, J. Measurement of optical transient losses and their effect on deployed optical systems in the BT network. *Electron. Lett.,* 30, (20), 1994, 1701–1703.

9 Dalgoutte, D G et al. OTIAN : an outside plant system for the access network. *Proc. European Conf. on Optical Communications and Networks (EFOC&N),* June 1994, 37–40.

10 Dalgoutte, D G, Lea-Wilson, N D. Installation cost issues in external plant deployment. *Proc. European Conf. on Optical Communications and Networks (EFOC&N), Brighton, UK,* June 1995, 3, 146–149.

11 Frankx, J F, Peacock, J. A new generation of products for the access infrastructure. *Proc. Conf. on European Optical Communications and Networks (EFOC&N), Brighton, UK,* June 1995, 3, 150–153.

12 Hale, P G et al. Modular optical plant for the access network : a practical solution. *Proc. Conf. on European Optical Communications and Networks (EFOC&N), Brighton, UK,* June 1995, 3, 158–161.

13 Hale, P G et al. Modular optical plant for the access network : operational aspects. *Proc. Conf. on European Optical Communications and Networks (EFOC&N), Brighton, UK,* June 1995, 3, 164–167.

14 ITU-T. *Study Group 6 draft recommendation document DT/23-E.* Geneva, March 1995.

Dependability in telecommunication networks

BY STEIN HAGEN AND KJELL E STERTEN

1 Background

1.1 Introduction to network reliability

Telecommunication networks in the nineties are characterized by:

- Integration/packaging of conventional services with "intelligent" and multimedia-based capabilities
- Transparent user access realized by consolidated access networks and interconnect agreements
- An increased use of telecommunications in society, continuously making more vital functions dependent on it
- Introduction of advanced network management systems facilitating supervision and protection against failures while at the same time introducing new failures, e.g. in software, possibly risking even network breakdowns.

The customer expectations to the overall network performance are rising and are further amplified by the increased competition due to national deregulation trends.

Network reliability is one of the most important network quality parameters and measures which affect the overall network performance (from a customer point of view). It is defined as *"the capability of the network to provide a required function under given conditions for a given time interval"* and is related to other network and customer oriented measures in ITU-T E.800 series recommendations.

Some of the issues which can form the basis for network reliability are:

- The coupling of network design objectives with network reliability targets and the methods/techniques/tools which are at the disposal of network designers for exploring the impact on reliability of different network architectures.
- How to tune and/or compromise network reliability in relation to the introduction of new network technologies and unforecast traffic demands.
- How to relate, maintain and monitor network reliability across the various network and management layers.
- Association of network reliability with network availability for the purposes of network costs and tariffs setting.
- Dimensioning of the level of network reliability with regard to cost and market demand.
- Unified or diversified levels of reliability for different customers?

1.2 Availability work in Telenor

The work in the area of reliability and availability in Telenor has been carried out for many years. The work has been spread out on different parts of the organization and has been given different emphasis.

Some important studies in this area are:

- **Policy for protection of the trunk junction network (1975)**
  The report recommended independent transmission routes from all exchanges above 5000 subscribers.
- **Dependability planning, Report 1–5, and the Dependability Handbook (1980)**
  The dependability planning project was a great effort on dependability in Telenor. It was a theoretical study made by the Research Department, and the project was summarised in the Dependability Handbook. The requirements in the handbook was determined by the number of subscribers being affected by a fault. The co-operation with the Technical Department, responsible for the planning process, was insufficient. The handbook was too theoretical and lacked reference to practical issues related to both the level of reliability and costs. In addition, the methods were difficult to implement in the planning process, and the recommended solutions were too expensive at that time.
- **Policy for protection of the trunk network (1985)**
  The report recommended protection routes for cable systems, protection switches and the use of protection channels for radio link systems. The recommendations were based on the assumption that the level of availability and the investments must be adapted to the customers' (or the society's) benefit of the improvement of the availability.
- **Risk analysis (1988)**
  Telenor was in 1986 directed by the Ministry of Transport to perform annual risk analysis in its installations. A risk analysis method (SBA) was developed in 1988 to calculate the risk of damaging an installation caused by external events. The work with the SBA-analysis was very important and gave a significant decrease in the risk of damaging an installation. Fire and sabotage was considered to be the most serious events and precautions against these events were given priority. The SBA-analysis does not cover "normal" faults and other interruptions in the network.
- **NT/R Nordic working group: Reliability Methodology (1979–1991)**
  A number of reports was issued. In the beginning the working group was influenced by the Norwegian Dependability project. Later on, the Swedish members were the leading persons in the group. The working group adopted a method for dependability planning developed in Sweden based upon the concept of balancing the cost of the operators and the benefit of the customers. The group was a major contributor to ITU and the method was the basis for ITU-T Rec. E.862.
- **The target network for the telephone network/ISDN, Report 1–5 (1990–1992)**
  The reports deal with the target network for the telephone network/ISDN. In particular report No. 2, "Choice of methods to ensure high level of grade of service" is based upon fault tolerance principles, combined with a simplified structure. The structure is characterized by extensive use of two separate routes towards superior exchanges. The number of routes across the hierarchical network was strongly decreased.

1.3 The situation of today

The analogue exchanges were very reliable. The structure of the exchanges ensured that only parts of the exchange were affected by faults. The introduction of digital exchanges introduced software failures and a decrease of the availability, in particular when introducing new software packages.

The development of the network in the eighties was based on the idea that the network should satisfy certain technical requirements and had little emphasis on cost. The service integration in the telephone/ISDN-network and the introduction of competition and deregulation led to a stronger control of the investments. Different quality for different services, customers and geographical location must be taken into account.

The target network for the telephony and ISDN network introduces a certain redundancy in the network. Introduction of SDH-technology will support rerouting-facilities simpler and even cheaper than before.

It is necessary to increase knowledge of availability in the network and to know the emphasis that customers give to interruptions in the service. It is also important to develop models for evaluating alternative solutions.

The experience from competitive environments shows that availability is an important factor in the competition. Major failure events will be brought to mass media and the competitors will use this to their advantage. A balanced attitude to availability issues are important in a situation with increasing competition for both customers and investments.

It is thus important for any operator to establish a strategy within this area which will be followed up by all parts of the organisation. There is a need for establishing methods for dimensioning and allocating dependability in telecommunication networks.

On 12 December 1993 the management of Telenor Network decided to establish a dependability project . The objective of the project was to develop principles and methods for calculating availability and practical planning rules for both transport and service networks.

2 Description of a dependability project in Telenor

2.1 Introduction

The decision to start the project was taken by the end of 1993. The project was established during the first half of 1994 with one project leader and 8 participants from both central and regional parts of the organization of Telenor Network. The project owner was the Planning Department of Telenor Network. A steering committee with 3 members was appointed.

The forecast manpower in the project was 3.5 man-years and the estimated project time was 2 years. Each project member should contribute 20 % and the project leader 30 % of the working hours during this period of time.

The formal constitution of the project was made on the first project meeting 23 June 1994.

The project was given the following mandate:

> *Develop principles and methods to choose availability actions in the planning of transport and service networks. The cost of the actions should be evaluated against the customers' need for such actions. The goal is that actions in both transport and service network should be co-ordinated. Practical planning rules for availability for the transport network and the ISDN/telephone network should be established. Goals should be established for the availability of the most important services delivered by Telenor Network.*

2.2 Organization of the project

The project was organized according to the PSO-model (Personnel, System and Organisation). This model has been developed by a Norwegian consulting company and was recommended to be used in project work in Telenor Network.

The PSO-model emphasizes precise goal and milestone descriptions and clarified responsibilities within the line organization.

The project was established with 4 different courses:

- M-course (Methods, M1 to M7)
- C-course (Customer, C1 and C2)
- S-course (Statistical, S1 to S3)
- O-course (Organisation, O1).

Each course was divided into several milestones named with the capital letter of the belonging course accompanied with an integer.

Below the different milestones in the project are listed:

- M1: Updating of the participants' competence within availability
- M2: Develop principles and methods to choose availability actions in the planning of transport and service networks
- M3: Propose protection/restoration mechanism in the transport network
- M4: Propose computer program for calculating network availability
- M5: Develop practical planning rules for the transport and the ISDN/telephone network
- M6: Establish goals for the availability of the most important services
- M7: Summary report of the project with conclusions
- S1: Preliminary availability data for network components
- S2: Developing routines for collecting and processing availability data
- S3: Updated availability data for network components in Telenor's network
- K1: Preliminary values for the customer estimated values of lost traffic and a traffic model for calculating the effect of reduced circuit capacity
- K2: Values of lost traffic based upon estimation from Telenor's customers
- O1: Establishing training activities for personnel in Telenor Network.

Each milestone is described in detail containing start and stop dates, milestone leader, participants, estimated manpower, detailed milestone activities, dependencies from other milestones and deliveries from the milestone, etc.

The project was finalized 1 September 1996. Due to some changes in the manpower situation during the project period all milestones were not completely fulfilled according to the milestone descriptions. All milestone activities and deliveries are described in separate reports. The main conclusions and proposals are also stated in the summary report (M7) which is now being discussed within the organisation of Telenor Network.

In Chapter 3 to 8 in this article the major discussions and the conclusions of the project are given.

Chapter 3 states the most significant definitions according to ITU-T Rec. E.800. In Chapter 4 a brief description of two different approaches to network availability are discussed, whilst Chapter 5 gives a short review of the ITU-T Rec. E.862: Dependability planning of telecommunication networks. In Chapter 6 protection and restoration methods in the SDH-network are described and Chapter 7 contains some examples of the dependability planning method used on different parts of Telenor's network. Finally, the major conclusions of the project are given in Chapter 8.

3 Definitions

3.1 Introduction

Standardization terminology is necessary for two main reasons:

- To avoid confusing the users of standards by introducing conflicting terms and definitions
- To assist alignment between the various groups involved in telecommunication standards development.

A consistent set of terms and definitions is required, therefore, to develop the important areas of quality of service, network performance and dependability standards pertaining to the planning, provisioning and operation of telecommunication networks.

In ITU-T Recommendation E.800 "Quality of service, network management and traffic engineering", terms and definitions related to quality of service and network performance including dependability are stated.

The intention of this Recommendation is to set out a comprehensive set of terms and definitions relating to these concepts. Associated terminology covering statistical terms, recommended modifiers, etc., is also included to ensure the broadest possible coverage in one document. These collective terms and definitions can be universally applied to all telecommunication services and the network arrangements used to provide them.

In paragraph 3.2 below a general concept to overall quality of service is given based upon the ITU-T Rec. E.800. In paragraph 3.3 definitions used in the dependability project in Telenor are given. Most of these definitions are also based on international standards such as ITU-T Rec. E.800.

3.2 General guide to concepts

Figure 3.1 (Performance concepts) is a framework intended to provide a general

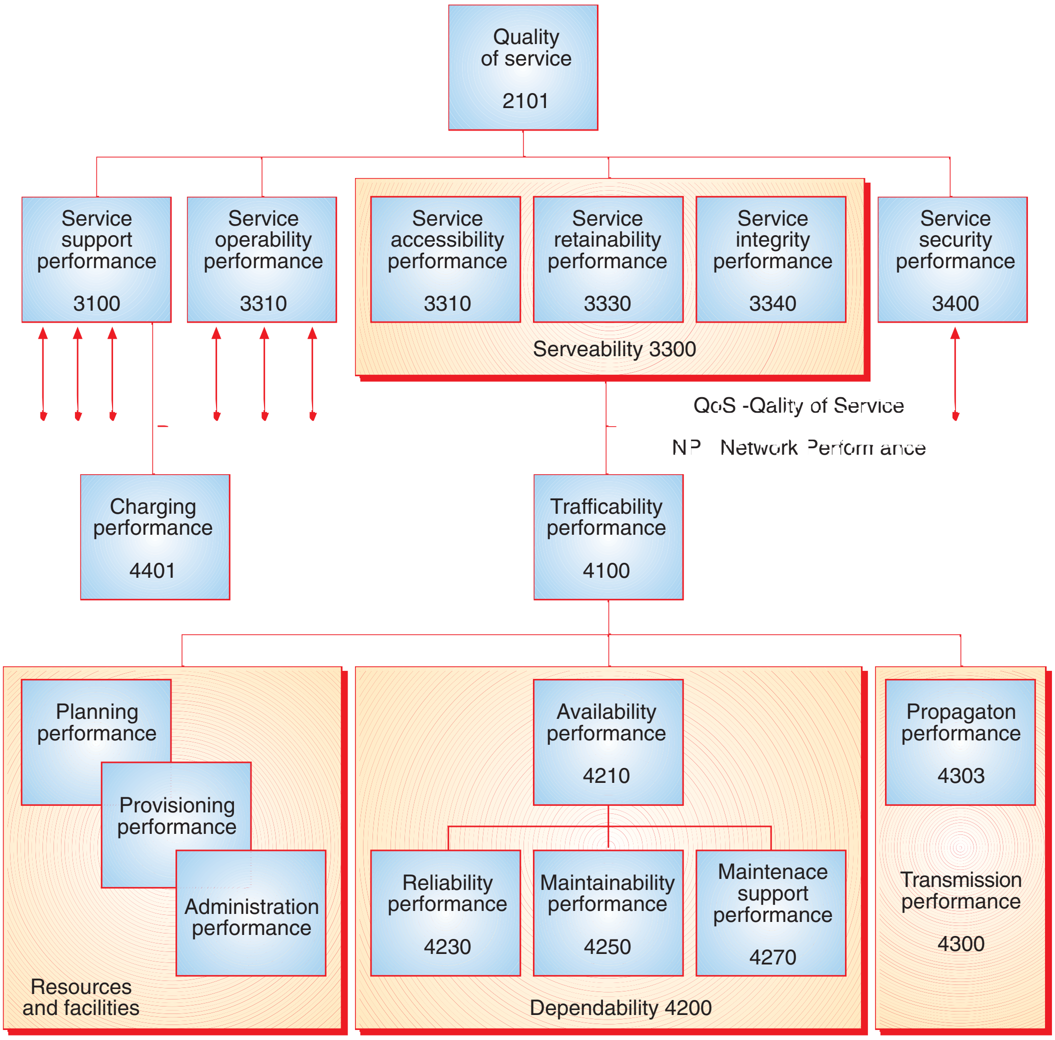

Quality of service 2101
Service support performance 3100
Service operability performance 3310
Service accessibility performance 3310
Service retainability performance 3330
Service integrity performance 3340
Service security performance 3400
Serveability 3300
QoS -Qality of Service
NP Network Performance
Charging performance 4401
Trafficability performance 4100
Planning performance
Provisioning performance
Administration performance
Resources and facilities
Availability performance 4210
Reliability performance 4230
Maintainability performance 4250
Maintenace support performance 4270
Dependability 4200
Propagaton performance 4303
Transmission performance 4300

Figure 3.1 Performance concepts

guide to the factors which contribute collectively to the overall quality of service as perceived by the user of a telecommunication service. The terms in the diagram can be thought of as generally applying either to the quality of service levels actually achieved in practice, to objectives which represent quality of service goals, or to requirements which reflect design specifications.

The figure is structured to show that one quality of service factor can depend on a number of others. It is important to note that the value of a characteristic measure of a particular factor may depend directly on corresponding values of other factors which contribute to it. This necessitates, whenever the value of a measure is given, that all of the conditions having an impact on that value be clearly stated.

The essential aspect of the global evaluation of a service is the users' opinion of the service. The result of this evaluation expresses the users' degree of satisfaction. This recommendation provides a framework for

- The quality of service concept
- Relating quality of service and network performance
- A set of performance measures.

It is obvious that a service can be used only if it is provided, and it is desirable that the provider has a detailed knowledge about the quality of the offered service. From the provider's viewpoint, network performance is a concept by which network characteristics can be defined, measured and controlled to achieve a satisfactory level of service quality. It is up to the service provider to combine different network performance parameters in such a way that the economic requirements of the service provider as well as the satisfaction of the user are both fulfilled.

In the utilisation of a service the user normally identifies the service provider only. The user's degree of satisfaction with the service provided depends on quality of service, that is on the perception of the latter of the following service performances:

- The support
- The operability
- The servability
- The security.

All are dependent on network characteristics. However, the servability performance is the most generally affected. It is further subdivided into three terms:

- Service accessibility performance
- Service retainability performance
- Service integrity performance.

Servability performance depends on trafficability performance and its influencing factors of resourcing and facility, dependability and transmission performance (of which propagation performance is a subset), as shown in Figure 3.1. The trafficability performance is described in terms of losses and delay times. Dependability is the combined aspects of availability, reliability, maintainability and maintenance support performance and relates to the ability of an item to be in a state to perform a required function. Propagation performance refers to the ability of the transmitting medium to transmit the signal within intended tolerances.

The resources and facilities box includes planning performance, provisioning performance and the related administrative functions. This highlights the importance of the network planning and provisioning aspects, etc. to the overall quality of service results.

3.3 Dependability definitions

While dependability is used only for a general description in non-quantitative terms, the actual quantification is done under the heading of availability performance, reliability performance, maintainability performance and maintenance support performance.

The properties expressed by these measures impact the measures relating to quality of service and network performance and are thus implicitly characterizations of these performance measures.

Measures are connected to events (failure, restoration, etc.), states (fault, up state, down state, outage, etc.) or activities (e.g. maintenance), with their time duration.

In the following definitions a number of essential concepts within the area of dependability are defined (in italics). For clarification, some of the concepts are accompanied by additional comments.

Availability performance

The ability of an item to be in a state to perform a required function at a given instant of time within a given time interval, assuming external resources, if required, are provided.

Note 1 – This ability depends on the combined aspects of the reliability performance, maintainability performance and maintenance support performance, see Figure 3.1.

Note 2 – The term availability is used as an availability performance measure.

Additional comment

The term availability performance defines the availability of an item. The item can be a single component, a transmission system or a telecommunication network. In more complex networks the resulting availability performance will depend on several conditions. This is illustrated in Figure 3.2.

For a complex system the resulting availability performance will thus be determined by the following 3 elements:

- Availability performance for an item

  The availability performance for an item will depend on the reliability performance of an item expressed in MTBF (mean time between failure) and the total maintenance aspects expressed in MTTR (mean time to repair).

- Influence from "unnormal events"

  The risk analysis will determine the influence from unnormal events like fire, sabotage, breakdown of external

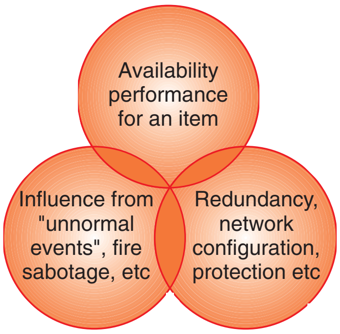

Availability performance for an item
Influence from "unnormal events", fire sabotage, etc
Redundancy, network configuration, protection etc

Figure 3.2 Illustration of the term resulting availability performance for a complex system

power supply, climatic systems, fault made be excavators, disturbances from EMC (electromagnetic compatibility), etc.

- Redundancy, network configuration, etc.

  This will depend on mechanisms like duplicating of equipment, network structures and different protection/restoration mechanisms.

Reliability performance

The ability of an item to perform a required function under given conditions for a given time interval.

Note 1 – It is generally assumed that the item is in a state to perform this required function at the beginning of the time interval.

Note 2 – The term reliability is used as a measure of reliability performance.

Additional comment
The reliability performance describes the ability of the item to be in an operative state (without failures) and is the parameter stated and guaranteed by suppliers. The guarantee is based upon the assumption of normal conditions regarding power supply, temperature, humidity, etc. Mean Time Between Failure (MTBF) and Mean Failure Intensity (MFI) are used as measures of the reliability performance.

Maintainability performance

The ability of an item under stated conditions of use, to be retained in, or restored to a state in which it can perform a required function, when maintenance is performed under given conditions and using stated procedures and resources.

Note – The term maintainability is used as a measure of maintainability performance.

Additional comment
The maintainability performance describes how easy the item can be repaired when failure occurs and how failures can be prevented. The maintainability performance is influenced by: Simple fault repair (e.g. alarms indicating the failure), simple fault repair by replacing cards and possibilities of preventive maintenance. Installing of new software will often introduce interruptions. In this case it represents decreased maintainability performance.

Maintenance support performance

The ability of a maintenance organization, under given conditions, to provide upon demand the resources required to maintain an item, under a given maintenance policy.

Note – The given conditions are related to the item itself and to the conditions under which the item is used and maintained.

Additional comment
The maintenance support performance describes the ability of a maintenance organization to repair failures. A number of elements will influence this ability: The supervision of the item, time to detect failure, sufficient maintenance staff, the competence of the staff, working hours of the staff, spare part storage, service agreements with central maintenance centre, etc.

Dependability

The collective term used to describe the availability performance and its influencing factors: reliability performance, maintainability performance and maintenance support performance.

Note – Dependability is used only for general descriptions in non-quantitative terms.

Additional comment
The term dependability is used to describe several characteristics, related to both equipment and maintenance. In addition, this term has a specific mathematical definition. To avoid confusion we recommend that the term dependability only be used as a collective term for the sphere when availability performance is discussed. The specific characteristics of an item relevant to its function will be described by the terms reliability performance and maintainability performance.

Risk analysis (SBA)

The risk analysis (SBA) expresses the risk of unwanted events presented to people, environment or equipment. In SBA the risk is expressed by the probability of an unwanted event and the consequence of the event.

Additional comment
The risk analysis (SBA) is a systematic way of describing and calculating the risk of unwanted events and is performed by doing a survey of the unwanted events which affect a service.

In the risk analysis (SBA) the following formula will be used:

$$Risk = frequency \cdot duration \cdot amount$$

where

$$duration \cdot amount = consequence.$$

Instantaneous availability

The probability of an item to be in a state to perform a required function at a given instant of time within a given time interval, assuming external resources, if required, are provided.

Additional comment
Instantaneous availability expresses the probability of an item to be error-free in a certain time interval and to be able to perform the required functions according to the specification.

Asymptotic availability

The limit, if it exists, of the instantaneous availability when the time tends to infinity.

NOTE – Under certain conditions, for instance constant failure rate and constant repair rate, the asymptotic availability may be expressed as:

$$A = \frac{MUT}{MUT + MDT}$$

where MUT and MDT are the mean up time and mean down time, respectively.

Availability

The term availability is used as a measure of the availability performance.

Additional comment
Asymptotic availability is the term describing the commonly used expression availability, which is a measure of availability performance. Availability denotes the part of a certain time period, i.e. one

year, when the item or system is available. Often, the complementary expression unavailability or mean accumulated down time (MADT) is used, where MADT is the accumulated down time in one year.

Unavailability can be expressed as:
$U = MADT$ (hours/year)

Unavailability can also be expressed as:
$U = (MADT / 8760) \cdot 100\ \%$

where 8760 is the number of hours in one year.

Availability can then be expressed as:
$A = 100\ \% - U$
$A = 100\ \% - (MADT / 8760) \cdot 100\ \%$

Unavailability and availability are often stated for a time period of one year.

4 Methods for dependability planning

4.1 Introduction

There are two fundamentally different methods for planning the dependability of a telecommunication network, here denoted method A and method B. In method A absolute requirements are applied to the different parts of the network. The network must then be designed to meet these requirements.

Method B is based on a valuation of lost traffic. The "profitability" of actions to improve the dependability is evaluated by comparing the costs of the action with the value of the traffic saved by performing the action.

The methods are described in more detail below.

4.2 Method A (absolute requirements)

In method A hypothetical reference circuits for connections through the network are used. Requirements are given for each separate part of the reference connection. A starting point may be that a connection shall satisfy a given end-to-end requirement. The requirements for the different parts of the connection are then determined in such a way that the end-to-end requirement is satisfied. Alternatively, the actual availability for the different parts of the connection is determined, and the requirements are fixed to comply with this. The end-to-end availability of the connection can then be calculated based on the requirements for the different parts of the connection.

Method A is in line with the methods traditionally used for transmission planning and results in a "worst case" description of the end-to-end availability.

When different telecommunication services are transported in the same network and different end-to-end availabilities are specified for each service, the resulting requirement for a particular part of the network will be determined by the service having the most stringent availability requirement.

Method A is used in I-ETS 300 416: "Availability performance of path elements of international digital paths", where availability requirements are given for the different parts of the international transport network, and also in ITU-R Rec. 695 and 696 where requirements for radio transmission systems are given.

Method A was recommended to be used in dependability planning in Telenor in the Dependability Handbook (1980).

4.3 Method B (economical method)

Method B is based on principles defining the object of dependability planning. The principles are realized through a quantitative model. The level of dependability is deduced by applying the model, taking into account all relevant factors in each planning case.

- Basic principle: The main objective is to find a balance between the customers' needs for dependability and their demand for low costs, based on solutions that are economically satisfactory for Telenor.
- Model: The consequences of a fault are expressed in terms of money and are included as additional costs in planning and cost-optimization. These additional costs reflect the customers' experience of faults in the network, quantified in terms of money, as well as the network operators' lost traffic revenue and the costs of corrective maintenance.
- Application: The method gives Telenor a method to integrate dependability as a natural part of planning and to take local information into account in each planning case. The method enables the preparation of simplified planning rules (thumb rules).

Method B is based on a valuation of the traffic lost because of a fault, given by:

1. The loss of revenue for the network operator caused by traffic not returning after the fault has been corrected.
2. An assessment of the economical loss of the average customer because of the affected traffic, given as Norwegian kroner (NOK) per Erlanghour (NOK/Eh).
3. A price tag reflecting the annoyance experienced by the average customer. This is important in a competing market.

The sum of 1. and 2. should reflect the price the average customer is willing to pay to avoid one Erlanghour of offered traffic delayed or lost due to a fault. The value of 1. will usually be very small compared to the sum of 2. and 3.

When valuating the lost traffic, it is possible to differentiate between different customer categories (business/residential) and between what services are affected by the fault. It is therefore possible to take account of this in those parts of the network where it is possible to obtain information about the customers and services affected. In other parts of the network, where such information is unavailable, the valuation of lost traffic must be based on an assessment of average traffic flows through the network.

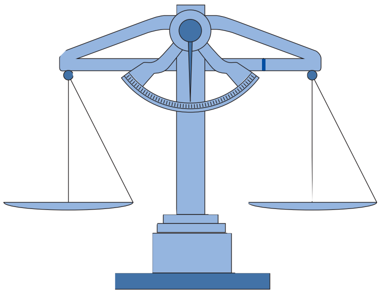


Figure 4.1 The main objective of dependability planning is to find a balance between the customers' needs for dependability and their demand for low costs

The valuation of lost traffic is based on the customers' assessment. The valuation used in dependability planning may, however, be adjusted according to the level of dependability chosen by Telenor. In addition to the factors 1. – 3. above, negative exposure in the media when faults occur may be taken into account when valuating lost traffic for special customers. The resulting valuation of lost traffic should therefore reflect the weight Telenor gives to this special service or customer category. The level of dependability in the network will increase when the value of lost traffic is increased.

Method B automatically gives priority to the dependability in those parts of the network having the largest traffic and for customers having large traffic and thus being most attractive for competing network operators.

For a given part of the network, the value of lost traffic (C_t) during a year, can be found from the formula:

$$C_t = \tau \cdot z \cdot A \cdot P \cdot c = MADT \cdot A \cdot P \cdot c$$

The factors are:

τ	- Duration of a fault, downtime per fault (hours/fault)
z	- Mean failure intensity (failures/year)
A	- Traffic intensity (Erlang, E)
P	- The part of the traffic being affected in the time τ
c	- The value of lost traffic (NOK/Eh)
MADT	$= \tau \cdot z$ (Mean Accumulated Down Time, hours/year)

A more comprehensive introduction to method B is stated in ITU-T Recommendation E.862 and will be presented in Chapter 5.

4.4 Comparison of method A and method B

Applicability

Method A:
When the requirements are fixed by the responsible department, the method is simple to use, but gives little flexibility.

Method B:
In addition to calculating the availability for different alternatives, it will be necessary to calculate the traffic volume affected by faults. The method is therefore somewhat more complicated than method A.

Determining requirements

Method A:
Determining the requirements for different parts of the network will, to a large degree, be based on rough estimates and will in practice be based on an estimate of what availability can be achieved by "acceptable" costs. The requirement will therefore often reflect "worst case" situations.

Method B:
There are no requirements for the different parts of the network. The costs for actions to improve the availability will in each case be compared with the value of the traffic affected by a fault. Actions will therefore be taken where the "profit" is largest.

Costs

Method A:
Because requirements are given for the different parts of the network, an increase of the availability exceeding the "worst case" situation will incur large costs, especially in sparsely populated areas where the traffic affected is small.

Method B:
The costs for actions to improve the availability will be weighted against the customer's valuation of traffic affected by a fault, so that a balance between the costs for the actions and the customers' need for availability (willingness to pay) is achieved.

Resulting availability (end-to-end, declared values)

Method A:
Based on the requirements for different parts of the network, the end-to-end availability may be calculated for different services. Such a calculation will, however, represent a "worst case" situation giving a bad impression of the availability, and a much lower availability than experienced by most customers. Instead, the availability for different services ought to be statistically calculated, based on knowledge of the network.

Method B:
The resulting end to end availability will generally be higher than for method A, because actions to increase the availability will be done where it is considered to be "profitable". The availability for different services will have to be statistically calculated, based on knowledge of the network. The availability will be higher for important customers having large traffic and in centrally situated areas, than by using method A, while customers in sparsely populated areas and customers having little traffic, will have about the same conditions as they would have if method A was used.

Emergency services

For both method A and method B it will be necessary to take special precautions concerning the availability of emergency services.

Flexibility

Method A:
The method is static and changing requirements and guidelines require large resources.

Method B:
The method is flexible. Changing the basis for planning is simple, and may be achieved by changing the value of traffic affected. The method is independent of changes in the traffic pattern and the design of the network.

4.5 Choice of method

Both methods are based on a calculation of the availability. In some parts of the network this may be complicated and require some labour.

Method A is simple to apply. Because the method sets down requirements for the different parts of the network, the result will be that weak requirements which can be achieved with low costs, will be chosen. The result will be a lower availability than what is desirable in densely populated areas and for important customers having large traffic. The method is static and changing requirements and guidelines require large resources.

Method B is somewhat more complicated to apply, but gives a rational basis for choosing solutions in the network, based on economic considerations. Method B has the great advantage that the dependability is expressed in terms of money and can be incorporated as part of an ordinary investment analysis. The method is simple and the basis for the considerations can easily be changed by changing the value of traffic affected. The method results in priority being given to those parts of the network having large traffic and to customers who are attractive to competing operators. The method is especially suitable for considering guidelines for combinations of actions, to increase the availability performed in different parts of the network, and makes it possible to evaluate the combination of actions in the transport network and in the service network. Method B will also give a basis for reconsidering existing guidelines and more causal actions (increasing the costs), but not being economically justified.

The project team concluded in the final report of Milestone M2 *"Develop principles and methods to choose availability actions in the planning of transport and service networks",* that method B should be introduced as the primary method for evaluating actions to improve the availability and for establishing the level of availability in the transport network and service networks of Telenor.

5 Dependability planning of telecommunication networks

5.1 Introduction

In Chapter 4 two different methods of dependability planning are briefly described. These two methods are denoted method A (absolute values) and B (economical method). In method A absolute requirements are applied to the different parts of the network. The network must then be designed to meet these requirements.

Method B is based on a valuation of lost traffic. The "profitability" of actions to improve the dependability is evaluated by comparing the costs of the action with the value of the traffic saved by performing the action.

The dependability project in Telenor recommended that method B should be used as a basis for dependability planning in Telenor. This recommendation was approved by the Planning Department in Telenor Network in June 1995. Method B is based on the principles of ITU-T Recommendation E.862 "Dependability Planning of Telecommunication Networks".

Below follows a description of the main parts of ITU-T Recommendation E.862.

The Recommendation is concerned with models and methods for dependability planning, operation and maintenance of telecommunication networks, and the application of these methods to the various services in the international network. ITU-T Recommendation E.862 was prepared by Study Group II and was approved under the Resolution No. 2 procedure on 16 June 1992.

The main reasons for developing this Recommendation are stated below:

- Economy is often an important aspect of dependability planning.
- The ability of achieving a certain level of dependability differs between network providers.
- Network providers often operate in a competitive environment.
- There is a need for establishing a method for dimensioning and allocating dependability in the telecommunication network.

ITU recommends that the procedures defined in Recommendation E.862 shall be used by administrations to plan, design, operate and maintain their networks.

5.2 General

The method described in E.862 is an analytical method based on principles defining the object of dependability planning. The principles are realized through a quantitative model. The level of dependability is deduced by applying the model, taking into account all relevant factors in each planning case.

Basic principle

The main object of dependability planning is to find a balance between the customers' needs for dependability and their demand for low costs.

Model

Fault consequences are expressed in terms of money and are included as additional cost factors in planning and cost-optimization. The cost factor reflects the customers' experience of faults in the network, quantified in terms of money, as well as the administration's costs for lost traffic revenue and corrective maintenance.

Application

The administration is provided with a method to integrate dependability as a natural part of planning, taking local information from the actual planning case into account. This method enables the preparation of simplified planning rules.

The application of the analytical method gives, economically, the best-balanced level of dependability, seen from the customers' point of view. This reduces the risk of customers' complaints and loss of business to competitors as well as the risk of unnecessary investments. It is, therefore, considered to be the best general way of planning dependability for the administration, as well as for the customers.

Recommendations for operational dependability objectives are needed in order to discover impairments and to check and compare dependability performance in the national and international network. Experience from the application of the analytical method may give reason to revise existing recommendations.

5.3 Generic measures for dependability planning

The dependability is described by measures defining the availability performance, the reliability performance and the maintainability performance of the network and its constituent parts as well as the maintenance support performance (for the maintenance of the network). The recommended measures are:

a) **Availability performance**
- Mean accumulated down time

b) **Reliability performance**
- Mean failure intensity

c) **Maintainability performance**
- Mean undetected fault time
- Mean time to restoration
- Mean active repair time

d)Maintenance support performance

- Mean administrative delay
- Mean logistic delay.

5.4 Planning for economic optimum

5.4.1 Economic dimensioning and allocation method

The main principle of dependability planning is to find actions (investments, increased maintenance, etc.) which maximise the total profit of the network:

max $\{LCR - LCC\}$

LCR – life cycle revenue
LCC – life cycle cost.

The revenues and costs are judged either by their effect on the operating company's (administration's) ability to reach its goals (a commercial evaluation) or by their effect on the welfare of all the members of society (a social evaluation). The principles for evaluating dependability may change and are to be regarded as a national matter.

An equivalent statement of the problem is to find actions that minimise the present value of the total costs of the network:

min $\{C_I + \Sigma\,(C_t + C_m + ...) \cdot d_i\}$

where:

C_I is the investment costs to achieve a certain degree of dependability

C_m is the expected maintenance costs of year *i*

C_t is the expected traffic disturbance cost (loss of revenue) for year *i*

d_i is the discount factor for calculating present value of costs occurring in year *i*.

C_t reflects the annoyance caused by faults and should be regarded as the basic service parameter which dimensions dependability in the network. A decrease in traffic disturbance cost represents an increase in life cycle revenue $(\Delta LCR = -\Delta C_t)$.

Unlike quantitative objectives for dependability performance (the intuitive method), this method is generally applicable and does not become out of date with technological advances, changes in cost structure, etc. Dependability is converted into one clear-cut measure (money) which makes it easier to evaluate actions to promote dependability and to compare and choose between different alternatives. The method is applicable for planning all parts of the national and international network and for dimensioning the dependability of network components and the level of maintenance support. It may be used in short and long term planning as well as to quantify scenarios in strategic planning.

5.4.2 A simplified model for quantifying traffic disturbance costs

The annual traffic disturbance cost is given by the interruption costs of circuit and packet switched traffic (first and second terms) and interruption costs of leased lines (last term):

$$C_t = z \cdot T \cdot E \cdot \alpha \cdot A \cdot c_s \\ + z \cdot T \cdot \lambda \cdot \beta \cdot r \cdot c_p + z \cdot T \cdot n \cdot c_l$$

where:

z is the failure intensity (failures per year)

T is the mean down time (hours)

A is the busy hour intensity of switched traffic (erlangs)

α is the factor reflecting the fraction of busy hour traffic demanded during the fault

E is the probability of congestion during the fault

c_s is the economic valuation of switched traffic (monetary unit per erlang-hour)

λ is the intensity of busy hour packet calls (packets per hour)

β is the factor reflecting the fraction of busy hour packet calls demanded during the fault

r is the probability of packet loss or delay during the fault

c_p is the economic valuation of a lost or delayed packet (monetary unit per packet)

n is the number of leased lines

c_l is the economic valuation of interruption of a leased circuit (monetary unit per circuit hour).

The model assumes that the parameters are stochastically independent and do not vary in time. However, this is seldom the case. If failures are more likely to occur at certain hours of the day, there may be a correlation between traffic and failure intensity (time is a common parameter).

Down time may be dependent on the time of day or week when the failure occurs. Correlations between parameters can be dealt with by assuming models of time variations of traffic, failure intensity, down time, etc. The problem is simplified if failures are assumed to be uniformly distributed in time. The fraction of busy hour traffic demanded during the fault is then equivalent to the average traffic and the values of α can be calculated if the traffic profile is known. ITU-T Recommendation E.523 defines standard traffic profiles for international traffic streams.

If only a fraction of the capacity is lost, the result is a state of increased congestion. The average probability of congestion or packet delay during a fault depends on the transmission capacity left and the traffic profile.

5.4.3 Economic assessment of disturbed traffic volume *(c)*

The factors c reflects the level of ambition of an Administration in dependability planning. High values of *c* will give a high level of dependability and vice versa. The objectives of the operating company (commercial or social) may influence the values. Important factors:

- The customers' willingness to pay for dependability
- The market structure (degree of competition, etc.)
- The category of customers and services affected
- The degree of congestion, delay or transmission disturbance
- The duration of the fault
- The accessibility to alternative communication means for the affected customer
- Time of day, week or year when the fault is in effect
- How often faults have occurred in the past, etc.

Administrations are recommended to make their own investigations among their customers in order to determine the values to be used in planning. If this is not possible, rough estimates may be obtained from information about actions taken previously in the network. The cost of actions is compared to the amount of traffic saved. Actions that intuitively are regarded as reasonable, give a lower limit of *c*, actions that obviously are unreasonable give an upper limit. The values derived in this way are then used under the assumption that they are valid also for planning the future network. If *c* is not possible to estimate at all, the method may still be used to make priorities among competing alternatives and thus roughly find an optimum allocation of a given amount of resources.

5.4.4 Planning procedure

Traffic disturbance costs are included as additional cost-factors in economical calculations for planning, thus integrating dependability as a natural part of planning.

The procedure of dependability planning is performed in four steps:

Step 1:
Plan a network attaining functional and capacity requirements
The starting point is a network planned and dimensioned in order to comply with the functional and capacity requirements, but without special consideration of dependability (zero-alternative). The second step is to identify what changes may be necessary to promote dependability.

Step 2:
Search for actions to promote dependability
There is a need for actions to promote dependability if traffic disturbance costs are high or if the actions can be taken at a low cost. A non-exhaustive list from which actions could be identified is given below:

- Protection of equipment in order to prevent failures
- Choice of reliable and maintainable equipment
- Modernization and reinvestment of worn out equipment
- Redundancy
- Overdimensioning
- Increase in maintenance support
- Network management actions to reduce fault effects.

Step 3:
Analyse the actions
Express improvements in terms of changes in traffic disturbance and maintenance costs ($\Delta C_t + \Delta C_m$) for each action. It is only necessary to calculate costs that differ between the alternatives.

Compare $\Delta C_t + \Delta C_m$ to the increased investment costs (ΔC_I) for each action, e.g. by the present value method.

Choose the best set of actions, i.e. the one which gives the lowest total cost.

Step 4:
Check that minimum requirements are complied with
A minimum service level may be stipulated by governmental regulations, by ITU-T Recommendations, for commercial or for other reasons. The establishment of any minimum requirements on the national level is a national matter. For planning of the international network the administration is recommended to check if dependability objectives deducible from existing ITU-T Recommendations are met. If not, the reasons for non-compliance should be examined more closely. If it is justified, the level of dependability should be adjusted.

5.4.5 Numerical example based on the above

In the following examples a Currency Unit (CU) is used as the monetary unit.

Step 1:
Network planned without special consideration of dependability
The network studied is the trunk between two exchanges as shown in Figure 5.1. The traffic between the two exchanges is *A* (Eh), the failure intensity *z* (failures/year) and the mean down time *T* (hours).

Step 2:
Search for actions to promote dependability
The action considered is to introduce a physically redundant cable as shown in Figure 5.2. The redundant cable is assumed to be dimensioned to carry the whole traffic load, i.e. a single failure will not disturb the traffic. The failure intensity and the mean down time is assumed to be the same for both cables.

Step 3:
Analyse the action
Assumptions:

Failure intensity:
$z = 0.1$ failures/year

Mean down time:
$T = 24$ hours

Mean offered traffic:
$A = 100$ E

Congestion:
$P = 1$ (without redundancy)
$P = 0$ (with redundancy)

Valuation of disturbed traffic volume:
$c = 100$ CU/Eh

Discount factor (lifetime 25 years, interest 5 % per year):
$d = 14$

Maintenance cost per failure:
$c_m = 1{,}000$ CU/failure

Cost of redundant cable:
$C_I = 100{,}000$ CU

Calculations:

Traffic disturbance costs for network without redundancy:

$C_t = P \cdot A \cdot z \cdot T \cdot c$
$C_t = (1) \cdot (100) \cdot (0.1) \cdot (24) \cdot (100)$
$= 24{,}000$ CU per year
Present value
$C_t \cdot d = (24{,}000) \cdot (14) = 336{,}000$ CU

Traffic disturbance costs for network with redundancy (the possibility of simultaneous faults is negligible):
$C_t = 0$

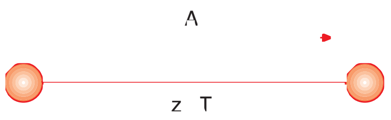


Figure 5.1 Trunk between two exchanges

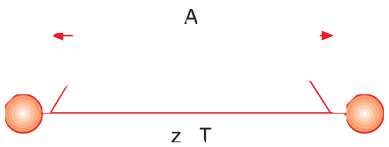


Figure 5.2 Trunks between two exchanges on physically redundant cables

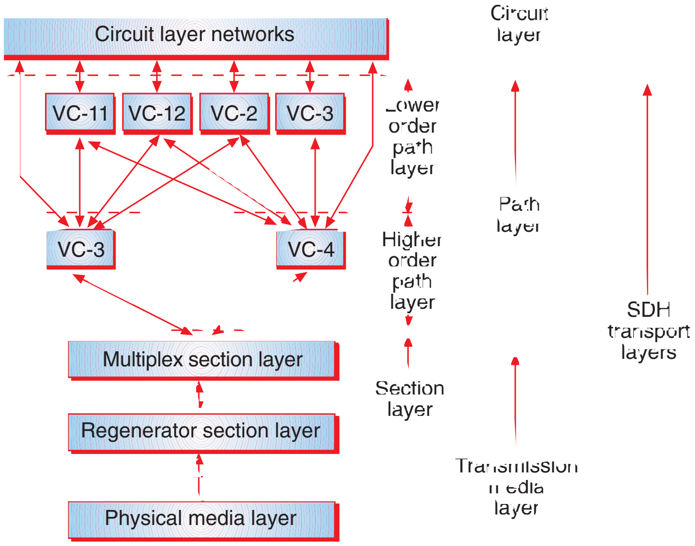


Figure 6.1 Layers in the SDH-based network

Change in traffic disturbance costs:
$\Delta C_t \cdot d = 0 - 336{,}000 = -336{,}000$ CU

Maintenance costs without redundancy:
$C_m = z \cdot c_m = (0.1) \cdot (1{,}000)$
$= 100$ CU per year
Present value
$C_m \cdot d = (100) \cdot (14) = 1400$ CU

Maintenance costs with redundancy:
$C_m = 2 \cdot z \cdot c_m = (2) \cdot (0.1) \cdot (1{,}000)$
$= 200$ CU per year
Present value $C_m \cdot d = (200) \cdot (14)$
$= 2{,}800$ CU

Change in maintenance costs:
$\Delta C_m \cdot d = 2{,}800 - 1{,}400 = 1{,}400$ CU
Cost reduction:
$\Delta C_t \cdot d + \Delta C_m \cdot d = -336{,}000 + 1{,}400$
$= -334{,}600$ CU

Change in total costs:
$\Delta C_I + \Delta C_m \cdot d + \Delta C_t \cdot d$
$= 100{,}000 - 334{,}600 = -234{,}600$ CU

Conclusion:
Since $\Delta C_I + \Delta C_m \cdot d + \Delta C_t \cdot d < 0$, the action is profitable. Whether or not it is optimal depends on whether there are alternative actions that are more profitable.

Step 4: Check minimum requirements

Any additional actions to meet minimum requirements, e.g. governmental requirements (for defence reasons, emergency, etc.), should be taken.

6 Protection/restoration methods in SDH-networks

6.1 Introduction

The telecommunication network of Telenor is the major carrier of all vital information which is distributed both domestic and abroad. The availability and the stability of the network is thus very important. The deregulation and the competition in the telecommunication market will strengthen the emphasis of the customers' needs and their willingness to pay for different services. The market demand for availability will be a critical factor in the growing competition. Availability and stability will be important in order to increase the ability of competition. Protection and restoration methods are elements used to achieve high availability performance.

In the PDH-network of Telenor today both DGROC (Digital Group Coupling) and PS (Protection Switch) are used.

DGROC (Digital Group Coupling)

- Consists of a central management system, a communication network and the switch
- Can be equipped with up to 16 traffic ports and 8 spare ports
- Operates only at the 140 Mbit/s level
- Automatic rerouting based on loss of signal (LOS), or alarm indication signal (AIS). Typical switching time is 10 seconds
- Operator based switching. Typical switching time is 1 second.

PS (Protection Switch)

- Consists of a switch and utilises the alarms and commands from the Alarm and Command system (A/C-system)
- Operates at 34 and 140 Mbit/s level (1+1)
- A single ended system (the two traffic directions are independent)
- Automatic rerouting based on LOS, bit error rate (BER) and AIS. Typical switching time is < 5 ms
- Operator based switching. Typical switching time is < 3 ms.

The introduction of SDH-networks represents a major improvement of the technology with increased functionality and mechanisms to maintain network protection. The challenge is to utilise the potential in the new technology and the human resources in order to develop cost optimum solutions for a reliable network.

6.2 Definition of protection and restoration

In the SDH technology the transport network is divided into three layers: The circuit layer, the path layer and the transmission media layer.

A further explanation of the layers is:

- A connection from the circuit layer is served from the path layer
- A connection from the path layer is served from the transmission media layer.

Figure 6.1 illustrates the relationship between the layers in the SDH-based network.

The division into layers in SDH introduces an opportunity to manage the network in a more simple and orderly way, also when failures occur. A fault can in a simple way be isolated to a specific part of the network and handled within that part.

The transport network of Telenor is divided into following sub-networks:

- Inter-regional network, where the nodes are connected to the secondary exchanges of the telephone network
- The regional network, where the nodes are connected to the local and primary exchanges of the telephone network

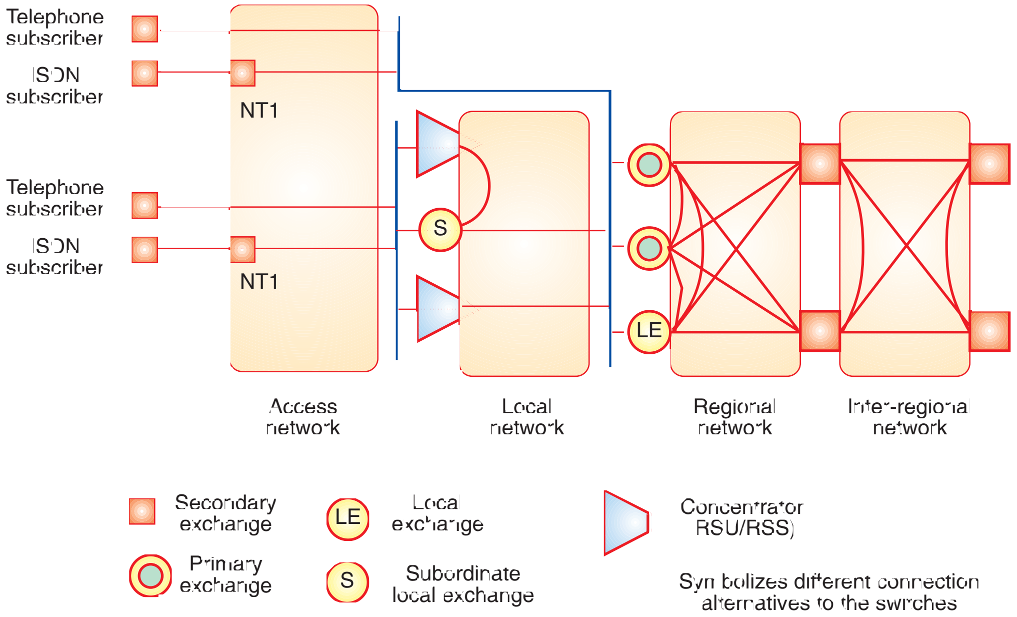


Figure 6.2 The transport and telephone network/ISDN of Telenor (schematic)

- The local network, where the nodes are connected to the concentrators and the local exchanges.

The different parts of the transport network and telephone network/ISDN of Telenor is shown in Figure 6.2. The outermost part of the network between the concentrators / local exchanges and the subscribers is denoted the access network, and is not considered to be a part of the transport network.

Within each sub-network specific protection or restoration mechanisms can be used in order to maintain the availability performance of the network. In the different types of SDH-equipment several protection mechanisms for rerouting to other transmission links are available.

Both protection and restoration are described in ITU-T Rec. G. 803. Following definitions (made by Telenor in order to avoid any confusion of the two terms) will be used:

Protection
The protection switch will be activated through alarms / fault states detected locally in the network element. The switching itself is completely independent from the Telecommunication Management Network (TMN). When the switching is performed, a message will be transmitted to TMN informing of this event. The TMN will thus be able to control all the routing of the traffic in the network.

Restoration
Restoration will be used only on the path layer. When alarms / fault states are detected by the TMN, both the fault states and the present status of the network will be analysed by the TMN. TMN transmits “coupling orders” to all relevant network elements. The affected traffic will be rerouted to spare connections. Restoration will be performed between the cross-coupling matrices in the network elements. When the switching is finalised a message will be transmitted to TMN in order to update the status of the network.

6.3 Capacity utilisation schemes

Three different schemes for capacity utilisation are identified when using protection and restoration mechanisms:

1 + 1, 1 : 1 and 1 : *n* scheme.

1 + 1 scheme (single-ended switching)

A fixed bridge on the transmitting side transmits the same signal to both the traffic and the spare route. The receiving end monitors the signal on both routes and can switch between the two signals. Each end switches independently of the opposite end. A 1 + 1 scheme will often be denoted as “single-ended switching”. In the case of a unidirectional fault, the two traffic routes will be transmitted through the network on different paths.

1 : 1 scheme (dual-ended switching)

1 : 1 scheme is denoted “dual-ended switching” because both ends switch when a fault occurs. A protocol is necessary to exchange the information between the two ends. It is possible to use the spare route for low priority traffic. If a fault occur on the traffic route (high priority traffic), the low priority traffic will be lost.

1 : *n* scheme (dual-ended switching)

One connection acts as the spare route for several traffic routes. This gives a

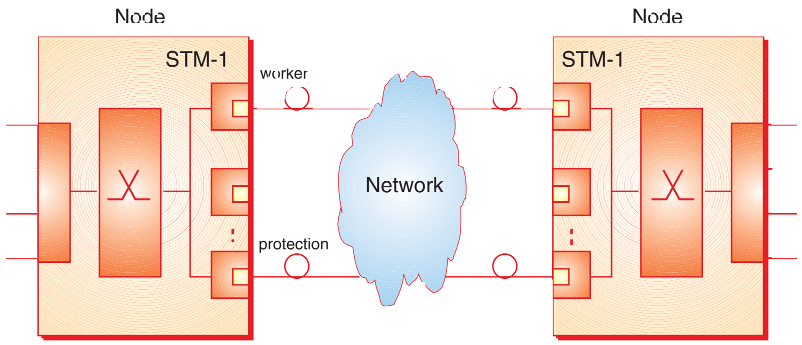


Figure 6.3 MSP 1+1 scheme

good utilisation of the capacity of the system, because only one spare route is allocated. The switching is “dual-ended”, and exchange of information between the ends is necessary through an information protocol (protection) or via TMN (restoration). The spare route can be used for low priority traffic when there are no faults in the system.

6.4 Protection and restoration schemes

There are mainly 4 different protection and restoration schemes. These are briefly described below. More details are given in the ITU-T Rec. G. 803.

MSP (Multiplex Section Trail Linear Protection)

MSP is a point-to-point protection mechanism which can be realised as 1 + 1, 1 : 1 or 1 : *n* protection schemes. Faults detected in MSOH (Multiplex Section Over Head) trigger the switch, and the payload in the transmission system (i.e. STM-1, STM-4 or STM-16) will be routed on a spare system in an alternative fibre path. The exchange of information between the two ends is performed through a standardised information protocol APS (Automatic Protection Switch protocol). An example of an MSP 1 + 1 scheme is shown in Figure 6.3.

MSP will manage failures both on the fibre, the laser and the receiving unit. MSP 1 + 1 and 1 : 1 schemes demand large capacity on the spare routes and will be very expensive to use on long distances.

MS SPRING (Multiplex Section Trail Shared Protection Ring)

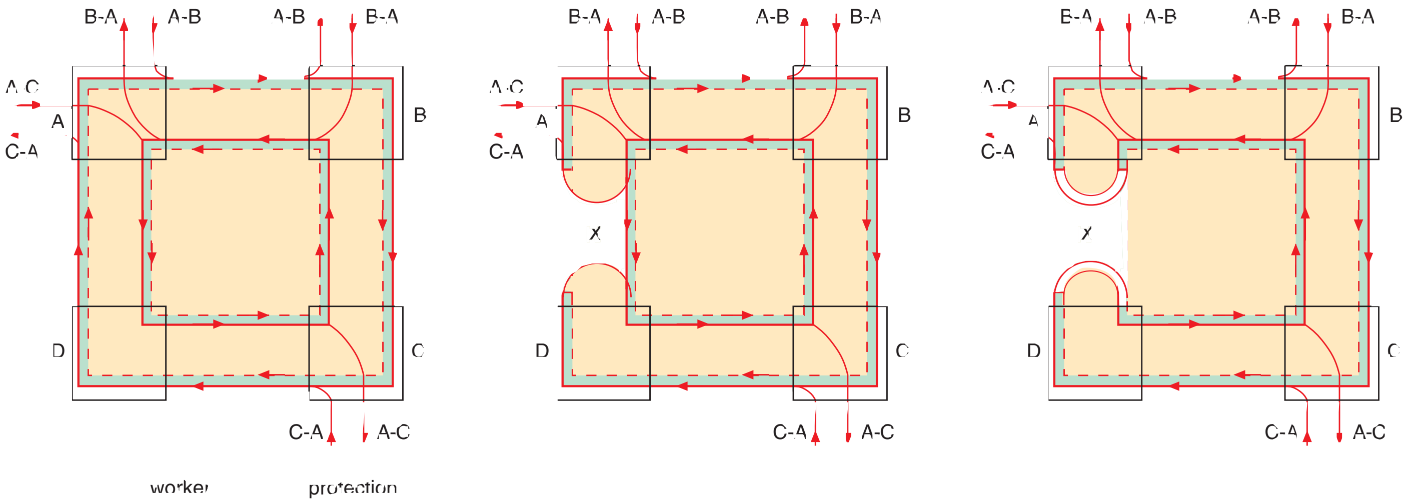
The MS SPRING scheme was earlier often denoted BSHR (Bi-directional Self Healing Ring). In an MS SPRING scheme the total Administrative Unit Group (AUG) payload capacity in a multiplex section is divided between the traffic and the spare capacity.

In a two-fibre ring half of the capacity (N/2 AUG) is used for traffic and the other half for the spare route. Faults detected in SOH (Section Over Head) trigger the switch. When a failure occurs, the network elements on both sides of the detected failure, loop the involved connections (AUG) to the spare route (dual-ended).

A two-fibre MS SPRING scheme is shown in Figure 6.4. MS SPRING is a 1 : *n* protection scheme, and 100% of the traffic in the ring is protected.

VC Trail Protection

VC Trail Protection is often named Path Protection. This is an end-to-end protection scheme, and the protection is valid between the two points where the actual trail is generated. Using VC Trail Protection, the traffic will be saved if a fault occurs or if the quality of the connection



a) Normal traffic

b) Unidirectional failure

c) Bidirectional failure

Figure 6.4 Two-fibre MS-SPRING scheme

decreases beyond a certain level (defined in advance).

VC Trail Protection can be used on VC-4, VC-3 or VC-12 trails. Since the two switching points can interpret status in POH both for the traffic and the spare trails, it is possible to use 1 + 1, 1 : 1 and 1 : *n* schemes. An example of a 1 + 1 VC Trail Protection scheme is shown in Figure 6.5.

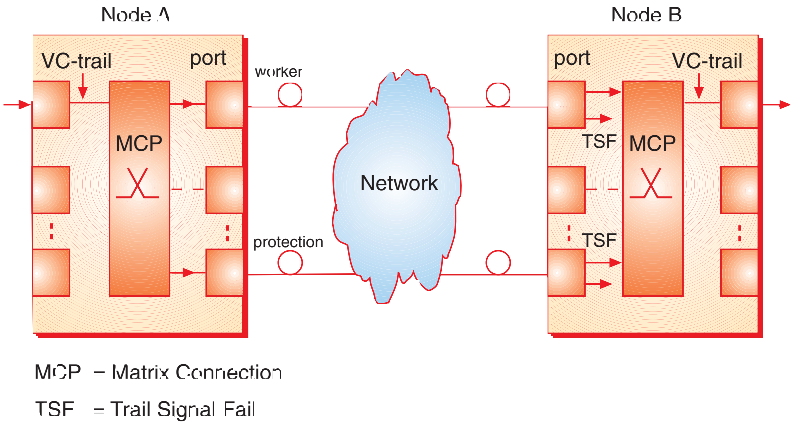
SNCP (Sub Network Connection Protection)

SNCP is a point-to-point protection scheme on the path layers and both 1 + 1, 1 : 1 and 1 : *n* schemes can be used.

The 1 + 1 SNCP scheme is a dedicated trail protection and the signal will be transmitted on both the traffic and the spare route, on two different paths. The switch in the receiving end operates based upon local information criteria. No APS protocol is so far specified and the SNCP scheme is single-ended. An example is shown in Figure 6.6.

The 1 : 1 SNCP protection scheme is based upon switching at both ends. An APS protocol is necessary. The spare route can be used for low priority traffic.

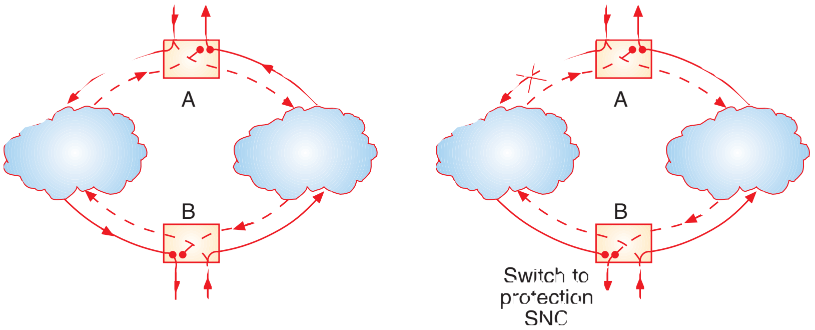
Using SNCP each single SNC can be configured on HOP and LOP, i.e. it is possible to protect a dedicated number of SNCs or all SNCs. One can choose whether an SNC shall be protected in parts of or through the whole SDH-network. SNCP can be used in all network elements containing switching matrices. SNCP is a very useful protection scheme expected to be used in a lot of SDH-networks.



MCP = Matrix Connection

TSF = Trail Signal Fail

Figure 6.5 1+1 VC trail protection scheme



a) Normal traffic

b) Unidirectional failure

Figure 6.6 1+1 SNCP protection scheme

7 Examples of dependability planning of telecommunication networks

7.1 Introduction

Below are given some examples of dependability planning in a telecommunication network. In the examples the economical method (method B) is applied to different situations and parts of the network. The following examples are shown:

- Establishing a parallel transmission system between a local exchange and concentrator in the local network
- Establishing a ring structure in the local network.

These two examples are based upon examples in the final report of Milestone M5: *"Develop practical planning rules for the transport and the ISDN/telephone network"*. In this final report a number of other examples from different parts of the network are shown.

7.2 Establishing a parallel transmission system between a local exchange and a concentrator

Case description

Between a concentrator (CONC) and a local exchange (LE) there is 5 km fibre cable. The concentrator has 2000 subscribers connected to it, each having an average two-way traffic of 0.07 Erlang in the busy hour. The total traffic between the concentrator and the local exchange is 140 Erlang in the busy hour. This situation is called alternative 0.

Will it be "profitable", using method B, to establish a parallel transmission system on a separate fibre cable? The situation with a parallel transmission system on a separate fibre cable is denoted alternative 1.

Dependability data:

Fibre cable
z = 0.002 failures/(km · year)
τ = 8.0 hours/fault

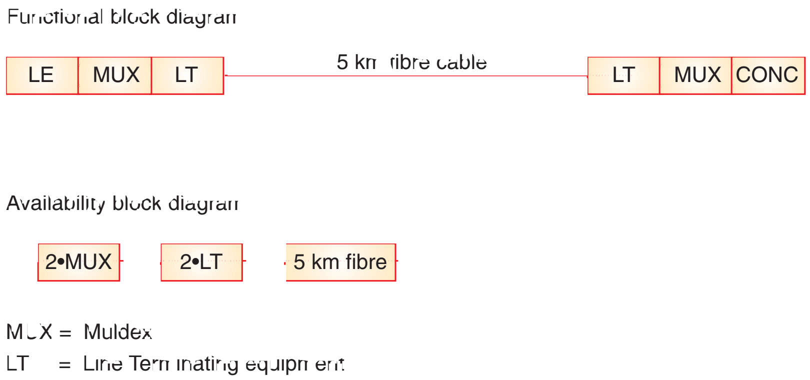


Figure 7.1 Single fibre cable

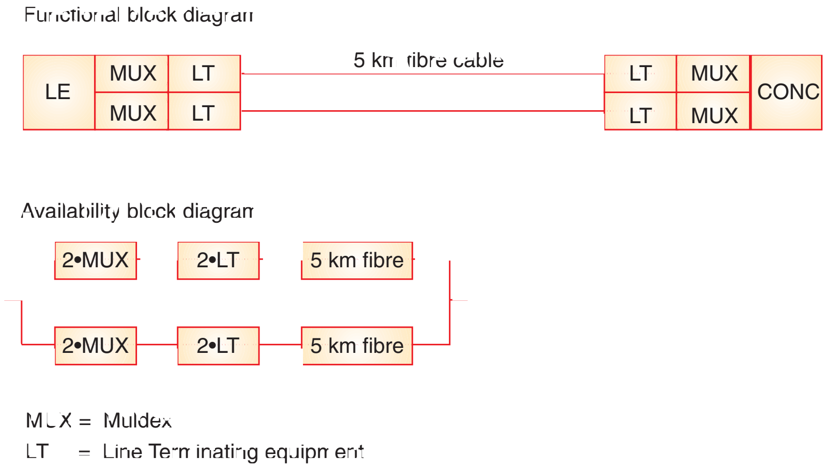


Figure 7.2 Parallel fibre cable

MUX
$z = 0.03$ failures/year
$\tau = 5.0$ hours/fault

LT
$z = 0.03$ failures/year
$\tau = 1.0$ hours/fault

Transmitted traffic
$A = 140$ E

Value of lost traffic
$c = 100$ CU/Eh

Calculations:

Mean accumulated down time *(MADT)*:

$$MADT = \Sigma z_i \cdot \tau_i = 5 \cdot 0.002 \cdot 8.0 + 2 \cdot 0.03 \cdot 5.0 + 2 \cdot 0.03 \cdot 1.0 = 0.44 \text{ hours/year.}$$

Mean failure intensity *(MFI)*:

$$MFI = \Sigma z_i = 5 \cdot 0.002 + 2 \cdot 0.03 + 2 \cdot 0.03 = 0.13 \text{ failures/year.}$$

Average lost traffic per hour is dependent on the utilization of the circuits *(U)* and relative circuit capacity *(k)*. Based on *U* and *k*, a traffic loss factor *(F)* giving the average traffic loss (per Erlang offered traffic) for one hour of unavailability in any hour of the year, may be found. The traffic loss factor will vary with the type of traffic, and we distinguish between residential traffic, business traffic and combined residential/business traffic. (The background for the traffic loss factor is explained in another article in this edition.) In this example the traffic loss factor for combined residential/business traffic is used because the customer population is unknown. It is assumed that failures occur with the same probability during a 24 hour period.

Lost traffic is calculated as: $C_t = Busy\ hour\ traffic \cdot MADT \cdot F$ (Eh/year).

For this alternative, all traffic will be lost when a failure occurs, i.e. a relative circuit capacity $k = 0$. Since $k = 0$, the utilization, *U*, has no influence. For combined traffic the traffic loss factor will be: $F = 0.375$. The value of lost traffic per year will be:

$$C_t = A \cdot MADT \cdot F \cdot c = 140 \cdot 0.44 \cdot 0.375 \cdot 100 = 2{,}310 \text{ CU/year.}$$

Maintenance costs are assumed to be: $c_m = 1{,}000$ CU/failure, i.e.

$$C_m = 1{,}000 \cdot 0.13 = 130 \text{ CU/year.}$$

The present value of lost traffic and maintenance costs are calculated using an economic time period of 10 years and a rate of interest of 7 % per year. This gives a present value factor equal to $C_{pv} = 7.02$.

Present value of lost traffic:
$C_t \cdot C_{pv} = 2{,}310 \cdot 7.02 = 16{,}216$ CU.

Present value of maintenance costs:
$C_m \cdot C_{pv} = 130 \cdot 7.02 = 913$ CU.

For alternative 0 the costs (present value) will be: 16,216 + 913 = 17,129 CU.

Alternative 1

In alternative 1 we will examine if it is profitable (according to method B) to establish a parallel transmission system on a separate fibre cable having a different routing. It is assumed that both transmission systems are capable of transmitting all the traffic.

Calculations:

Mean accumulated down time *(MADT)*:

$$MADT = (z \cdot t \cdot z \cdot t) / 8760 = (0.44 \cdot 0.44) / 8760 = 2.2 \cdot 10^{-5} \text{ hours/year}$$

Mean failure intensity *(MFI)*:

$$MFI = 2 \cdot z_i = 2 \cdot 0.13 = 0.26 \text{ failures/year.}$$

For this alternative, one will have a relative circuit capacity $k = 1$ in case of one failure. Assuming a utilization $U = 0.74$, the traffic loss factor for combined traffic will be: $F = 0$. This means that no traffic will be lost in case of one failure. Taking two simultaneous failures into account, one gets a relative circuit capacity of $k = 0$, giving a traffic loss factor for combined traffic: $F = 0.375$. The value of the lost traffic per year (C_t), will for alternative 1 be:

$$\begin{aligned} C_t &= A \cdot MADT \cdot F \cdot c \\ &= 2.2 \cdot 10^{-5} \cdot 140 \cdot 0.375 \cdot 100 \\ &= 0.12 \text{ CU/year} \end{aligned}$$

Maintenance costs are assumed to be: c_m = 1,000 CU/failure, i.e.

$$C_m = 1{,}000 \cdot 0.26 = 260 \text{ CU/year}$$

The same maintenance costs are used here as in alternative 0. In practice, one will in alternative 0 have to correct the faults immediately, so that overtime and extra costs for week-end work should be included. In alternative 1 one has complete back-up for the traffic, and therefore has the possibility to do the correction of faults in normal working hours, thereby reducing the maintenance costs.

Present value of lost traffic:
$C_t \cdot C_{pv} = 0.12 \cdot 7.02 = 0.84$ CU.

Present value of maintenance costs:
$C_m \cdot C_{pv} = 260 \cdot 7.02 = 1{,}821$ CU.

For alternative 0 the costs (present value) will be: 0.84 + 1,821 = 1,822 CU.

Conclusion

The difference in present value for lost traffic and maintenance between alternative 0 and alternative 1 will be

17,129 CU - 1,822 CU = 15,307 CU.

If the investments necessary to establish the parallel transmission system on a separate cable do not exceed 15,307 CU, the investment will be "profitable" according to method B and the investment should be done.

The transmission system can be established on a buried cable or on a suspended cable on existing posts.

Using suspended cable, the total investment will be about 90,000 CU.

For buried cable, the total investment will be 170,000 CU.

The investments for establishing a separate transmission system on a separate cable are much bigger than the gain in lost traffic and maintenance costs. The investment is therefore not "profitable" according to method B, and the investment should not be done.

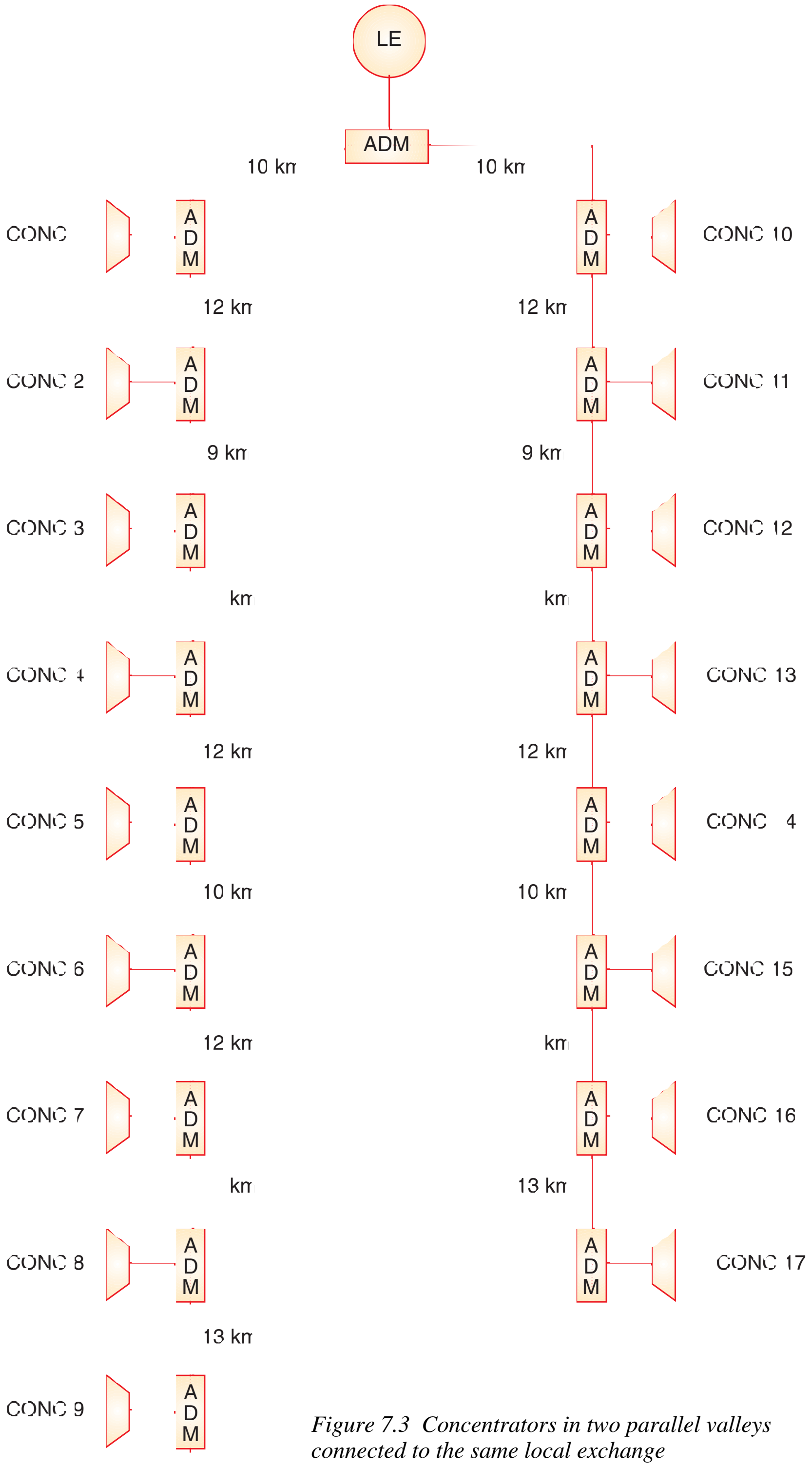


Figure 7.3 Concentrators in two parallel valleys connected to the same local exchange

Dependability data

Network element	Failure intensity (failures/year)	Down time (hours/fault)	Accumulated down time (hours/year)
ADM	0.025	4	0.10
Fibre cable	0.003 per km	8	0.024 per km

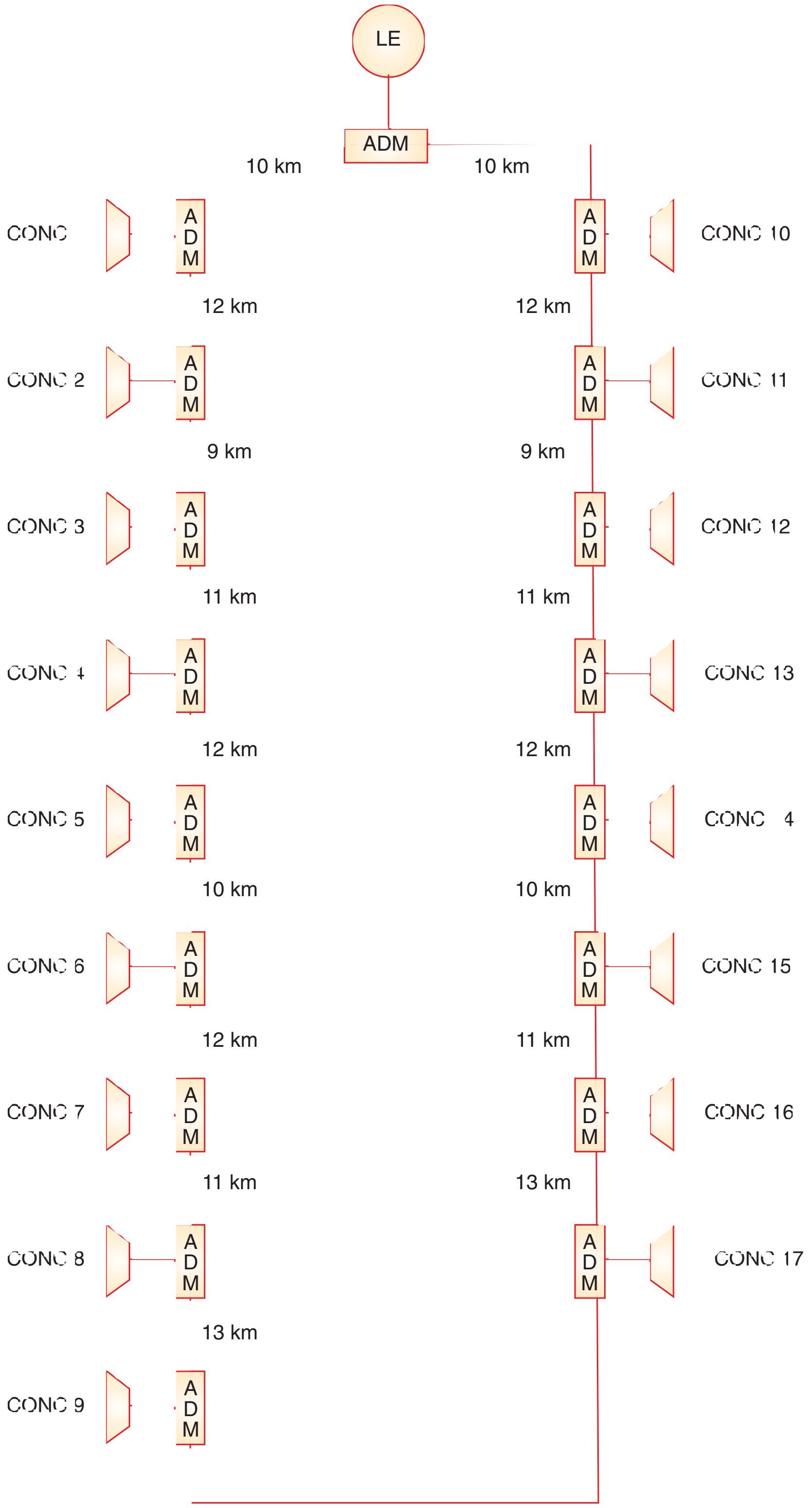


Figure 7.4 Concentrators in two parallel valleys connected to the same local exchange via an ADM ring

7.3 Establishing a ring structure in the local network

Case description

A local exchange (LE) has 17 concentrators (CONC) connected to it via two add-drop multiplexer (ADM) chains on fibre cables in parallel valleys, as shown in Figure 7.3. The figure also shows the distance between the ADMs.

Following alternatives will be evaluated:

Alternative 0:
The local network shown in Figure 7.3.

Alternative 1:
A ring structure is established by establishing 10 km fibre cable between concentrator 9 and 17, as shown in Figure 7.4. Diversity routing is used for the circuits between each concentrator and the local exchange.

Alternative 2:
A ring structure as in alternative 1 is established. 100 % back-up using subnetwork connection protection (SNCP) is established for the circuits between each concentrator and the local exchange. The back-up circuit is routed in the opposite direction of the circuit it is back-up for.

Traffic:
Traffic to/from each individual customer in the busy hour (*A*max) is assumed to be 0.07 E (two-way). Average lost traffic per hour is dependent on the utilization of the circuits *(U)* and relative circuit capacity *(k)*. Based on *U* and *k*, a traffic loss factor *(F)*, giving the traffic loss (per Erlang offered traffic) for one hour of unavailability in any hour of the year, may be found. The traffic loss factor will vary with the type of traffic, and one distinguishes between residential traffic, business traffic and combined residential/business traffic. (The background for the traffic loss factor is explained in another article in this edition.) In this example the traffic loss factor for combined residential/business traffic is used because the customer population is unknown. It is assumed that failures occur with the same probability during a 24 hour period.

The maintenance costs are assumed to be 1,000 CU per fault. In this example one will only consider the differences between the different alternatives. In practice, this means that the difference in maintenance costs will stem from the increase in the number of faults because of the establishment of 10 km fibre cable between concentrator 9 and 17. This gives an increase in the maintenance costs of 0.003 · 10 · 1,000 = 30 CU/year. This is negligible compared to the value of lost traffic. In the calculations one therefore neglects the maintenance costs.

The value of lost traffic is set to 100 CU per Eh. It is assumed a mixture of residential and business traffic (combined traffic).

Alternative 0

In this example it is assumed that not more than 300 customers are connected to each concentrator, giving a maximum traffic in the busy hour (Amax) of 21 E for any concentrator. This is equivalent to a dimensioning traffic of Ad = 21 / 1.05 = 20 E. Using congestion dimensioning, this means that a 2 Mbit/s circuit is necessary between the concentrator and the local exchange. Since the lowest capacity in the SDH system is 2 Mbit/s (VC 12), this means that one will have one 2 Mbit/s circuit between each of the concentrators and the local exchange. Totally, one will have 17 circuits (2 Mbit/s) in the local network. It will be sufficient to use an STM-1 system (with a capacity of 63 2 Mbit/s circuits).

In practice, one will in the local exchange have 2 DTMs for each concentrator, so that the number of DTMs will be: 3 · 17 = 51.

Equipment needed:

- 18 ADM-1s
- 51 DTMs.

This represents a cost of about 600,000 CU.

The availability of a concentrator is calculated using an availability block diagram. As examples, Figure 7.5 shows the availability block diagram for concentrator 1 and 17.

For alternative 0 one will have a complete loss of traffic in case of a fault in an ADM or on a cable section, i.e. a relative circuit capacity $k = 0$. This gives for combined traffic a traffic loss factor $F = 0.375$. Lost traffic is calculated as: *Busy hour traffic · MADT · F.*

Tables 7.1 and 7.2 show traffic, down time and the value of lost traffic *(C_t)* for each concentrator and totally for concentrators 1 to 9, and 10 to 17, respectively.

For all concentrators together, this gives the following value of lost traffic:

6,143.0 + 4,554.8 = 10,697.8 CU/year.

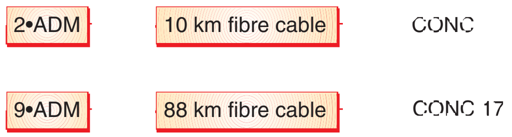


Figure 7.5 Availability block diagram for concentrator 1 and 17

Table 7.1 Value of lost traffic for concentrator 1 - 9

Concentrator	Number of subscribers	Traffic (Amax, E)	Down time (hours/year)	Traffic loss factor F	Value of lost traffic (C_t, CU/year)
CONC 1	100	7.0	0.44	0.3750	115.6
CONC 2	250	17.5	0.83	0.3750	543.4
CONC 3	100	7.0	1.14	0.3750	300.4
CONC 4	300	21.0	1.51	0.3750	1,187.6
CONC 5	150	10.5	1.90	0.3750	746.6
CONC 6	100	7.0	2.24	0.3750	587.0
CONC 7	200	14.0	2.62	0.3750	1,377.6
CONC 8	50	3.5	2.99	0.3750	392.2
CONC 9	100	7.0	3.40	0.3750	892.6
Sum	1350	94.5			6,143.0

Table 7.2 Value of lost traffic for concentrator 10 - 17

Concentrator	Number of subscribers	Traffic (Amax, E)	Down time (hours/year)	Traffic loss factor F	Value of lost traffic (C_t, CU/year)
CONC 10	300	21.0	0.44	0.3750	346.6
CONC 11	50	3.5	0.83	0.3750	108.6
CONC 12	200	14.0	1.14	0.3750	600.6
CONC 13	100	7.0	1.51	0.3750	395.8
CONC 14	200	14.0	1.90	0.3750	995.4
CONC 15	50	3.5	2.24	0.3750	293.4
CONC 16	150	10.5	2.60	0.3750	1,023.8
CONC 17	100	7.0	3.01	0.3750	790.6
Sum	1150	80.5			4,554.8

The present value of lost traffic are calculated using an economic time period of 10 years and a rate of interest of 7 % per year. This gives a present value factor equal to $C_{pv} = 7.02$. The present value of lost traffic will be:
10,697.8 · 7,02 = 75,099 CU.

The total costs for alternative 0 will be:
600,000 + 75,099 = 675,099 CU.

Alternative 1

A 10 km long fibre cable is established between concentrator 9 and 17 in order to establish an ADM ring, as shown in Figure 7.4.

Diversity routing is used for the circuits between each concentrator and the local exchange. In practice, this means that one will have 2 circuits (2 Mbit/s) between each concentrator and the local exchange, routed in opposite directions in the ADM ring.

Totally, one will now have 34 circuits (2 Mbit/s) into the local exchange. It will still be sufficient with an STM-1 system in the local network.

The number of DTMs will be:
2 · 34 = 68.

Equipment and cable needed:

- 18 ADM-1s
- 68 DTMs
- 10 km fibre cable.

This represents a cost of about 1,000,000 CU.

In this alternative one will have a complete loss of traffic when a fault occurs in the ADM connected to the local exchange, and a complete loss of traffic to/from each individual concentrator when a fault occurs in the ADM that the concentrator is connected to. This means that one will have a relative circuit capacity $k = 0$ for all the traffic in the down time of 2 ADMs. Total traffic in the busy hour (*A*max) can be found from Table 7.1 and 7.2 and will be:
94.5 + 80.5 = 175 E.

For faults on the cable and the other ADMs (16), one will get a relative circuit capacity $k = 1$. Assuming that congestion dimensioning is used, giving a utilization of $U = 0.74$, one gets for combined traffic a traffic loss factor $F = 0$, which means that no traffic will be lost in these situations. (One does here neglect the possibility of having two simultaneous faults.)

For a relative circuit capacity $k = 0$, one will have a traffic loss factor $F = 0.375$ (combined traffic).

For alternative 1 the value of lost traffic will be:

$$C_t = 175 \cdot 2 \cdot 0.1 \cdot 0.375 \cdot 100$$
$$= 1{,}313 \text{ CU/year.}$$

In present value the value of the lost traffic for alternative 1 will be:
1,313 · 7.02 = 9,215 CU.

Table 7.3 Summary results of the calculations

Alternative	Equipment and cable needed	Equipment costs (CU)	Value of lost traffic (present value, CU)	Total costs (CU)	"Profitability" relative to alt. 0 (CU) 1)
0	18 ADM-1 51 DTM	600,000	75,099	675,099	-
1	18 ADM 68 DTM 10 km cable	1,000,000	9,215	1,009,215	-334,116
2	18 ADM 51 DTM 10 km cable SNCP in 18 ADM	900,000	9,215	909,215	-234,116

1) Negative "profitability" means that the alternative is less "profitable" than alternative 0.

Total costs for alternative 1:
1,000,000 + 9,215 = 1,009,215 CU.

Alternative 2

A 10 km fibre cable is established between concentrator 9 and 17 in order to establish an ADM ring, as shown in Figure 7.4.

One has 100 % back-up for the circuits between the concentrators and the local exchange. The back-up circuit is routed in the opposite direction of the circuit it is back-up for. SNCP is used to establish the back-up circuits.

Totally, one will now have 34 circuits (2 Mbit/s) into the ADM that is connected to the local exchange. The number of circuits into the local exchange will, however, be the same as for alternative 0, i.e. 17 circuits (2 Mbit/s). It will still be sufficient with an STM-1 system in the local network.

The number of DTMs will be:
3 · 17 = 51.

Equipment and cable needed:

- 18 ADM-1s
- 51 DTMs
- SNCP in 18 ADMs
- 10 km fibre cable.

This represents a cost of about 900,000 CU.

In this alternative one will get exactly the same conditions as for alternative 1: One will have a complete loss of traffic when a fault occurs in the ADM connected to the local exchange, and a complete loss of traffic to/from each individual concentrator when a fault occurs in the ADM that the concentrator is connected to. This means that one will have a relative circuit capacity $k = 0$ for all the traffic in the down time of 2 ADMs. Total traffic in the busy hour (*A*max) can be found from Tables 7.1 and 7.2 and will be:
94.5 + 80.5 = 175 E.

For faults on the cable and the other ADMs (16), one will get a relative circuit capacity $k = 1$. Assuming that congestion dimensioning is used, giving a utilization of about $U = 0.74$, one gets for combined traffic a traffic loss factor $F = 0$, which means that no traffic will be lost in these situations. (One does here neglect the possibility of having two simultaneous faults.)

For alternative 2 the value of the lost traffic will be the same as for alternative 1, i.e.

1,313 CU/year and in present value: 9,215 CU.

The total costs for alternative 2:
900,000 + 9,215 = 909,215 CU.

Conclusions

The results of the calculations above are summed up in Table 7.3.

Table 7.3 shows that alternative 0 is the most “profitable” alternative. The reason for this is that one in alternative 1 and 2 must establish 10 km fibre cable between concentrator 9 and 17. This is equivalent to a cost of about 300,000 CU.

If one disregards the costs for establishing the cable, alternative 2 (100 % backup), will be the most “profitable”. Because of the increase in the number of DTMs, alternative 1 will still be less “profitable” than alternative 0.

The conclusion is that the traffic in this local network is too low to justify the establishment of an ADM ring for dependability reasons. The value of saved traffic must be increased by approximately a factor of 6 in order to justify the investment of establishing a ring.

8 Summary

The dependability project was closed 1 September 1996. Each milestone in the project was finalised with a status report. In addition, the major results and recommendations from the project are given in the status report of milestone M7: “Summary report of the project with conclusions”. All the status reports were submitted to the Planning Department of Telenor Network (project owner), at the completion of the project. Below are summarised the major results and recommendations from the project.

Based on the economical method described in ITU-T rec. E.862, the dependability project has evaluated existing planning rules for the transport network and the telephone network/ISDN in the context of availability. The method is considered to be a good tool to determine the level of availability in different parts of the network. During variation of the value of the lost traffic dependent on the customers and the grade of competition, an adaptation of the level of availability is possible.

For the transport network, existing planning rules are laid down in the report from another project “Superior plan for SDH development co-ordinated with the replacement of PDH”. The dependability project concludes that the guidelines given in this report ensures good availability in the transport network. Additional guidelines concerning the use of spare capacity and recommendations related to protection/restoration, are given for

- The inter-regional network
- Regional networks.

For local networks the conclusion is that the establishment of SDH-rings will have to be considered in each case. Simple guidelines for using the economical method to decide whether an SDH-ring should be established or not, from an availability point of view, has been evolved.

In the telephone network/ISDN, the existing planning rules are given in “The target network for the telephone network/ISDN”. The dependability project concludes that the guidelines given here ensures good availability, but that they should be supplemented by availability actions in the transport network. Some of the guidelines should be modified and changed and the dependability project also proposes supplementary guidelines.

In addition to this, the dependability project points to several problem areas revealed during the work of the project and especially asks for further study of these problem areas within the organisation of Telenor Network.

The dependability project has shown that the statistics concerning the number of faults, repair time and thus the availability of different types of equipment, is important in order to be able to make accurate calculations for the availability in parts of the network and for different services. It is strongly recommended that this area should be studied further in order to come up with routines and methods to achieve a precise knowledge of the availability of different equipment and services in connection with the introduction of new systems for fault management in Telenor Network.

Although the dependability project did not achieve all the goals laid down in the mandate, the work represents a major thrust forward in the area of availability planning and awareness about dependability problems in Telenor. It is hoped that the project will be followed up, to give the day to day planner in Telenor the necessary tools to make availability evaluations and in the end, offer the customers the availability they want and at a price the customers are willing to pay.

The conclusions and recommendations from the project are now under consideration by the head of the Network Operations Planning Department in Telenor Network, who is responsible for initiation of further actions and follow-up.

A traffic model for calculating the effect of reduced circuit capacity on the availability of the ISDN/telephone network

BY ARNE ØSTLIE

1 Introduction

When is a network or a service available from a customer's point of view? We have no direct answer to this important question. The ITU definitions consider the availability of items which may be regarded as building blocks of network elements like exchanges and transmission lines. According to the ITU definitions *an item is available if it is in the state to perform a required function*. This definition is OK as long as the item is either working or not working. However, very often a faulty network element will result in reduction of the circuit capacity. Depending upon the traffic offered the effect of the fault on the traffic flow may vary from no influence to very high congestion. So, what is required function for the telephone service? This is a very good question. The answer is not obvious and we will discuss this later in the article (Chapter 3).

In a dependability project in Telenor we have approached this problem from another angle as described in the ITU-T Rec. E.862 "Dependability planning of telecommunications networks" (see the article "Dependability in telecommunication networks"). In Rec. E.862 the traffic lost due to failures and the value of the lost traffic are used as parameters. When comparing different network structures, the present values of lost traffic are compared with the investment costs. But how should the amount of lost traffic be calculated? Is it the traffic during busy hour multiplied by the down time? This approach is often used, but it has certainly several limitations. In the ITU-T Rec. E.862 model a proportion of the traffic offered during the failure may be lost. But there is no guideline to determine what proportion is lost.

The target network for Telenor's telephone network is indicated in Figure 1.

The example shows how a subscriber is connected to the long distance network through a concentrator *(C)*, to a local exchange *(LE)* which again is connected to two transit exchanges *(TE)* which describes a TE-region. The 14 TE-exchanges are connected by a mesh network.

The network elements may be divided into two types depending on the consequence of a failure:

1. A failure on one element affects all the traffic between the subscribers *A* and *B*. There will be no traffic capacity left and all traffic will be lost. Failures on network elements like concentrators, subscriber exchanges and the access network will in most cases have this effect. The traffic lost may be calculated as the traffic offered (not including repeated call attempts) during the fault multiplied by the down time. If we assume a uniform failure distribution over all the hours of the year, the traffic lost per year (Erlang-hours) due to failures may be calculated as the average traffic intensity (Erlang) multiplied by the down time per year (hours).
2. A failure on a transit exchange or a transmission system in the long distance or regional network normally will reduce the circuit capacity by a circuit capacity factor, e.g. to 50 % or 67 %. This is a result of the target network with double or triple homing. The effect of the reduced circuit capacity during a fault depends upon the traffic offered and indirectly also upon the dimensioning. If the failure occurs during low traffic hours, we may expect that no traffic is affected. However, during high traffic periods, the traffic loss cannot be neglected. Intuitively, we accept that the expected lost traffic depends upon factors like

- The profile of the traffic
- The down time distribution, in this paper we will assume a uniform down time distribution, that is the probability of a fault is constant and not time dependent
- The circuit capacity during the fault.

Our objective is to establish a traffic model combining type 1 and type 2 above. This means that we want to calculate the expected traffic loss depending upon the dimensioning, the circuit capacity factor and the traffic profile. We assume that type 1 is a special case of type 2. In other words, we want to calculate how much traffic is lost, e.g. on a group of circuits due to failures on exchanges and transmission systems given the data above.

This is not an easy task, however, the use of such a model is very promising. We see the possibilities to compare different alternatives of protection mechanisms or combinations of protection mechanism by calculating the present value of saved traffic and the investment of each alternative. We may compare e.g. double homing in the network, protection in the transmission network and a combination of double homing and protection. We may also vary the dimensioning parameters and the percentage of the transmission lines protected. Another question that may be addressed is the use of diversity routing. Or generally: Which protection mechanisms should be combined where in the network? However, before jumping to conclusions, the accuracy of input data, e.g. availability figures and the value of lost traffic, have to be considered.

At the end of the paper we will return to the problem of calculating an end-to-end availability and discuss different approaches to this problem.

2 Traffic model

2.1 Traffic loss during a fault

Our first problem is to estimate the traffic loss given

- The traffic demand, *A* (= traffic offered without repeated call attempts)
- The number of circuits, *N'*, carrying traffic during the failure.

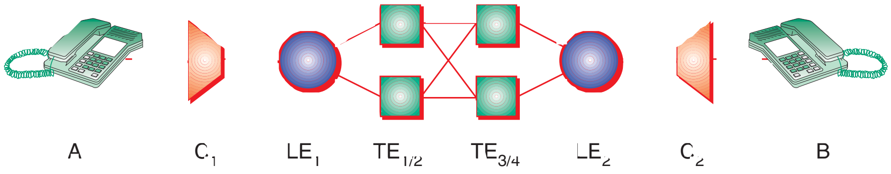


Figure 1 Example of the Telenor target network structure for telephone/ISDN

Figure 2 shows the traffic loss during congestion comparing the following models:

1. Erlang's loss model (no repetition of calls). Normal loss formula $E = E(N,A)$
2. Erlang's delay model. Implies that lost traffic = A - N for A - $N > 0$ and 0 if A - $N < 0$. The Erlang delay model may be interpreted in the following way. The subscribers have infinite patience during the failure period. This means that they are not giving in, but are repeating the call until they eventually succeed. The effect of this is that if the traffic demand is higher than the number of circuits, all working circuits will be filled with traffic. Traffic demand higher than the number of circuits will be lost.
3. Heuristic formula as a "ratio" of the two above. Lost traffic = A - 0.97 * N.

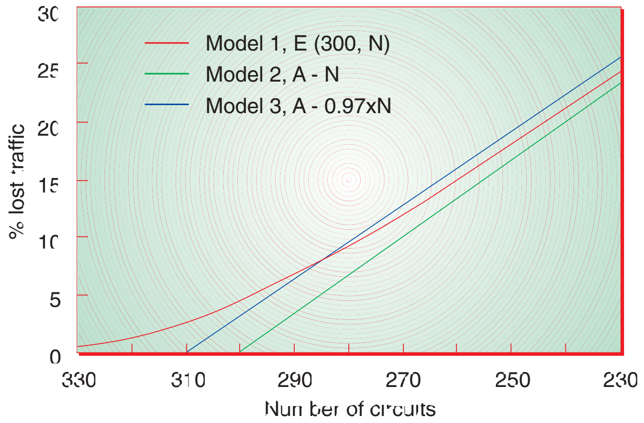


Figure 2 Comparison of traffic loss for different traffic models. Traffic A = 300 Erlang

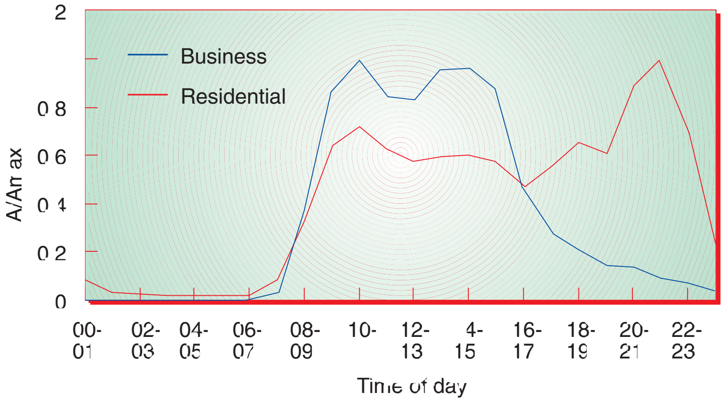


Figure 3 Typical traffic profiles for a working day

We expect that the traffic lost should be between the two Erlang models. However, for short fault duration, some of the traffic will be successful after the failure is restored. In this case, all three models will overestimate the traffic loss. Generally, traffic research has shown that the traffic lost may vary from about 30 % for short fault duration to 100 % for long. We do not have recent measurements of subscribers' persistence. However, we assume that today's subscribers are more accustomed to a high grade of service, and also that the calls are more a result of the moment's wish and have less value if they cannot be set up immediately. If we compare with the situation 15 – 20 years ago, when most of the research on repeated call attempts were done, we will expect today that a lower proportion of the traffic will be repeated after normalisation. The customer expects the call to be set up at the time he tries. Even if he manages to get through some time later, he will probably be annoyed. We have therefore made the simplifying assumption that all the traffic is lost independent of the fault time.

In the dependability project we have chosen the heuristic model. However, the results are not very dependent upon the choice of model.

2.2 Traffic profile

In Figure 3 are shown two traffic profiles for typical residential and business exchanges for working days. The figure shows that residential customers have a 2 hour traffic peak period in the evening, while business customers have a much longer busy period during the day. In Norway, this is normally reflected in similar profiles also for circuit groups.

We assume that we have traffic measurements for all the 8,760 hours of a year for two circuit groups carrying traffic with a business or a residential profile. For each group we sort the measurements in declining order. In Figure 4 are shown two traffic profiles for a year based on 16 days' measurements. The figure then tells the proportion of hours in a year with higher traffic than A. The traffic profiles are sequentially linearized. This is a simplification, but not far from the truth.

2.3 Expected traffic loss

In Figure 5 the same traffic profiles are shown. In the figure a line indicating the number of intact circuits N' in a failure situation is shown. According to the Erlang delay model, the traffic above the line would be lost. If we want to use model 3 instead, we only replace N' with 0.97 * N'. The figure tells us how much traffic will be carried and how much traffic will be lost in a failure situation with N' circuits.

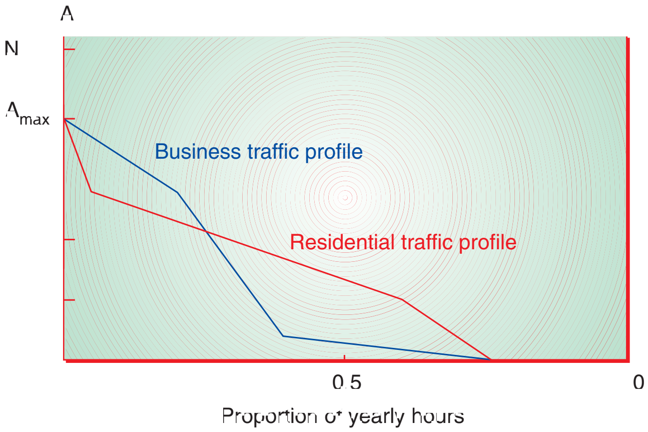


Figure 4 Traffic profiles for business and residential subscribers

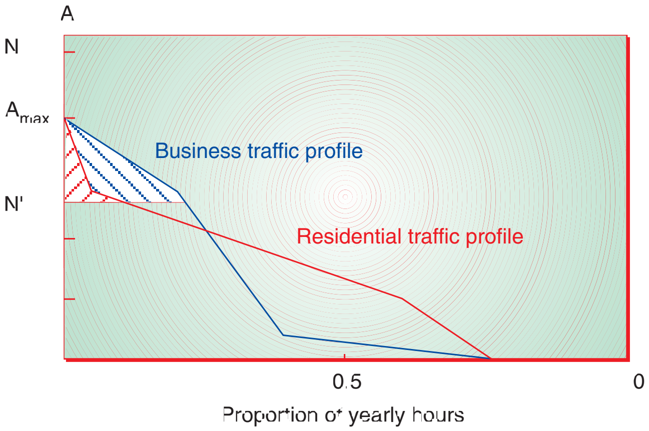


Figure 5 Traffic lost with N' circuits intact during the failure situation

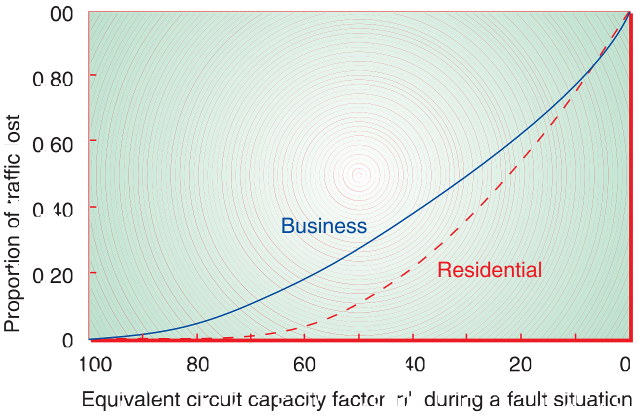


Figure 6 Proportion of traffic lost as a function of equivalent circuit capacity n' during a fault situation (%)

If we assume that the down time is uniformly distributed, the failure intensity will be the same for all hours of the year. The traffic volume defined by the hatched part of the diagram can be interpreted as the traffic lost. The total traffic volume in all the hours of the year is represented by the total area under the curve. Therefore, the proportion of lost traffic assuming uniformly distributed down time is

$$P_L = \frac{Hatched\ \ area}{Total\ \ area}$$

From Figures 4 and 5 we see that business traffic has many more hours of high traffic than residential traffic. In the example, the lost traffic will be much higher for business traffic than for residential traffic, which mainly has a short traffic peak during the evening.

The linearization of the curve makes it simpler to calculate the areas under the curves and make tables for different capacities relative to the maximum traffic. Curves comparing business and residential profiles are given in Figure 6. It shows how much traffic is expected to be lost relative to the offered traffic, assuming the down time probability is the same for all hours. The input parameter is what we will call equivalent circuit capacity

$$n' = 0.97 * N' / A_{MAX} \qquad (1)$$

where model 3 for traffic loss is assumed. A consequence of the model is that the proportion of traffic lost is not dependent on the circuit group size directly, only the ratio N'/A_{MAX}.

Up till now we have used A_{MAX} and N' as parameters. In a practical situation we would like to use

- Circuit capacity factor due to failure, $n = N' / N$. Different fault protection methods will have different circuit capacity factors, e.g. 0.5 for double homing or diversity routing, 0.67 for triple homing.
- Circuit utilisation $U = A_D / N$. This is a parameter that is characteristic for each circuit group as well as for the network. It depends upon the dimensioning principle used and the practice for augmentation of circuit groups. This is generally more important than the circuit group size. The utilisation is therefore a very convenient parameter to use.

In Telenor, we estimate that the maximum traffic, A_{MAX}, is 1.05 higher than the dimensioning busy hour traffic, A_D. Now we are able to calculate the equivalent circuit capacity *n'* from our preferred parameters as follows:

$$\begin{aligned} n' &= 0.97 * N' / A_{MAX} \\ &= 0.97 * (N' / N) * (N / A_D) \\ &\quad * (A_D / A_{MAX}) \\ &= 0.97 * n / (U * 1.05) \end{aligned}$$

$$n' = n / (1.08 * U) \qquad (2)$$

Example

A circuit capacity factor $n = 0.5$ (e.g. diversity routing) and circuit utilisation $U = 0.70$ gives $n' = 0.66$. Tables similar to the curve in Figure 6 give a traffic loss for the business traffic profile of 13.3 % and for the residential traffic profile 2.2 %. It must be remembered that these are average figures assuming that the down time is uniformly distributed.

According to Figure 5 the maximum lost traffic for both traffic profiles would be

$$(1 - n') = 1 - 0.66 = 34 \%.$$

We may compare with a circuit group of 300 circuits. During a fault the circuit group is reduced to $N' = 150$. $A_{MAX} = 0.70 * 300 * 1.05 = 220$ Erlang. If we use Erlang's loss formula we find a congestion *B* without repeated call attempts = 33 %. Due to repeated call attempts the congestion should be expected to be even higher.

If we from Figure 5 calculate the proportion of hours when there will be a loss of traffic, we will find about 31 % for the business profile and about 3 % for the residential profile.

2.4 Typical traffic loss figures

In Table 1 typical traffic loss factors with circuit capacity factor 0.5 and different values of circuit utilisation are calculated.

It is very important to stress that the figures in column 2 and 3 of Table 1 gives *average* loss in a failure situation with 50 % circuit capacity assuming that the down time is uniformly distributed over all the 8,760 hours in a year. The traffic loss during the busiest hour is much higher, as we can see from column 4.

From Table 1 we see that the expected traffic loss is much smaller for the residential than for the business profile. The reason is the short traffic peak in the evening for residential subscribers. If a failure occurs during other hours there will be capacity to carry the traffic.

Table 1 Expected and maximum traffic loss with uniformly distributed down time and 50 % circuit capacity factor

	Expected traffic loss (%)		Maximum traffic loss (%)
Utilisation (A_D/N)	Residential Profile	Business Profile	During busiest hour
0,6	1	6	23
0,7	2	13	34
0,8	5	20	42

3 End-to-end availability

As long as the duration of a fault is low and the faults are independent of each other, we can with very good accuracy neglect the effect of two or more simultaneous faults. With unavailability times of up to 10 hours a year, the probability of two simultaneous faults will be of the order 1000 less than the probability of one fault. However, we have to be aware of incidents where faults will not be independent, e.g. due to weather conditions.

This approximation makes it possible to calculate the end-to-end unavailability as the sum of the unavailabilities caused by faults on each network element. Each network element *i* has an unavailability figure in hours T_i per year and a corresponding unavailability reduction factor $P_{R,i}$. (The unavailability reduction factor is dependent upon how the element, e.g. transit exchange, is used.) The fault effect of one element may be defined as the product of the unavailability and the unavailability reduction factor. The resulting weighted unavailability T_w we get by adding the fault effects of each fault:

$$T_w = \Sigma (T_i * P_{R,i}) \qquad (3)$$

However, we have to determine the unavailability reduction factor. We are back to the questions in the introduction to this paper: When is the network available? What is the required function of the telephone service? In the following, we suggest three alternatives to define the unavailability reduction factor:

1. The telephone service is available if the grade of service is better than a specified percentage.
2. The telephone service is available if the traffic loss is less than a specified percentage. This definition is near to method 1, however, subscriber persistence giving repeated call attempts are not taken into account. It is only the

Table 2 Unavailability reduction factor (%) comparing method 2 with three different levels of traffic loss with alternative 3. Three levels of utilisation, U, during normal situation with dimensioning busy hour traffic. 50 % circuit capacity during the fault in all examples

	Residential profile				Business profile			
U	2/0 %	2/10 %	2/20 %	3	2/0 %	2/10 %	2/20 %	3
60 %	2	2	1	1	13	8	3	6
70 %	7	3	2	2	20	15	10	13
80 %	17	10	3	5	32	19	15	20

capacity of the network to carry the traffic normally demanded by the subscribers that matters. Both in alternative 1 and 2 the unavailability reduction factor is defined as the proportion of time that the specified percentage is exceeded.

3. A third approach is to define the unavailability reduction factor as the traffic loss factor. This means that the unavailability reduction factor is defined as the proportion of traffic lost during faults. This method is not connected to any service level defining availability. Instead, the down time is reduced according to the proportion of traffic which is lost. This is a reasonable point of view.

Method 1 is suitable for measurements on live traffic, but it is not suitable in a model to calculate the unavailability reduction factor. This is due to the influence of subscriber persistence during periods with low grade of service. However, with a traffic loss of 10 to 20 %, the traffic loss from method 2 and the rate of unsuccessful calls from method 1 are not far from each other. Due to the simplicity and suitability of the traffic model described earlier, we will prefer method 2 to method 1.

In Table 2 methods 2 and 3 are compared. For method 2 three levels of traffic loss are considered: 0 %, 10 % and 20 %. For the case with 50 % circuit capacity during faults the results are calculated from Figure 5 by determining the percentage of hours that will have more traffic loss than the level considered. The figures for the fault effect for method 3 are the same as already found as traffic loss figures in Table 1.

Which level of traffic loss should be used to differ between available and unavailable service? Due to new services and competition, the acceptable loss is certainly lower than it was some years ago. If the traffic loss is 20 %, the repeated call attempts cannot be neglected and the congestion gets higher. We may argue that the service is not acceptable and we may define it as not available. However, if the traffic loss is 10 %, the repeated call attempts are expected to be low. We are likely to define the service as available. In our example we find that the unavailability reduction factor from method 3 (the proportion of traffic lost) is between method 2 with traffic loss 10 and 20 % as service levels. This means that with our profiles, method 3 fits reasonably well with method 2. As we already use the traffic loss as a parameter in the calculation of lost traffic, we will recommend it to be used for end-to-end calculation in formula (3).

In the Telenor network the same dimensioning principles are used together with load sharing and double homing. Calculation of end-to-end availability is in our model then reduced to

- Determine which elements are involved in the set up or transfer of the traffic
- Find the down time per year for these elements
- Find the traffic loss factors when an element fails.

A connection between subscriber A and B in Figure 1 may then consist of the following elements (standard values for down time are used for illustration of the method):

- Two access net parts; down time = 4 hours/year, traffic loss factor = 1.0
- Two end offices; down time = 0.5 hours/year, traffic loss factor = 1.0
- 4 transit exchanges; down time = 0.5 hours/year, traffic loss factor = 0.1
- 200 km transmission, down time = 4 hours/year, traffic loss factor = 0.1
- 1000 km transmission, SNCP, down time = 20 hours/year, traffic loss factor = 0

$$\begin{aligned} T_w &= 2 * 4 * 1 + 2 * 0.5 * 1 + 4 * 0.5 \\ &\quad * 0.1 + 4 * 0.1 + 20 * 0 \\ &= 8 + 1 + 0.2 + 0{,}4 + 0 \end{aligned}$$

$$T_w = 9.6 \text{ hours/year}$$

4 Example with diversity routing

We want to compare two alternatives of diversity routing with direct routing. The direct route is used as route 1 in both cases of diversity routing. Route 1 has 1 hour down time per year. In alternative 2 we assume that the two parallel routes have the same down time, 1 hour/year. In alternative 3 we assume that route 2 has a down time of 5 hours/year. We assume a traffic loss factor 0.2 for diversity routing.

- Alternative 1. No diversity. Down time 1 hour/year. 100 % loss.
  Weighted unavailability:
  1 * 1.0 = 1 hour/year
- Alternative 2. Diversity routing. Down time route 1 is 1 hour/year.
  Down time route 2 is 1 hour per year.
  Same circuit capacity on both routes.
  Traffic loss factor 0.2.
  Weighted unavailability:
  1 * 0.2 + 1 * 0.2 = 0.4 hours/year
- Alternative 3. Diversity routing. Down time route 1 is 1 hour/year.
  Down time route 2 is 5 hours per year.
  Same circuit capacity on both routes.
  Traffic loss factor 0.2.
  Weighted unavailability:
  1 * 0.2 + 5 * 0.2 = 1.2 hours/year

We see that alternative 3 gives higher weighted unavailability than alternative 1, showing that diversity routing is not always an advantage according to this model.

5 Conclusions

In this article is shown a traffic model to calculate the traffic lost in failure situations. For a traffic profile assuming uniformly distributed down time, the proportion of traffic lost can be calculated when we know the circuit utilisation during dimensioning traffic and the circuit reduction factor during the fault. This is used as input to the ITU-T Rec. E.862 for calculation of the value of lost traffic. Examples are shown in the article “Dependability in telecommunication networks”. The models make it possible to compare methods like fault tolerant dimensioning in the telephone network with diversity routing and protection switching in the transport network. We can also calculate the effect of combination of methods.

The traffic between two subscribers may be served by a number of network elements. We assume that the failures on each network element are independent. In our model each network element has an unavailability reduction factor which is chosen as the traffic loss factor. The resulting weighted unavailability is calculated by adding up the products for each element of down time per year and the corresponding unavailability reduction factor.

QoS differentiation in ATM networks – a case study[1]

BY BJARNE E. HELVIK AND NORVALD STOL

Asynchronous Transfer Mode (ATM) is an enabling technology for enabling differentiation of the QoS provided to various customers with respect to availability and reliability (servability). Hence, network operators may gain a competitive edge by providing "tailor-made" bearer services with respect to QoS and corresponding cost. This article introduces the concept of QoS differentiation and contains the results from an initial case study of using QoS differentiation in ATM. A hypothetical future Norwegian ATM network is used as a basis.[2] The basic strategy used is to pre-allocate one back-up path for each end-to-end VP through the network. The back-up path may share nodes but no links with the end-to-end VP. When a link fails, traffic is moved to alternative paths following a certain algorithm, e.g. low priority traffic may be removed to give room for higher priority traffic.

The main conclusion is that the approach seems feasible and may have potential for increased revenue for a service provider.

1 Background and motivation

Introduction of more service providers on the Norwegian as well as the international markets leads to an increased demand for competitive and "tailor-made" services for potentially very different types of telecommunication customers. Large public institutions and large businesses will typically demand (and be willing to pay for) a very high availability of offered services, e.g. an on-line ticket-ordering system or access to databases with critical information. Private citizens are usually at the other end of the "demand" scale, i.e. cheaper services with lower availability demands are preferred, e.g. to be used for accessing the Internet.

This differentiation in QoS requirements should be exploited when new services are planned and new networks are implemented. From a service providers point of view this may also be seen as an opportunity for increased income: spare capacity used to ensure high availability for demanding customers may be sold as a lower grade service (in terms of availability) when the network is failure free.

Asynchronous Transfer Mode (ATM) is the enabling technology for this service differentiation strategy. Pre-establishment of additional (back-up) Virtual Paths (VPs) through a network secures a very quick rerouting of high QoS traffic (typically 100 to 150 ms) without end-to-end disconnects. This may be done without actually having reserved any *unused* capacity. The capacity is instead used by lower QoS traffic. This lower QoS traffic is then "thrown out of the network" when the capacity is needed by the high QoS traffic.

Service integration is an inherent part of ATM, as is the possibility of allocating virtual resources (back-up VPs). In this initial study, it is assumed that a simple back-up path strategy based on these principles should have a relatively low implementation cost in an ATM network. Also the Operation and Management costs should be low. A summary of related work is given at the end of this section.

The paper presents some first results from examining these service provider opportunities. It introduces the basic theoretical framework for dealing with differentiated servability, with respect to availability, in ATM networks, see Section 2 and 3. Section 4 presents a hypothetical Norwegian ATM backbone network offering services with differentiated QoS. In this network, links between the nodes may fail, while switching nodes are assumed failure free. A static link disjoint back-up VP for each end-to-end VP through a network is used. Some results for this network, focusing on availability of end-to-end connections, are presented in Section 5, before some concluding remarks are given in Section 6. This work is obviously an element in a wider network planning and optimization context. This will not, however, be discussed.

The work reported here is strongly related to techniques for achieving self-healing ATM networks. Some studies of using pre-allocated back-up paths in ATM networks have been done by others, see e.g. [2], [9] and [7]. In [2] a number of back-up paths, not necessarily disjoint, may be pre-allocated for every end-to-end VP. A selection of the best one to use is then necessary when a failure occurs. In [9] each end-to-end VP is allocated one disjoint back-up path. More back-up paths may, however, be allowed for high-priority VPs. [7] discusses pre-allocation of back-up VPs to establish a self-healing ring network. The use of *flooding,* i.e. to search extensively for an alternative path *after* a failure has been detected has also been studied, see e.g. [4], [11] and [5]. In [4] and [6] a simulator is developed to study a specific flooding algorithm. [11] describes a single-search algorithm, i.e. the flooding algorithm searches for an alternative path from only one side of the failed link. In [5] a double-search algorithm is described.

2 Definitions and basic formalism

To facilitate a structured discussion and identification of further work, a semi-formal description of the problem is introduced. See also Figure 1 for illustration of some of the concepts.

2.1 Network topology

A general network is considered. The nodes are denoted $i, j, ...$, and the links between them $\{i, j\}$, $\{l, m\}$, ... The terminal nodes, which are assumed to be fault free, are denoted s, d. The link between a terminal node and the network may fail.

A network failure is defined as the set of elements which have lost their traffic carrying capability. For instance, a node m and a link $\{i, k\}$

$$f_x = \{m, \{i, k\}\}$$

For simplicity, denote the fault free state $f_0 = \{ \}$. A notational convention used is to let the index x denote decreasing probability of the corresponding failure, i.e. $x > y \Rightarrow P(f_x) \leq P(f_y)$. It is chosen to define network failures as done above to facilitate and encourage analysis of dependent failures of network elements. The capacity of link $\{i, j\}$ is denoted $C_{i,j}$. It is assumed that there is no capacity restrictions within the nodes/switches.

[1] *The work reported here is carried out under contract with Telenor R&D.*

[2] *This network is in no way based on development plans or data from Telenor.*

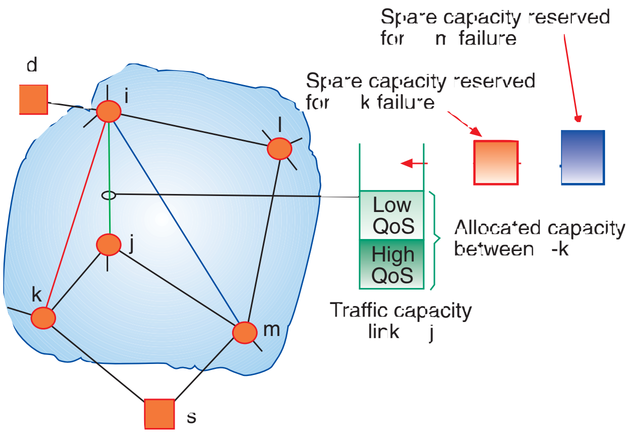


Figure 1 Illustration of the concept of using back-up VPs

2.2 Traffic offered

Traffic is offered to the network between its terminal nodes. For tractability, this traffic is quantified in terms of effective bandwidth, for instance as introduced by Kelly [10]

$$\alpha(z,t) = \frac{1}{zt}\log E\left(e^{zX(t)}\right) \qquad (1)$$

where *X(t)* is the traffic volume offered during a random interval of length *t*. It is assumed that the offered traffic in this measure is (may be treated as) additive.

We consider several QoS classes. Hence,

$$A^{(q)}_{s,d}(z,t)$$

is the offered traffic from *s* to *d* with QoS level *q*. QoS level zero, i.e. $q = 0$, is the best, and the QoS is strictly decreasing with increasing *q*. In the study presented in Section 4, only two levels will be used for the sake of simplicity.

2.3 Routing and rerouting

The objective of the routing is to map the offered traffic as described in Section 2.2 onto the topology defined in Section 2.1 when the network is fault free, i.e. f_0. The rerouting is the re-establishment of the paths through the network affected by a failure, say f_x. For modelling and analysis it is seen that:

- The rerouting after a failure may be different from the routing chosen if all paths were defined from scratch. In the latter case it is likely to achieve a better routing. This corresponds to a “repacking” of the network, which seems unrealistic for a network in operation.
- Under the assumption that the network is not “repacked”, it is seen that actual routing with failure f_x depends on the sequence of failures leading to it. At least for analytical studies this represents a major complication. Hence, with respect to rerouting we will assume that the network was fault free, f_0, before f_x.

The routing and rerouting algorithms will determine the traffic streams carried on all links when the network is fault free, or a certain failure exists. This is represented by a set of triplets, {*s, d, q*}, for each link, i.e. $L_{i,j}(x) = \{..., \{s, d, q\}, ...\}$. In this formulation, it is implicitly assumed that end-to-end streams which cannot be fully carried are completely lost. (This conservative assumption is not used in Section 4.) The traffic carried on link {*i, j*} under f_x becomes

$$a_{i,j}(x;z,t) = \sum_{\{s,d,q\}\in L_{i,j}(x)} A^{(q)}_{s,d}(z,t) \quad (2)$$

and where for a feasible routing it is required that $a_{i,j}(x;z,t) < C_{i,j}$.

2.4 Unavailability requirements

It is foreseen that the QoS guarantee given to the end-user is, in this context, in terms of an unavailability. Denote the unavailability of communication between s and d given that service level q is subscribed for $U^{(q)}_{s,d}$. It is then required that

$$U^{(q)}_{s,d} \le U^{(q)}, \forall\{s,d\} \qquad (3)$$

3 Theoretical model

Figure 1 indicates how we foresee an improved network utilization in the coming network.

- The network is dimensioned so that in a fault free operating state it may carry all the (predicted) traffic of all service classes with a nominal rejection rate, due to CAC, or better.
- The links of the network may also be dimensioned with a certain spare capacity to carry re-routed traffic from failed links and nodes. This spare capacity may be nil if the link nominally carries sufficient “low QoS traffic” that may be removed.

3.1 Network states considered

Identify the failure states in the network, f_x, and their corresponding probabilities, $P(f_x)$, in increasing order as defined in Section 2.1 The entire number of failure states in the network will be very large, 2^{n+m} = All, where *n* is the number of nodes and *m* is the number of links. To restrict the computational effort, the number of states considered should be restricted as much as possible, for instance up to $F_{\max}$, All ≥ $F_{\max} \ge F^{(q)}, \forall q$, where $F^{(q)}$ is implicitly defined in (4) as the index of the least likely failure state where traffic from all service levels may be handled.

For each of the network states determine the (re)routing. This should be done under the routing strategy considered, e.g. fixed pre-defined back-up or flooding.

3.2 Network performance

The unavailability requirement of Section 2.4 may be reformulated in the following terms. The probability of the network failures where QoS level *q* cannot be handled, should not be greater than $U^{(q)}$. Formally, this may be written as

$$1-\sum_{x=0}^{F^{(q)}} P(f_x) \le U^{(q)} \tag{4}$$

where a satisfactory handling of QoS level q implies

$$\sum_{\{s,d,u\}\in L_{i,j}(x)} A_{s,d}^{(u)}(z,t) < C_{i,j},$$

$$\forall(u \le q),\ \forall\left(x \le F^{(q)}\right),\ \forall\{i,j\} \tag{5}$$

Lost traffic (and corresponding revenue) due to failures is bounded by [17]

$$\hat{B}^{(q)}_{\text{Upper}} = \sum_{x=0}^{F_{\max}} P(f_x)\cdot B_x^{(q)} + \left(1-\sum_{x=0}^{F_{\max}} P(f_x)\right)\cdot B_{All}^{(q)}$$

$$\hat{B}^{(q)}_{\text{Lower}} = \sum_{x=0}^{F_{\max}} P(f_x)\cdot B_x^{(q)} \tag{6}$$

where $B_x{}^{(q)}$ is the traffic of service quality level q which is not carried in failure state f_x, i.e. assuming that all paths are rejected at the border of the network. $B_{All}{}^{(q)}$ is the traffic lost when all network elements have failed and are obviously equal to the offered traffic. More precisely, $B_x{}^{(q)}$ is defined in (7), where the operator $\forall j$ accounts for dual (multiple) homing of the source nodes, e.g. terminal node s in Figure 1.

$$B_x^{(q)}(z,t) = \sum_{s}\ \sum_{\{s,d,q\}\in L_{s,j}(x),\forall_j} A_{s,d}^{(q)}(z,t) \tag{7}$$

The expressions above are also valid if lost traffic (or unavailability) for *a certain source-destination pair* is focused: indexes s,d must then be added to all lost traffic variables (B). Note also that the resulting upper unavailability value for a given source-destination pair, based on (6), is found by adding the term

$$1-\sum_{x=0}^{F_{\max}} P(f_x) \tag{8}$$

to the lower unavailability value. This term will be equal for all source-destination pairs.

For a given network topology and dimensioning, the routing strategy that meets the dependability requirement (4) and (5) and minimizes the lost traffic (6) and (7) is the best.

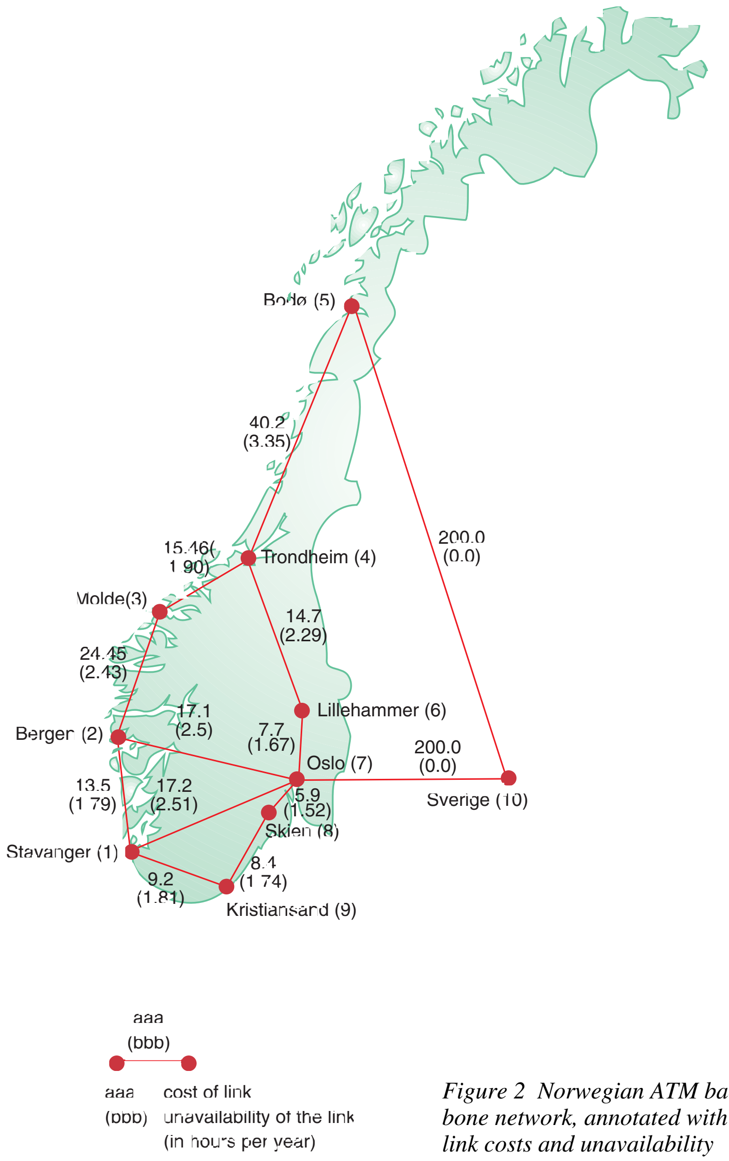


Figure 2 Norwegian ATM backbone network, annotated with link costs and unavailability

4 Hypothetical Norwegian ATM backbone network

4.1 Network structure and parameters

This section contains a description of a hypothetical Norwegian ATM backbone network at a certain level of development. From a "structural" point of view it probably represents a nearly finished network since most geographical areas are included.[3] From a "traffical" point of view it represents some early point (1 – 2 years?) after the introduction of the network; i.e. traffic will probably increase to many times the amounts used below.

However, the actual traffic level is not the important issue in this study. The important task is to show (in principle) what is possible to accomplish, regarding QoS differentiation, even with very simple back-up path strategies.

[3] *A possible addition would be Tromsø.*

Table 1 Reference traffic matrix (sum of traffic in both directions)

	Bergen	Molde	Trondheim	Bodø	Lillehammer	Oslo	Skien	Kristiansand
Stavanger	101.1	47.4	75.0	92.7	73.6	259.9	70.8	47.0
Bergen		74.0	116.7	144.9	114.9	405.9	110.8	73.4
Molde			54.7	67.9	53.8	190.19	51.8	34.4
Trondheim				107.09	84.9	300.0	81.7	54.2
Bodø					105.4	372.2	101.4	67.3
Lillehammer						295.2	80.4	53.4
Oslo							284.0	188.6
Skien								52.1

For the same reason, and to avoid obscuring the results with too much data, we focus on two QoS levels in this study. These two levels will be referred to as H-QoS traffic and L-QoS traffic, respectively: i.e. H-QoS traffic is high priority traffic to get the highest QoS, and L-QoS traffic is to get a lower QoS. The results should in principle be extendible to more levels.[4]

The ATM backbone network used as a starting point for our case study is taken from [3]. All the Swedish nodes have been replace by one common node, to be used by back-up paths only. The actual geographical placement of each node is not important in this study; the number of nodes and connectivity of these nodes is more important. Principal results should be representative even with different structures.

To establish primary (least cost) and secondary (back-up) paths through the network, a *cost of each link* must be defined. Only the relative relationships between these costs are used when defining the paths, not the actual costs.

Our initial procedure for defining these costs are simple: we assume that the cost is proportional to the length of the link. However, we multiply some of the links with 1.5 to account for the topography. Links via "Sverige" are given a very high cost since these should be used as a last resort back-up only. The resulting costs on the links are included in Figure 2.

Even though the method does not require the independent link failure assumption, as stated in Section 3, this assumption is made here for simplicity. The *unavailability of each link* is assumed to consist of two contributions:

- One contribution independent of the length of a link, i.e. reflecting the dependability of the link's end points and a part of operational errors, and
- One contribution proportional to the length of a link, i.e. reflecting the reliability of the physical medium and other length dependent events leading to failure.

The first contribution is assumed equal for all links, and is assumed to result in an average mean down time of 1 hour per year. The mean value of the part proportional to the link length is assumed to give a mean contribution to the down time in about the same numerical area. The resulting mean down time for each link is shown in Figure 2 (as the values inside parentheses). The links via "Sverige" are assumed failure free. As a result, the probability that the network is *without* any failure is 0.99732149, that *one and only one* failure is present is 0.00267527, and that there is *more than one* simultaneous failure in the network is $3.24\ 10^{-6}$. This value may also be added to the lower limits found for unavailability between two nodes in the network to get an upper limit for the unavailability.

4.2 Traffic matrix and network dimensioning

The traffic matrix for the ATM network is obtained as follows:

1. The relative relationships between traffic generated between any two switching nodes of the network is assumed proportional to the product of the population of the geographical areas around each node.
2. The traffic between Trondheim and Oslo is assumed to be 150 Mbit/s in each direction. These values are not necessarily mean values but rather *effective bandwidth*, see (1) in Section 2.2.
3. All other end-to-end values[5] are adjusted relative to Trondheim ↔ Oslo.

This leads to the total traffic values between each end-to-end point of the network as given in Table 1.

For dimensioning of the network 50 % of this traffic is assumed to be H-QoS. The basic dimensioning criteria is that so much capacity is allocated to each link that any single link failure can be handled without loss of H-QoS traffic with the back-up strategy presented in 4.3. Hence, the secondary (back-up) paths are able to handle any single link failure, giving zero unavailability for all H-QoS traffic connections in the network. However, if the

[4] *With some extensions of the resource allocation algorithms.*

[5] *Except "Sverige", which is only a transit node for back-up paths.*

amount of H-QoS traffic is larger than 50 %, loss may occur, even with the same total traffic matrix.

Figure 3 shows the resulting link capacities. An additional value is given (inside the parentheses) for each link, either with a leading "+" or a leading "–". A "+" indicates excess capacity, i.e. how much less a link could have been and still given zero unavailability for H-QoS traffic. A "–" indicates how much capacity has been added to the capacity needed to carry H-QoS and L-QoS traffic in a fault free network.

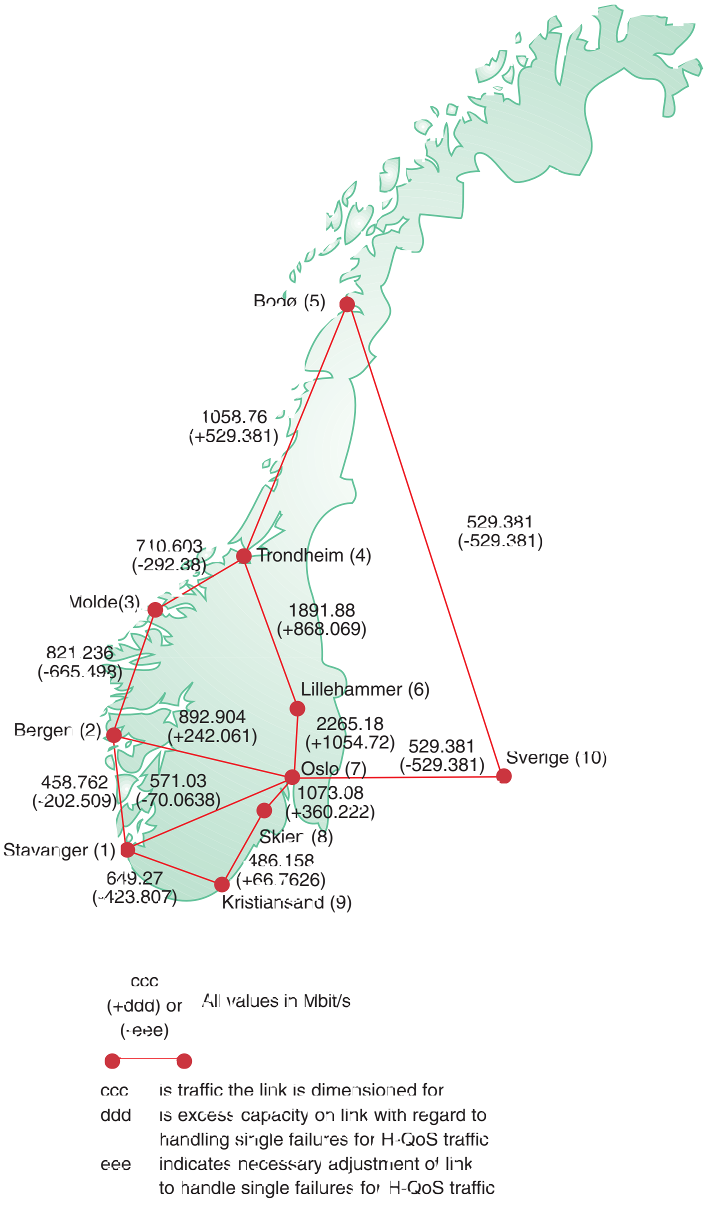


Figure 3 Link capacities (added capacities "–" or excess "+")

4.3 Back-up path strategy

The choice of link disjoint paths is partly motivated by [2]. The alternative path strategy used in this study is simple:

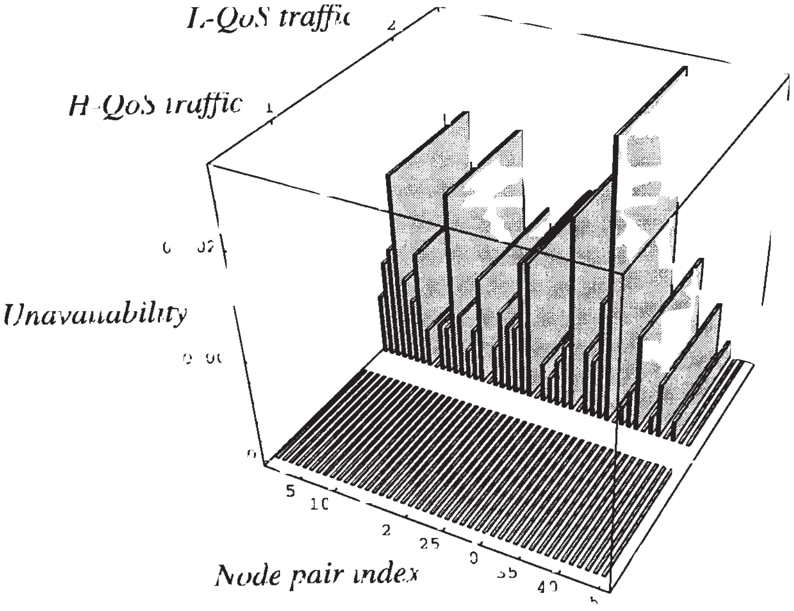
1. The primary path between each source destination pair, (s,d) in Figure 1, is the minimum cost path between these pairs.
2. Each primary end-to-end path through the network will have a single secondary (back-up) path, if such a path can be found in accordance with 3. and 4. below.
3. The secondary (back-up) path between any two given switching nodes (end-to-end) is totally disjoint with the primary path between the same two nodes, with regard to links. The primary and secondary paths may, however, traverse the same nodes.
4. The secondary (back-up) path is found as the minimum cost path with primary path links removed from the network.

If node failures contribute significantly to overall unavailability, both link and node disjoint back-up paths may be used. This would, however, result in an overall more expensive network.

When a single link in the network fails, the procedure below is followed:

1. All primary paths traversing the failed link is identified.
2. The secondary (back-up) paths for the affected connections are identified and sorted in an ascending order with regard to the number of links to be traversed; i.e. a shorter secondary path is prior to a longer one (in number of links to be traversed).
3. Affected H-QoS traffic is re-allocated for secondary paths in the sorted order above. If necessary, L-QoS traffic is disconnected to give room for H-QoS traffic. When disconnecting L-QoS traffic, connections using longer (in number of links traversed) primary paths are removed before shorter ones. Potential loss of H-QoS and L-QoS traffic is registered.
4. Affected L-QoS traffic is re-allocated for secondary paths in the sorted order above (in 2.). No disconnection of existing traffic is allowed. Potential loss of L-QoS traffic is registered.

Note that in the current version of the algorithm L-QoS traffic which is removed is not given any chance to be re-routed via any back-up path.



Node pair index:

	Bergen	Molde	Trond-heim	Bodø	Lille-hammer	Oslo	Skien	Kristian-sand	Sverige
Stavanger	1	2	3	4	5	6	7	8	9
Bergen		10	11	12	13	14	15	16	17
Molde			18	19	20	21	22	23	24
Trondheim				25	26	27	28	29	30
Bodø					31	32	33	34	35
Lillehammer						36	37	38	39
Oslo							40	41	42
Skien								43	44
Kristiansand									45

Figure 4 Unavailability with 50 % offered H-QoS traffic, i.e. the dimensioned proportion

5 Source destination unavailabilities

The unavailability for H-QoS and L-QoS traffic under the assumptions and dimensioning in Section 4 is shown in Figure 4. Upper unavailability limits may be found by adding 3.24 10^{-6} to the lower limits, see Section 4.1, and are therefore not included. This is also the case for the remaining figures.

It is now of interest to study what happens with the unavailability for both H-QoS and L-QoS traffic when we increase the relative percentage of H-QoS traffic above the level the network is dimensioned for, but keep the total traffic matrix constant, i.e. reduce the amount of L-QoS traffic correspondingly. Results for 55 % and 60 % H-QoS traffic are given below.

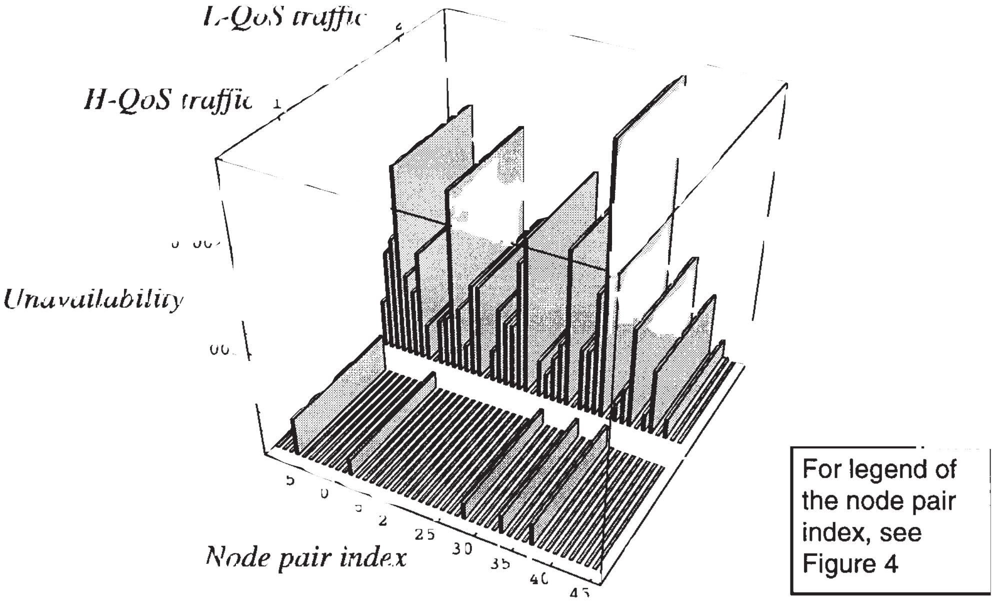
5.1 High QoS traffic constitutes five per cent points more than dimensioned

Figure 5 shows lower unavailability limits for end-to-end H-QoS and L-QoS traffic connections when H-QoS traffic constitutes 55 % of the total, cf. Table 1. For H-QoS traffic, we see that 6 of the 36 connections, e.g. between *Stavanger* and *Bodø,* get unavailabilities larger than zero, of which 5 are in the 1–3 hours per year area. The last one is very low.

For L-QoS traffic, the lower limits for unavailability are in the 1–17 hours per year area. No major differences can be observed in comparison with Figure 4. Most values are slightly higher, but the unavailability for some L-QoS traffic connections, e.g. between *Stavanger* and *Bergen,* are reduced. This is due to some L-QoS traffic being able to using link capacity freed by now unsuccessful H-QoS traffic connections.

5.2 High QoS traffic constitutes ten per cent points more than dimensioned

Figure 6 shows lower unavailability limits for end-to-end H-QoS and L-QoS traffic connections when H-QoS traffic constitutes 60 % of the total, cf. Table 1. The upper unavailability limits are found by adding 3.24 10^{-6} to the lower limits, as earlier, see Section 4.1. For H-QoS traffic, we now see that 10 of the 36 connections get unavailabilities larger than zero, of which 8 are in the 1–4 hours per year area. The other two are an order of magnitude less.



For legend of the node pair index, see Figure 4

Figure 5 Unavailability with 55 % offered H-QoS traffic

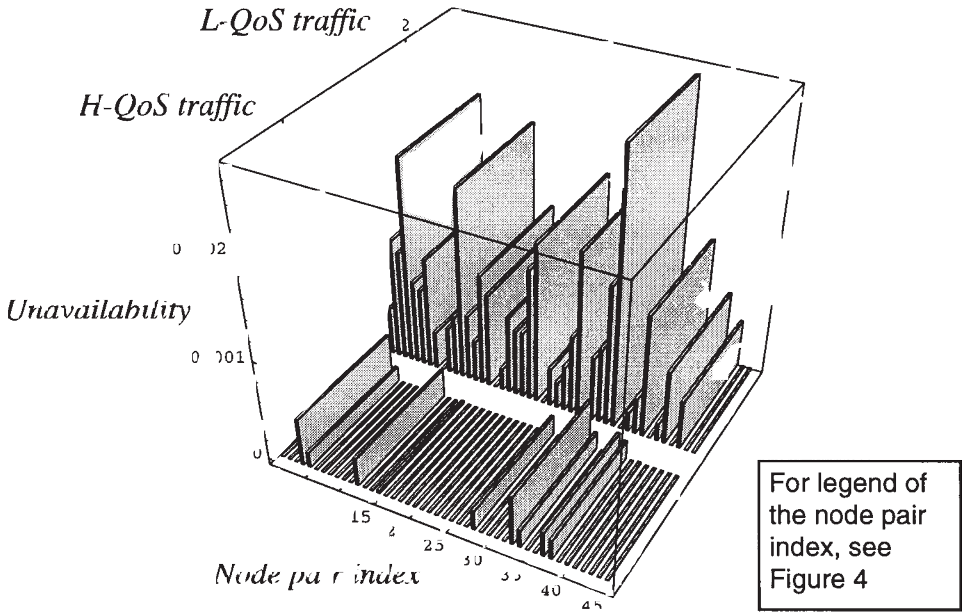
For L-QoS traffic, the lower limits for unavailability are in the 1–24 hours per year area. No major differences can be observed by comparing these unavailabilities to Figure 4 and Figure 5. Some of the unavailabilities increase slightly, but again, the situation for a few L-QoS traffic connections are improved.

5.3 All cases combined

Figure 7 shows the development of a mean unavailability, found by dividing the total lost traffic due to link failures by total offered traffic. Our main observation is that even for the H-QoS traffic, the unavailability increases very rapidly with a too high relative percentage of offered traffic. It is, however, still an order of magnitude less (on the average) than L-QoS traffic.

6 Conclusions

This study has demonstrated the potential gain of QoS differentiation. A main observation from the case study is that the end-to-end unavailabilities for the high QoS traffic is very sensitive to a proper dimensioning of the network. Thus, a tight dimensioning for the highest QoS cannot be recommended. Keeping in mind that "excess capacity" may be sold with lower QoS, more than two QoS levels are expected to improve the economical benefits of the approach, but this remains to be investigated. To guide the dimensioning process, a cost function including income from carrying the traffic of various QoS levels and penalties for not being able to deliver a promised availability should be established. This should be in a form usable for optimization with regard to income for the network operator.



For legend of the node pair index, see Figure 4

Figure 6 Unavailability with 60 % offered H-QoS traffic

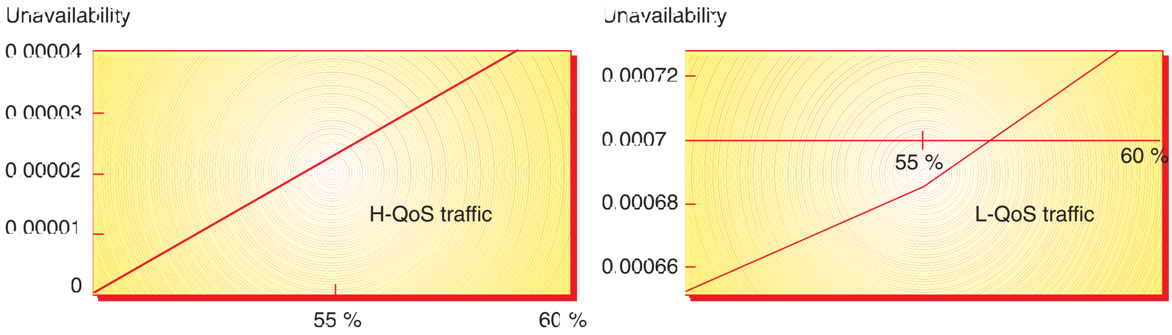


Figure 7 Average unavailability with an increasing relative portion of H-QoS traffic

More advanced back-up path strategies should also be considered. This will lead to a trade-off between the complexity of re-routing in case of failure, versus an increase of network cost. Hence, for a successful implementation of a QoS differentiated ATM network, the following issues should be considered:

- An implementation oriented study with respect to variants of pre-defined back-up path concepts (or more advanced methods, like flooding, could be included).
- Studies of improved dimensioning and optimization methods, hereunder uncertainties with respect to offered traffic and total investment gain trade-offs.

References

1 Anderson, J et al. Fast restoration of ATM networks. *IEEE JSAC,* 12, (1), January 1994.

2 Crochat, O, Le Boudec, J Y, Przygienda, T. A path selection method in ATM using pre-computation. *International Zürich Seminar 1996 (IZS'96),* February 19–23, 1996, ETH Zürich.

3 Kamsvåg, I. Spearhead against Telia. *Computerworld Norge,* No. 46, page 7, December 1995. (In Norwegian.)

4 Fagge, H. *Simulator for studies of self healing networks based on ATM mechanisms.* NTH diploma thesis, IDT, 1996. (In Norwegian.)

5 Fujii, H, Yoshikai, N. Restoration message transfer mechanism and restoration characteristics of double-search self-healing ATM network. *IEEE JSAC,* 12, (1), January 1994.

6 Gylterud, G, Solhaug, A. *Implementation and simulation of flooding algorithms to achieve self-healing in ATM networks.* NTNU project work, Div. of Telematics, 1996. (In Norwegian.)

7 Kajiyama, Y, Tokura, N, Kikuchi, K. An ATM VP-based self-healing ring. *IEEE JSAC,* 12, (1), January 1994.

8 Kawamura, R, Sato, K, Tokizawa, I. Self-healing ATM networks based on virtual path concept. *IEEE JSAC,* 2, (1), January 1994.

9 Kawamura, R, Tokizawa, I. Self-healing virtual path architecture in ATM networks. *IEEE Communications Magazine,* 33, (9), September 1995.

10 Kelly, F P. *Modelling communication networks, present and future.* The Clifford Paterson Lecture 1995, Proc. R. Soc. Lond. A (1995) 444, 1–20, 1995.

11 van Landegem, T, Vankwikelberge, P, Vanderstraeten. A self-healing ATM network based on multilink principles. *IEEE JSAC,* 12, (1), January 1994.

12 May, K P et al. A fast restoration system for ATM-ring-based LANs. *IEEE Communications Magazine,* 33, (9), September 1995.

13 Murakami, K, Hyong, S K. Virtual path routing for survivable ATM networks. *IEEE/ACM Trans. on Networking,* 4, (1), February 1996.

14 Nederlof, L et al. End-to-end survivable broadband networks. *IEEE Communications Magazine,* 33, (9), September 1995.

15 Oki, E, Yamanaka, N, Pitcho, F. Multiple-availability-level ATM network architecture. *IEEE Communications Magazine,* 33, (9), September 1995.

16 Shyur, C-C, Tsao, S-H, Wu, Y-M. Survivable network planning methods and tools in Taiwan. *IEEE Communications Magazine,* 33, (9), September 1995.

17 Sanso, B, Soumis, F, Gendreau, M. On the evaluation of telecommunications network reliability using routing models. *IEEE Transactions on Communications,* 39, (10), October 1991, 1494–1501.

ATM network performance parameters and methodology for deriving their objectives

BY INGE SVINNSET

1 Introduction

A major problem of B-ISDN is to guarantee the Quality of Service (QoS) given by the wide range of services that the B-ISDN will support. In order to plan, design and operate the broadband networks and services in an efficient way, an appropriate QoS framework has to be defined. This framework must have in mind the QoS requirements of a broadband service user. Specific QoS parameters for each type of service can be classified by using a 3x3 matrix proposed by ITU-T [12] and shown in Table 1.

Speed describes the time interval that is used to perform the function or the rate at which the function is performed with the desired accuracy. Accuracy describes the degree of correctness with which the function is performed with the desired speed. Dependability describes the degree of certainty with which the function is performed regardless of speed or accuracy, but within the given observation interval.

There appears to be some dispersion in the understanding of the concepts QoS, Network Performance (NP) and Grade of Service (GoS) and the relationships between them. Sometimes this leads to confusion when experts within slightly different technical fields have discussions and interchange results

Network Performance has been defined in ITU [11] as "the ability of a network or network portion to provide the functions related to *communications* between users". It is also stated that "network performance measures are meaningful to network providers and are quantifiable at boundaries of network portions to which they apply. Quality of service measures are only quantifiable at a service access point". Accordingly, network performance constitutes a set of impairment measures which shall either be directly measurable, or obtained as statistics from long term measurements, at certain defined measurement points at the edges of or within the interior (but in this case on the border between different network operators) of the public network. Note that these points are different from the service access points.

On the other hand, *Quality of Service* is defined as "collective effect of service performance which determine the degree of satisfaction of a user of a service." Quality of service (QoS)-parameter sets are defined as follows:

"For each QoS-parameter, a set of 'sub-parameters' is defined from among the following possibilities:

a. A *target* value which is the QoS value desired by the calling user;

b. The *lower quality acceptable value* which is the lowest QoS value agreeable to the calling user. (When the lowest quality acceptable refers to throughput, the term "minimum" may be used, while when it refers to transit delay, the term "maximum" may be used.);

c. An *available* value which is the QoS that the network is willing to provide;

d. A *selected* value which is the QOS which is the QOS value to which the called user agrees."

From the above definition we conclude that the QoS is a complex measure for user satisfaction with the service in question. This includes variables of the types which are also characterising network performance. But one should bear in mind that these are observable at different measurement points. Thus the values of the parameters may be different.

Table 1 ITU-T 3x3 QoS parameters matrix

	Performance criteria		
Function	Speed	Accuracy	Dependability
Access	Access speed	Access accuracy	Access dependability
Information transfer	Information transfer speed	Information transfer accuracy	Information transfer dependability
Disengagement	Disengagement speed	Disengagement accuracy	Disengagement dependability

Finally, let us consider *grade of service*. It has been defined in ITU [9] as follows: "A number of traffic engineering variables used to provide a measure of adequacy of a group of resources under specified condition; these grade of service variables may be the probability of loss, dial tone delay, etc."

By referring to traffic engineering variables it is understood that as far as these will be measurable at MPT/MPIs (see Section 3) or service access points, GoS variables may be identical to NP or QoS variables. However, GoS may also be characterised by parameters which are not measurable. One such case is when a circuit group operating as a delay system is characterised as a loss system, e.g. for dimensioning purposes.

Thus, there is no general correspondence between network performance and grade of service, and one should be cautious when attributing parameter values to the one concept or the other.

This article concentrates on the network performance (NP) of ATM networks. A general methodology for quantifying NP and QoS is given in Section 2. Both a "top-down" and a "bottom-up" procedure is described. Basic definitions of measurement points, reference events, reference connections and apportioning of end-to-end requirements are then given in Sections 3 and 4. Section 5 elaborates on the ATM layer information transfer performance and on the different factors that influence this performance. The possibility of offering different performance to different connections is discussed in Section 6 through the definitions of ATM layer QoS classes. QoS classes may also be introduced at the call level giving rise to priorities between calls what regards accessibility and dependability (QoS differentiation). Such differentiations are not treated in this paper. Section 7 contains definitions of connection processing performance parameters and a "bottom-up" procedure for calculating end-to-end connection processing performance. In Section 8 we treat long term characterisation of network performance, introducing the concepts of control plane availability and user plane availability and retainability. Finally, in Section 9 we make references to measurements for assessing and controlling network performance with some first indicative values for international connections within Europe.

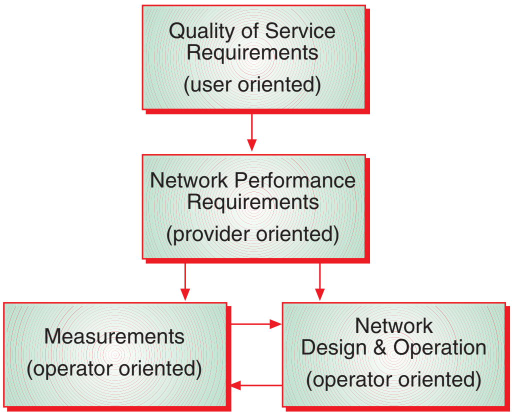


Figure 1 Process of establishing QoS and NP parameter requirements

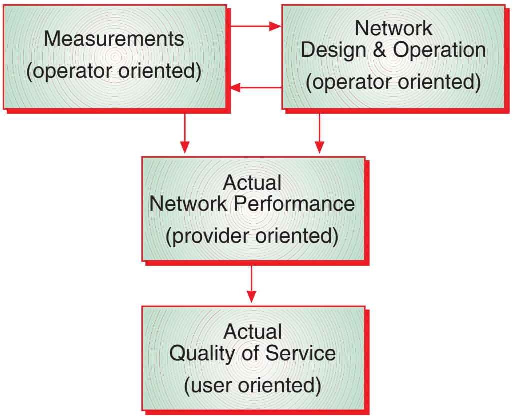


Figure 2 Process of establishing actual QoS and NP parameter values

2 The processes of quantifying NP and QoS parameters

These processes will be different depending on whether one is aiming at required values or actual values.

In the case of required values the logical point of departure is to consider a selected set of services, denominated e.g. "benchmark services" and determine the QoS required by *each* of these services, possibly represented as QoS parameter sets containing "target", "lower quality acceptable" and "selected" values corresponding to the ITU definition cited above. This requires a thorough exploration of the service characteristics and how the users perceive them. Typically, such results may originate from application trials. The results concerning the various benchmark services should then be "enveloped" together into consistent characteristic sets of QoS parameter requirements, each defining a QoS class.

The NP parameter requirements should then be specified on the basis of these characteristic sets of QoS requirements, making allowances for the customer premises network and the terminals. The NP requirements may then be used when designing the network (e.g. dimensioning) and the operations and measurement capabilities. As such they will have to be considered by the equipment providers.

The "top-down" procedure has been depicted in Figure 1.

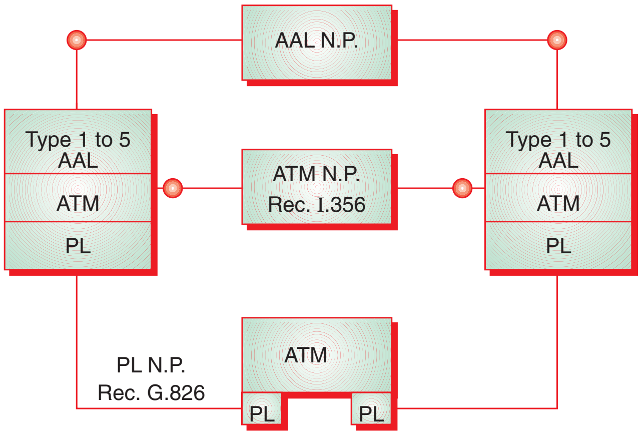


Figure 3 ATM Network Performance related to AAL NP and Physical Layer (PL) NP

In this case one is aiming at actual values one should perform a traffic study of the network, subject to a specified traffic environment, in order to determine the actual NP parameter values. If the network is already operating, information from measurements should also be taken into account.

The actual (or "available") QoS parameter values may then be established by taking into account the customer premises network and the terminal capabilities. This is a "bottom-up" procedure compared with the preceding one.

The QoS requirements of future B-ISDN services depend amongst others on layer processing and workload in host systems. Therefore, it is necessary to introduce a set of performance parameters whose properties indicate the nature and the requirements in the layered protocol stack. These parameters are defined for each layer. Figure 3 indicates the relation between the ATM layer network performance, respectively the physical layer performance and the ATM Adaptation Layer (AAL) performance.

3 Measurement points and reference events

As indicated above all B-ISDN information transfer performance parameters are specified at measurement points (MPs).

In [15] two types of measurement points (MP) for broadband ISDN are defined:

- *Ingress MP* – which is located at the input of the first equipment which accesses the ATM layer in a network operator domain,
- *Egress MP* – which is located at the output of the last equipment which accesses the ATM layer in a network operator domain.

Beside this classification, MPs can be divided into:

- *Measurement Point T (MPT)* – located at a User Network Interface (UNI)
- *Measurement Point I (MPI)* – located at an interface that terminates an ATM transmission system at an International Switching Centre (ISC).

For broadband ISDN, the location of MPI is on the international side of ISC at:

- The last egress MP in a given country
- The first ingress MP in a given country.

The establishment of the MP on the national side of the ISC and its performance allocation in the national portion are national matters, depending on the network topology of each country.

For the purpose of performance management, ATM connections are consequently divided into three types of connection portions:

- National Portions – national portion can be divided into two types:
  - MPT-MPI type, where the MPI is an egress MPI
  - MPI-MPT type, where the MPI is an ingress MPI.
- International Transit Portions – is delimited by a pair of MPIs, the first one is an ingress MPI and the second is an egress MPI, both located in the same transit country.
- International Inter-Operator Portions – is delimited by a pair of MPIs, where the first is an egress MPI and the second an ingress MPI, located in adjacent countries. Such a portion links
  - National Portion to an International Transit Portion
  - Two adjacent International Transit Portions
  - Two adjacent National Portions.

The set of International Transit Portions and International Inter-Operator Portions constitutes the international portion of a connection. The MPIs delimiting the International Portions is the first (egress) MPI and the last (ingress) MPI.

All performance measurements definitions, in this document, are defined in terms of corresponding reference events as defined in [15]. Reference events occur at the MPs, and when this occurs a measurement can be initiated or terminated. Two different reference events are defined in [15]: "An *exit event* occurs when a unit of control or user information crosses the MP exiting the SSN or CEQ into the attached transmission system". "An *entry event* occurs when a unit of control or user information crosses the MP entering the SSN or CEQ from the attached transmission system".

The exit and entry events may occur at either the same MP or at different MPs. The first corresponds to a single point measurement and the latter to a multiple point measurement, where the exit event is generated at one MP and the entry events occur at other MPs.

The semantics of reference events are different for each of the layers in the B-ISDN Protocol Reference Model (PRM). At the signalling layer, i.e. the OSI layer 3 – which corresponds to the upper layer of the control plane in the B-ISDN Protocol Reference Model (PRM), a reference event corresponds to a CEQ or SSN sending a signalling information unit. This information unit may have a size larger than 48 octets of information, and fill more than one cell. Reference events at the signalling layer are called Protocol Reference Events (PRE) [14].

The performance of the Control Plane in a VC-switched ATM network is assessed based upon measurement of time intervals between and relative frequency of significant events. These events refer to arrivals of specific messages belonging to the signalling protocol, according to the relevant ITU recommendations.

A consequence of using protocol messages as reference for event detection is the need to terminate the protocol as part of the detection process. The signalling protocol is terminated in Layer 3, which means that the messages are not observable at the layers below.

One frequently finds references to time instances of arrival for last or first bit of an information element of interest. Apart from bit there is a choice of octet, cell, AAL PDU, or the whole message as a delimitation unit. In the definition above it is stated a unit of control or user information. This must also be seen in view of what is practical and in view of the required accuracy.

If we e.g. assume that the values for the various Control Plane performance parameters are in the range of 100 ms, an accuracy of 1 % for the measurement would require a timing inaccuracy of < 1 ms. For feasible implementations this seems to limit the choice of delimiting event for definition of measurements to the instant of decoding of a signalling message in Layer 3. The reason for this limitation is the difficulty of assuring this level of accuracy in a "trace back" through the underlying layers.

In [12] it is stated that network performance is measured at the T or combined T/S reference point MPT, because this is where the private part is separated from the network. Hence, to determine the true values of network performance parameters they must be measured at these points.

However, the MPT is physically located at the user's premises, which means that it may be difficult for the network operator to access the physical MP. Since limited functionality is implemented in the Network Termination 1 (NT1) in the first phase of the ATM network, the delay in the NT1 will probably be negligible. The measurement points can thus be moved to some other locations of the access line, without reducing the measurement accuracy significantly. A more feasible location, from the network operator's point of view, is at the V reference point, i.e. the measurements are located at the access node.

Finally, the network performance measurements can be divided into two different types, *in-service measurements* and *out-of-service measurements*. In-service measurements are performed by monitoring performance parameters for "real life" traffic. Out-of-service measurements are based on measurements of performance parameters on artificially generated traffic.

4 Reference connection and apportioning of end-to-end requirements

Reference connections of connection-oriented bearer services depend on the location of VC switches. VC switches can be accessed in two ways:

- Direct access from user/CEQ to the VC switch
- Access through a VP cross-connect (or several cross-connects).

We assume that network elements that support signalling (i.e. switches) in most cases can also do cross-connecting (that is establishment by the management system without signalling). In the reference connection depicted in Figure 4 we have thus used a combined VC switch / VP cross-connect at the edges of the network. In this way we get a better resource utilisation than what is the case

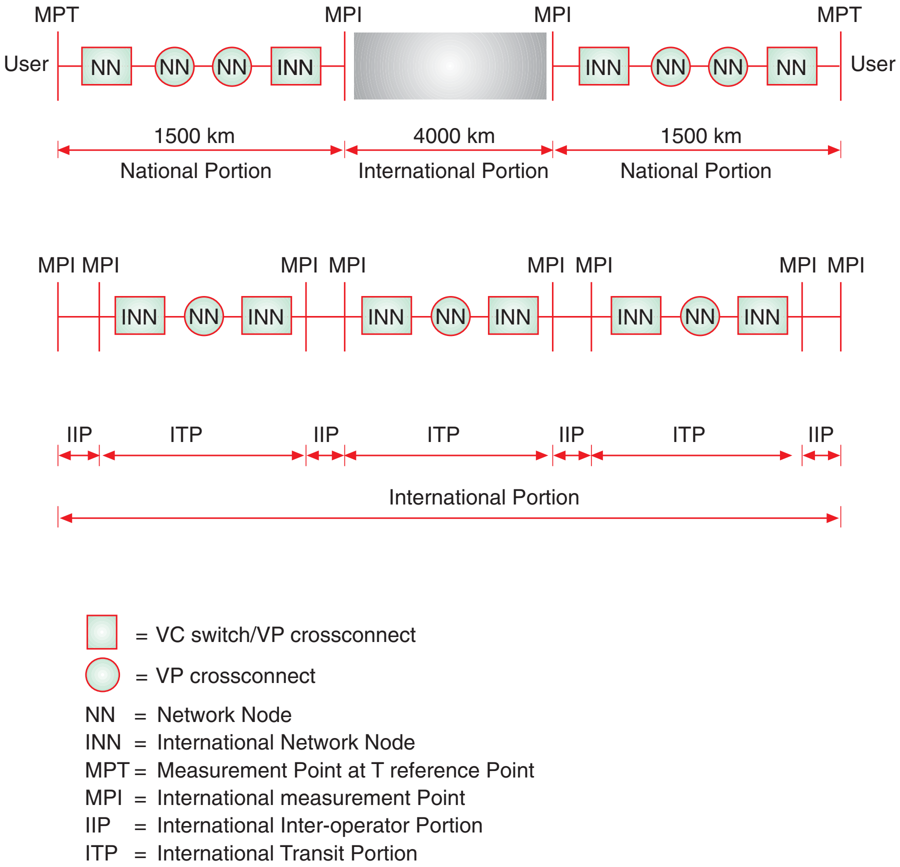


Figure 4 Hypothetical reference connection when VC switches are at the edges of the network

with VP cross-connects at the edges of the network at the cost of more VC switches in the network. Also problems of QoS support and UPC/shaping problems can be observed with the other solution, due to the fact that the first node does not access the VC layer. With this configuration the MPT is located at the user side of the national VC switch / VP cross-connect and it is possible to observe VCs on an individual basis.

In the international portion are used VP cross-connects and combined VC switch / VP cross-connect elements. In that way some resources can be saved in the case of point-to-multipoint connections to several countries at the same time.

From the description of the ABR service [1] it is easy to conclude that the end-to-end delay depends heavily on the distance between source and destination, number and type of network elements (NE), and the use of virtual source / virtual destination (VS/VD) functionality in intermediate nodes. If the ABR service is strictly implemented without VS/VD functionality in transit nodes, round-trip delay for RM cells may be huge. If these delays can be reduced the buffers needed for managing ABR connections may also be reduced. A reference connection for the ABR service should thus indicate the use of virtual source / virtual destination functionality in the nodes.

A virtual source assumes the behaviour of an ABR source end point. Backwards RM-cells received by a virtual source are removed from the connection. A virtual destination assumes the behaviour of an ABR destination end point. Forward RM-cells received by a virtual destination will not be forwarded to the next segment of the connection but sent back towards the source, with or without added information. In this way the connection is segmented into several ABR segments, which can be separately controlled. The use of control segments is a weighting of control overhead against buffer sizes, i.e. the longer (and fewer) control segments the longer buffers are needed. The coupling between two adjacent ABR control segments associated with an ABR connection is implementation specific.

The fact that ATM Forum's proposal allows NE manufacturers to choose freely any conceivable implementation and ABR resource allocation procedure increases the complexity of the problem for PNOs. It is thus proposed to implement VS/VD functionality at the network ingress. Since a network operator has no direct control of a network node which is located in an adjacent (private) network, it may also be practical to consider VS/VD functionality at the network egresses.

Taking into account all these facts, we may propose the following for the reference connection of the ABR service:

- In the national portion at least VS/VD functionality in the International Network Node (INN) and at the networks ingress/egress
- In the international portion each INN should be a VS/VD node.

Applying these proposals to the reference connection depicted in the figure means to add VS/VD functionality to each combined VC switch / VP cross-connect.

[17] gives rules for allocation of the end-to-end network performance objectives to the standardised network portion. That means that for each combination of QoS class and network performance parameter, the performance degradation between the portions indicated in the reference connection has to be allocated, in our case (Figure 4)

- 2 national portions,
- 3 international transit portions, and
- 4 international inter-operator portions.

For all the network performance parameters except CDV the sum of the network portion values will be the end-to-end network performance value. Within a network portion the values can be further deallocated to requirements on network equipment. A network operator can thus deallocate the allowed degradation within his domain depending on the way he wants to configure his network.

5 ATM layer NP parameters

The ATM layer NP is measured by a set of parameters intended to characterise the performance of an ATM layer connection. One or more values of these performance parameters may be offered on a per connection basis giving rise to QoS classes. Support of multiple different performance objectives can be done by routing the connection to meet different objectives, or by implementation-specific mechanisms within individual network elements.

The following information transfer performance parameters on the ATM layer are defined in [17]:

- Cell Error Ratio (CER)
- Cell Loss Ratio (CLR)
- Cell Misinsertion Rate (CMR)
- Severely Errored Cell Block Ratio (SECBR)
- Cell Transfer Delay (CTD)
- Cell Delay Variation (CDV).

These parameters are defined either end-to-end or for a connection portion, in both cases based on observations in two measurement points MP1 and MP2. They are defined with respect to reference events and cell transfer outcomes. According to [17] the two basic reference events are cell exit events and cell entry events. The main cell transfer outcomes, resulting from the analysis of the monitored reference events, are successfully transferred cell, errored cell, lost cell, misinserted cell, and severely errored cell block outcomes.

The outcome definitions are illustrated in Figure 5, where the cell reference event (CRE1) is the entry event at measuring point MP1 and CRE2 is the exit event at measuring point MP2. The CRE2 must occur within a specified time T_{max} after CRE1.

The evaluation of the network performance parameters based on the above analysis should include calculations of the following [17]:

- Error related parameters (cell error ratio, cell loss ratio, cell misinsertion ratio, severely errored cell block ratio)
- Delay related parameters (cell transfer delay, cell delay variation).

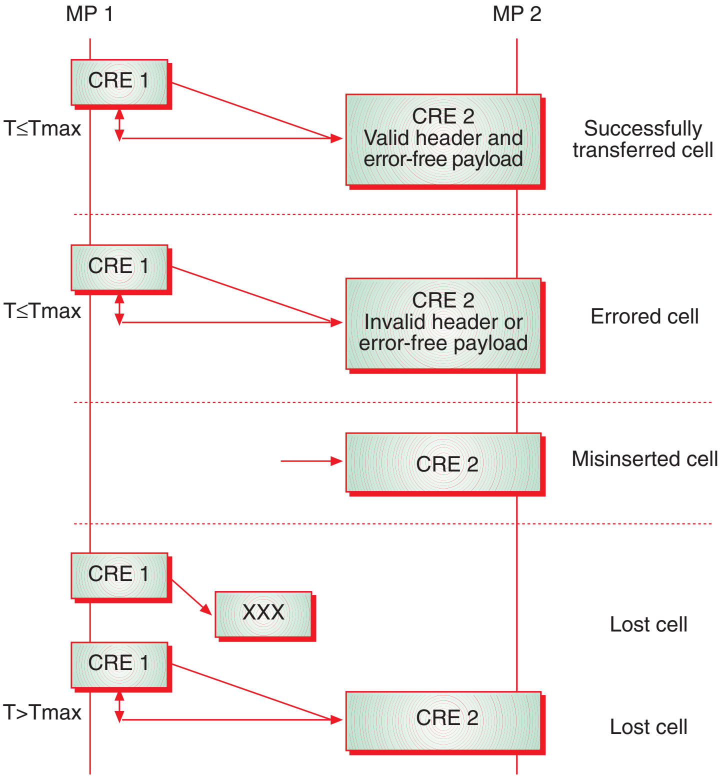


Figure 5 Cell transfer outcomes

An additional performance parameter, which may be considered as delay related, is the number of non-conforming cells, i.e. cells that arrive at intervals shorter than those allowed by the traffic contract specified in [19].

The *Cell Error Ratio* for an ATM connection is defined as:

$$CER = \frac{Errored\ \ Cells}{Successfully\ \ Transferred\ \ Cells + Errored\ \ Cells}$$

Successfully transferred cells and errored cells contained in cell blocks counted as severely errored cell blocks are not counted in the calculation of cell error ratio.

The main contribution to CER comes from transmission.

The *Cell Loss Ratio* for an ATM connection is defined as:

$$CLR = \frac{Lost\ \ Cells}{Total\ \ Transmitted\ \ Cells}$$

Lost cells and transmitted cells in cell blocks counted as severely errored cell blocks are not counted in the calculation of cell loss ratio.

Requirements for the CLR performance parameters are defined for a cell stream which complies to the traffic contract. When a connection is not complying to the contract no QoS can be guaranteed. However, if the number of non-conforming cells is below a given threshold, guarantees can be given taking into account that the non-conforming cells can be deleted by the network and that these cells should not be counted in the mentioned performance measures. A method for calculating an upper bound for the number of non-conforming cells is given in [17].

In [17] distinctions are made between cell loss ratio for the various cell streams within a connection defined by the CLP bit. Three special cases of interest are described, CLR_0 (the CLR for high priority cells), CLR_{0+1} (the CLR for aggregate cell stream and CLR_1 (the CLR for low priority cells).

The *Cell Misinsertion Rate* for an ATM connection is defined as:

$$CMR = \frac{Misinserted\ Cells}{Time\ Interval}$$

This performance parameter is defined as a rate (rather than ratio), since the mechanism producing misinserted cells is independent of the number of transmitted cells received on the corresponding connection.

Misinserted cell is caused either by miscorrected cell header or by disfunctioning of the ATM nodes (e.g. misroutings). Therefore, cell misinsertion can only be observed in a node.

The *Severely Errored Cell Block Ratio* for an ATM connection is defined as:

$$SECBR = \frac{Severely\ Errored\ Cell\ Blocks}{Total\ Transmitted\ Cell\ Blocks}$$

A cell block is a sequence of *N* cells transmitted consecutively on a given connection. A SECBR outcome occurs when more than *M* errored cell, lost cell, or misinserted cell outcomes are observed in a received block. The values for *M* and *N* have been given in [17] as:

$$M = N / 32$$

$$N = 2^{\lceil \log_2 (PCR/25) \rceil}$$

where PCR = peak cell rate in cells/s (1600 < PCR < 819200).

The minimum block size is 128 cells, i.e. for PCR < 3200 cells/s the block size *N* should be equal to 128. It is not yet agreed whether to have block sizes larger than 1024.

This definition of SECBR is only valid for conforming traffic. In cases with unspecified CLR objectives for the aggregated (CLP = 0 + 1) cell stream the estimation of the SECBR has to be based on the CLP = 0 cell stream only. To be more specific, the aggregated cell flow shall be counted to determine the cell block, but lost CLP = 1 cells are not counted. Errored and misinserted CLP = 1 cells, however, are counted. In cases with lower rates the definition of severely errored seconds becomes problematic, see the section on availability.

The *Cell Transfer Delay* is defined as elapsed time between a cell exit event at a given measurement point MP1 (usually the *T* reference point at the sending side) and the corresponding cell entry event at another measurement point MP2 (usually the *T* reference point at the receiving side) for a particular connection. The CTD between two MPs is the sum of the following:

- Total inter-ATM node transmission delay – caused by transmission related equipment between two adjacent ATM nodes (e.g. SDH cross-connect is considered to be part of this delay).
- The total ATM node processing delay between the actual MPs – caused by queuing, switching, routing. Due to queuing in ATM nodes, this component is variable on a cell-by-cell basis within one ATM connection, causing *cell delay variation* (CDV).

There are two performance parameters associated with cell delay variation:

- 1-point cell delay variation – describes variability in the pattern of cell arrival events observed at a single MP with reference to the negotiated peak rate as defined in [19]. It includes the variability present at the cell source (customer equipment) and the cumulative effects of variability introduced (or removed) in all connection portions between the cell source and the specified MP. It can be related to cell conformance at the MP and to network queues. It can also be related to the buffering procedures that might be used in AAL to compensate CDV.
- 2-point cell delay variation describes variability in the pattern of cell arrival events at the output of a connection portion (MP2) with reference to the pattern of the corresponding events observed at the input to the connection portion (MP1). It includes only the delay variability introduced within the connection portion. It provides a direct measure of portion performance and an indication of the maximum (aggregate) length of cell queues that may exist within the portion.

The CDV can be expressed by $k\tau$, where τ is the cell service time, and it is a function of the bitrate.

In setting NP parameter objectives, the factors which influence the NP parameters have to be considered. These factors can be called sources of NP degradation, and can be the following:

- Propagation delay – delay caused by the physical media which transports the bits comprising ATM cells between UNIs and ATM network nodes. It depends on the distance between source and destination.
- Link rate – the lower the link rate, the higher the CDV degradation.
- Physical media errors – is the bursty bit errors that are introduced by physical media.
- Switch architecture – the overall architecture of the switch can have significant impact on performance. Some aspects that may be considered are switching matrix design, buffering strategy and the switch characteristics under load.

Table 2 Degradation of QoS Parameters

ATTRIBUTE	CER	CLR	CMR	CTD	CDV
Propagation Delay				x	
Link rate					x
Physical Media Errors	x	x	x		
Switch Architecture		x		x	x
Buffer Capacity		x		x	x
Number of Nodes	x	x	x	x	x
Traffic Load		x	x	x	x
Failures		x			
Resource Allocation		x		x	x

- Buffer capacity – the actual capacity of the buffer in units of cells supporting a link, within an ATM matrix or in other elements of an ATM switch.
- Traffic load – load offered by the set of ATM VPC/VCCs that share part of the same route as the VPC/VCC under consideration.
- Number of network nodes – the number of network nodes in the reference connection effects the NP objectives.
- Resource allocation – this is the capacity allocated to the VPC/VCC or to a set of VPC/VCCs. The resource allocation strategy effects the NP.
- Failures – events that impact availability, such as port failures or link failures, failures that introduce cell losses.

How various sources of degradation can influence the performance parameters is depicted in Table 2.

6 ATM layer Quality of Service classes

Within an ATM layer QoS class CTD, CDV and at most two CLR parameters may be specified, i.e. for the CLP = 0 cell stream (high priority cells) and for the aggregated cell stream (both CLP = 0 cells and CLP = 1 cells) of the ATM connection.

The QoS class scheme and candidate values for international (end-to-end) information transfer performance parameters, proposed by ITU-T (ref. ITU) are given in the Table 3.

As indicated CER, CMR and SECBR do not differ between the QoS classes. This is due to the fact that these are merely a consequence of errors in switching and transmission equipment. *U* means unbounded or unspecified. Operators may indicate a value but this should be very relaxed.

Also in the *U* (unspecified) QoS class no objective is specified for performance parameters. However, the network provider may determine a set of internal objectives for the performance parameters. In fact, these internal performance parameter objectives do not have to be constant during the call duration. Services using unspecified QoS class may have explicitly specified traffic parameters. The *U* for SECBR for this class is a consequence of having no requirement for CLR.

An example application of the unspecified QoS class is the support of so-called "best-effort" service. For this type of service, the user has to select the best-effort capability, the unspecified QoS class and the traffic parameter for the Peak Cell Rate.

Class 1 is intended for delay sensitive services. The 400 msec delay is for an intercontinental reference connection including a satellite hop. For a terrestrial connection, using short buffers as would be the case for delay sensitive services, an upper bound for the transmission delay could be estimated by

$$\text{CTD} < 6.25\ \mu\text{s} \cdot s + 300\ \mu\text{s} \cdot n$$

where s is the length of the connection in km and n is the number of ATM switching and cross-connect nodes. The estimation of 6.25 µs/km for transmission includes delay in transmission terminal equipment and transmission switches. This gives beneath 50 ms for a European terrestrial connection as indicated in the reference connection above.

The requirement for CLR and (CER) is a common agreement within ITU based on experiences within pilot networks. We must foresee that this can be improved in the near future, at least with the use of the best available equipment. An end-to-end CLR requirement of 10^{-8} for the highest quality class and a CER in the order of 10^{-7} should thus be possible to offer in the near future.

The difference between class 2 and class 3 is that class 2 is intended for services not using the CLP bit. Class 3 is for services using the CLP bit. In the latter case there is no requirement for the low priority cells (CLP = 1).

All requirements are defined for the lifetime of the connection. In practice, this means that a measurement will end whenever the connection is disconnected or a maximum time has expired. The maximum length of the measurements is for further study. The CLR, CER and SECBR are computed as the ratio of lost cells, errored cells and severely errored cell blocks over the measurement interval. The CMR requirement is the requirement for the mean value over the specified measurement interval. When a small number of cells are observed it is of course possible that the computed values will be higher than the specified values.

The CTD requirement is a requirement for the mean observed CTD during the measurement period. Bearing in mind that CDV is rather low compared to CTD for class 1 services, the maximum observed CTD could just as well have been chosen. ITU-T has defined 2-point CDV requirement as an upper bound on the difference between the 10^{-8} and the $1 - 10^{-8}$ quantile of the CTD distribution. For measurement purposes the following definition could be used: The CDV requirement is a requirement for the difference between the maximum observed CTD and the minimum observed CTD during the measurement period. The length of this measurement period should be fixed but is yet to be defined.

QoS depends on the nature of the impairment, i.e. the distribution or burstiness. A given CLR may thus give rise to different observed phenomena at the user site depending on the distribution of cell losses and errors. Such effects are only partly covered by the definition of

Table 3 ITU-T proposal for QoS classes (U = Unbounded or Unspecified)

	Class 1	Class 2	Class 3	Class U
CTD	400 msec	U	U	U
CDV	3 msec	U	U	U
CLR 0+1	$3 \cdot 10^{-7}$	10^{-5} or TBD	U	U
CLR 0	none	none	10^{-5}	U
CER	$4 \cdot 10^{-6}$	$4 \cdot 10^{-6}$	$4 \cdot 10^{-6}$	$4 \cdot 10^{-6}$
CMR	1/day	1/day	1/day	1/day
SECBR	10^{-4}	10^{-4}	10^{-4}	U

SECBR and much works remains in this area. SECBR is also a basis for the definition of user plane availability. This will be covered in Section 8.

7 NP parameters to characterise Control Plane Performance

Call processing performance for B-ISDN is currently an issue in ITU-T SG 2, SG 11 and SG 13. Whereas SG 11 is addressing the call processing performance in the ATM exchanges, both SG 2 and SG 13 address the end-to-end performance. In [10] (SG 2) the end-to-end performance is covered from a dimensioning point of view with objective values for normal load (i.e. dimensioning load), high load and under failure conditions. [13] (SG 13) addresses design objectives for worst case, represented by a reference connection, end-to-end call processing performance. Co-ordination is taking place between these study groups to assure consistency between the recommendations.

In B-ISDN networks a call is defined as an association between two or more parties. This association contains connections (zero connections is possible) and relations between these connections. The handling of a call may thus be decomposed into the handling of a set of connections and parties. The call processing performance can therefore be deduced from the connection processing performance and the add / drop of parties performance, once the composition of the call is known. For the single connections objective values may be defined. Objective values may also be defined for the process of adding and dropping parties in a call.

The connection processing performance parameters concern the establishment, release and parameter change of connections and parties, i.e.

- Connection rejection probability
- Add party rejection probability
- Connection establishment delay
- Party establishment delay
- Answer signal delay
- Look ahead response delay
- Connection release delay
- Party release delay

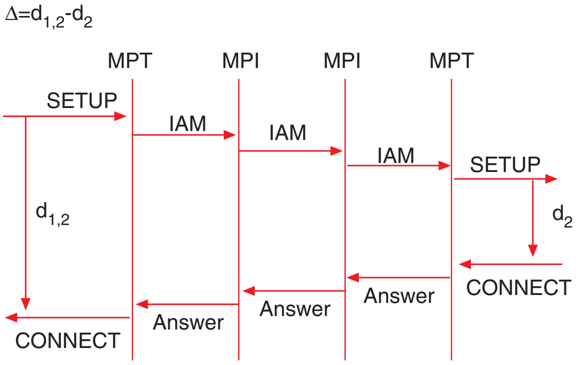


Figure 6 Connection establishment delay

- Connection modification rejection probability
- Connection modification delay.

Some initial definitions of these parameters are given below. They are based on information flows from the signalling protocols for release 1 and 2 B-ISDN now being developed by ITU-T SG 11 ([22] and [23] for the release 1 signalling protocols). Some of these parameters are being defined by ITU-T SG 2 in [10] and by ITU-T SG 13 in [13] and these definitions have been used as a basis. We here propose some additions. Our definition of connection rejection probability is more in line with the SG 13 approach since it corresponds both to lack of resources and to other errors. The exact events defining the performance parameters should be Protocol Reference Events observed at the relevant measurement points.

The parameters are divided into two main groups:

- End-to-end parameters
- Local parameters (grade of service parameters).

7.1 End-to-end parameters

The *connection rejection probability* may be estimated as the ratio of unsuccessful VCC establishment attempts to the total number of VCC establishment attempts in a specified period of time. An unsuccessful VCC establishment attempt is an establishment attempt to a valid number, properly ‘dialled’, for which a called party ALERT or CONNECT message or a RELEASE or RELEASE COMPLETE message initiated by the receiving side is not observed at the MPT on the calling user’s side within T_{max} from the instant when the SETUP message necessary for outgoing VP/VC selection is observed at the same MPT. The value of T_{max} should be chosen in accordance with relevant time-out values (i.e. second consecutive expire of timer T303).

The RELEASE and RELEASE COMPLETE messages contain a cause information element with a location value indicating which part of the network is causing the release. Cases where customer premises equipment is not connected should be excluded in the statistics. In case the VCC is rejected only due to the network’s inability to establish other VCCs belonging to the same call, the connection establishment attempt shall be counted as successful for the considered VCC. This means that for the establishment of a point-to-point bi-directional call, the connection rejection probability shall be given for each direction separately. From this the rejection probability of the bi-directional call can be calculated.

Lack of control plane resources during the connection set-up phase may give rise to end-to-end blocking. If we assume peak rate allocation or equivalent bandwidth allocation only, the connection rejection probability is dependent on the required bandwidth of the connection and the required QoS. The connection rejection probability should thus be given as a set of curves, one for each QoS class (at the ATM layer), with the connection rejection probability as a function of the bandwidth demand (or one curve for the mean value and one for the 95th percentile value). The connection rejection probability also depends on the load and could be specified/measured both for

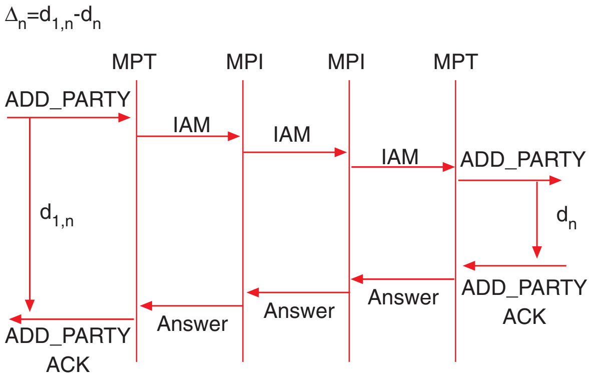


Figure 7 Party establishment delay

normal (i.e. dimensioning) load and overload (see Figure 9). Instead of curves objective values for classes of connections could be given, e.g. a maximum connection rejection probability of 10^{-3} per node for bandwidth demands less than 0.5 Mb/s.

The connection rejection probability objective determines the dimensioning, under the assumptions of resource allocation policy and traffic environment. In most cases the operators will specify their own objective values for this parameter depending on their dimensioning policy. But in future it may also be that this parameter will be part of a call level QoS framework giving some customers priorities in accessing the network.

The *add party rejection probability* is a similar parameter for adding of new parties to a call. It may be estimated as the ratio of unsuccessful ADD PARTY attempts to the total number of ADD PARTY attempts in a specified period of time. An unsuccessful ADD PARTY attempt is an ADD PARTY attempt to a valid number, properly ‘dialled’, for which a PARTY ALERTING or ADD PARTY ACK message or an ADD PARTY REJECT message initiated by the receiving side is not observed at the MPT on the calling user’s side within T_{max} from the instant when the ADD PARTY message is observed at the same MPT.

Similar observations as for connection rejection probability can be made for this parameter.

The parameters for connection establishment delay defined by ITU-T SG 2 and SG 13 are slightly different. [10] defines post-selection delay as the time interval from the initial SETUP message is passed by the calling terminal until indication of call disposition is received by the calling terminal. Another way to define this parameter is as follows.

The *post-selection delay* is the period starting with the instant when the SETUP message is observed at the MPT on the calling user’s side, and finishing with the instant when an ALERT or CONNECT message or a RELEASE or RELEASE COMPLETE message initiated by the receiving side is observed at the same MPT.

This end-to-end set-up message response delay will include some processing time in the far end CEQ. The ALERT message will be used in case of manual answering terminals and for some automatic answering terminals. In these cases the time delay for delivering the CONNECT message from the called user to the calling user (answer signalling delay as defined below) is not included.

[13] defines *connection establishment delay* as follows (illustrated in Figure 6 for a successful connection establishment).

The connection establishment delay associated to a measurement point MP is defined as the time from the instant a SETUP/IAM message is observed at the MP until the corresponding CONNECT/ Answer message is observed at the MP.

The connection establishment delay between two measurement points MP1 and MP2 (numbered from originating towards destination side) is defined as a difference between the connection establishment delay associated to MP1 and the connection establishment delay associated to MP2. The *end-to-end connection establishment delay* is the establishment delay between the MPT at the two sides of the connection. The end-to-end connection establishment delay (Figure 6) is thus

$$\Delta = d_{1,2} - d_2$$

This definition of set-up delay allows for allocation of set-up delay to the network portions. The end-to-end connection establishment delay is based on simultaneous measurements at both MPTs involved and does not include the processing time at the customer premises. In practice, measurements may take place at the two sides independently. The first measurement gives the end-to-end establishment including the response time at the far end side. This figure may be corrected by separate measurements giving the customer equipment response time statistics.

More exact reference events are given in [13]. In case of access signalling [23] the relevant messages are SETUP and CONNECT. In case of network signalling [22] the relevant messages are IAM and Answer.

If an adequate VP with sufficient bandwidth cannot be provided, procedures for setting up a new VP or to renegotiate VP bandwidth will be involved. This will affect the total set-up time.

[13] also defines the delay for establishment of an extra party to a call (Figure 7). The *party establishment delay* associated to a measurement point MP is defined as the time from the instant an ADD PARTY/IAM message is observed at the MP until the corresponding ADD PARTY ACK/Answer message is observed at the MP.

The party establishment delay between two measurement points MP1 and MP2 (numbered from originating towards destination side) is defined as a difference between the party establishment delay associated to MP1 and the party establishment delay associated to MP2. The end-to-end party establishment delay is the establishment delay between the MPT measurement points at the two sides of the connection, i.e. for party *n* it is defined as:

$$\Delta_n = d_{1,n} - d_n$$

More exact reference events are given in [13]. In case of access signalling (Q.2931) the relevant messages are ADD PARTY and ADD PARTY ACK. In case of network signalling (Q.2762) the relevant messages are IAM and Answer.

Answer signal delay is defined in [10] as the time interval from the CONNECT message is sent from the user at the receiving side until the CONNECT message is received by the user at the calling user's side. More precisely, it is the time interval from the instant that the CONNECT message is observed at the MPT on the called user's side, and finishing with the instant when the corresponding CONNECT message is observed at the MPT on the calling user's side.

The look-ahead procedure ([20] and [25]) allows the network to check whether the addressed terminal(s) are compatible and free. This procedure is supposed to be used in many broadband services in order to increase QoS and network utilisation. It may be initiated by the access switch at the calling users side upon receipt of a SETUP message and shall then be used prior to the establishment of the call/connection. The switch will then only send an IAM message towards the called user upon receipt of a positive response of the look-ahead message.

The *look-ahead response delay* is measured in the access switch of the initiating user. It is the time interval from the instant the look-ahead message LA Invoke is observed at the measurement point of the outgoing SVCC towards the receiving user and until the instant when the look-ahead response message LA Result is observed at the measurement point of the incoming SVCC from the same user.

The *Connection Modification Rejection Probability* is yet defined by neither SG 2 nor SG 13. It may be estimated as the ratio of unsuccessful modification attempts to the total number of modification attempts in a specified period of time. An unsuccessful modification attempt is a valid modification attempt for which a MODIFY CONNECTION ACK is not observed at the MPT at the initiating user's side within T_{max} seconds from the instant when the MODIFY CONNECTION message was observed at the same MPT.

A non valid modification attempt will be rejected with a cause value specifying the reason. These attempts shall not be counted. The value of T_{max} should be chosen in accordance with relevant timeout values (i.e. expiry of timer T360).

If a RELEASE message is initiated by either side before a MODIFY CONNECTION ACK/REJECT message is received by the initiating user's signalling system and within T_{max} seconds from the MODIFY CONNECTION message was received by the network, the attempt shall not count in the statistics.

The first parameter to be considered for renegotiation is VC bandwidth. The rejection probability will then depend on the extra bandwidth needed, the QoS of the connection and the network and control plane load.

Even in the case when reduction of bandwidth is demanded, the response may be delayed or lost due to errors or heavy load in the control plane. Any requirement for rejection probability in such cases will be different from requirements for increasing bandwidth by orders of magnitude.

It may also be possible to renegotiate QoS.

The *end-to-end connection modification delay* is defined as the time from the instant a MODIFY CONNECTION message is observed at the MPT of the initiating user until the corresponding MODIFY CONNECTION ACK message is observed at the same MPT.

The response time in renegotiations should be much less than the set-up time.

In B-ISDN users are allowed to negotiate the cell rate traffic parameters during the call/connection establishment phase ([21] and [24]). For a point-to-multipoint call it is only possible during establishment of the first party. The user will add either an alternative ATM traffic descriptor or a minimum acceptable ATM traffic descriptor to the SETUP message. In this way he will indicate the possibility of accepting lower bandwidth if the demanded bandwidth is not available. The alternative ATM traffic descriptor indicates a single alternative whereas the minimum acceptable ATM descriptor gives a range of acceptable values (restricted to negotiation of peak cell rates).

The possibility of negotiation during connection set-up will decrease the connection rejection probability. The probability of rejection of the connection request equals the probability of asking for the minimum bandwidth. The probability that the default values are accepted equals the probability of success given that the negotiation option is not used. The probability of success given that the default values cannot be accepted is a conditional probability. This value will very much depend on allocation policies like use of service protection methods.

7.2 Local parameters

These parameters are intended for performance characterisation and require-

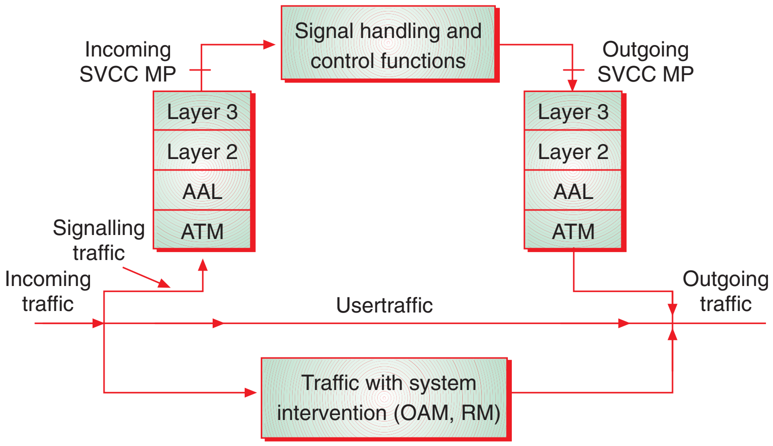


Figure 8 Location of measurement points for control plane measurements in a signalling and switching node

ments for B-ISDN switches. End-to-end requirements (based on service requirements) can be mapped to individual components in the network (Section 4), and will in this way give input to specifications for network equipment. We can also go the other way around and estimate end-to-end performance based on figures describing performance of switches and transmission equipment. A way to do this is described in Section 7.3.

As explained before a consequence of using signalling protocol messages as reference for the event detection is the need to terminate the protocol as part of the detection process. In Figure 8 the protocol stack with ATM layer, ATM Adaptation Layer (AAL) and Layer 2 and Layer 3 of the relevant signalling protocol is indicated. The signalling protocol is terminated in Layer 3, which means that the observations must be made between this termination and the signalling handling and control functions in the node, as near the Layer 3 termination as possible.

In what follows are given some definitions of local control plane performance parameters.

The *SSN connection establishment delay* is the time interval from the instant when the SETUP/IAM message is observed at the measurement point of the incoming signalling VCC (SVCC) until the instant when the message has been processed, outgoing VPI/VCI has been selected and the SETUP/IAM message is observed at the measurement point of the outgoing SVCC towards the receiving user.

The *SSN party establishment delay* is the time interval from the instant when the ADD PARTY/IAM message is observed at the measurement point of the incoming SVCC until the instant when the message has been processed, outgoing VPI/VCI has been selected and the ADD PARTY/IAM message is observed at the measurement point of the outgoing SVCC towards the receiving user.

The *SSN answer signal delay* is the time interval from the instant when the CONNECT/Answer message is observed at the measurement point of the incoming SVCC until the instant the CONNECT/Answer message is observed at the measurement point of the outgoing SVCC towards the receiving user.

The *SSN connection modification delay* is the period starting with the instant when the MODIFY CONNECTION message is observed at the MP of the incoming SVCC, and finishing with the instant when the message has been processed, necessary resources have been reserved and a MODIFY CONNECTION message has been observed on MP of the SVCC outgoing towards the other user.

Connection release delay is defined in [10] and [13].

The *connection release delay* is the time interval starting with the instant when the RELEASE message is observed at the MPT on the disconnecting user's side, and finishing with the instant when a RELEASE COMPLETE message is observed at the same MPT.

This parameter has only local significance, since the RELEASE COMPLETE message does not imply an acknowledgement of clearing from the remote user.

Party release delay is defined in [13]. It is the time interval starting with the instant when the DROP PARTY message is observed at the MPT on the disconnecting user's side, and finishing with the instant when a DROP PARTY ACK message is observed at the same MPT.

This parameter has also only local significance, since the DROP PARTY ACK message does not imply an acknowledgement of clearing from the remote user.

7.3 Calculations of end-to-end connection processing performance

The connection processing performance will depend on the performance of both the signalling networks and the ATM switching nodes. The connection processing includes essentially the same tasks as for narrowband, but we have to take into account that

- Necessary bandwidth must be calculated and reserved if possible (CAC)
- Policing parameters must be handled
- Outgoing link and VP must be selected taking into account available bandwidth, QoS requirements and routing criteria
- VCI value must be selected.

Network failures or extraordinary traffic conditions may give rise to a call/connection processing overload in some (smaller or larger) parts of the network. This kind of phenomenon, and the potential of possible counter measures, must be taken into account when specifying connection processing performance requirements.

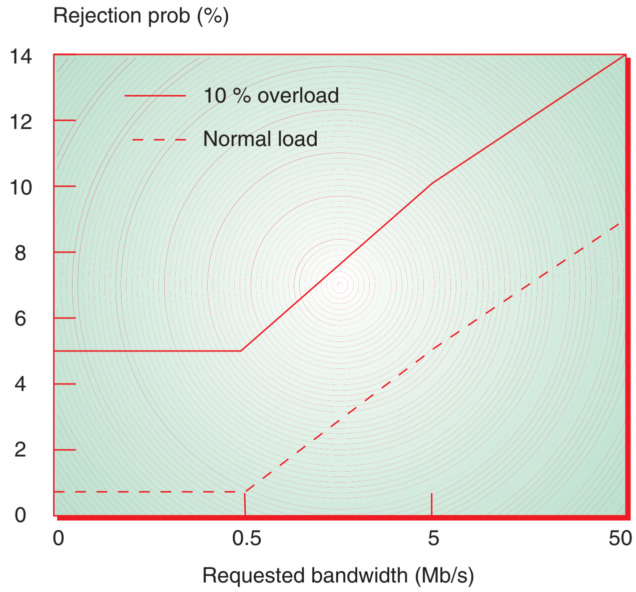


Figure 9 Example of objective values for connection rejection probabilities (STM-4 link)

Table 4 Transit exchange cross-office handling time

Message type	Exchange call attempt loading	T_{ch} (ms)	
		mean	95 %
Simple	Normal	80	160
(e.g. answer)	+15 %	120	240
	+30 %	200	400
Processing intensive	Normal	160	320
(e.g. IAM)	+15 %	240	480
	+30 %	400	800

One way to do this is to specify parameter requirements with respect to various levels of connection processing overload. The choices should be based on realistic assumptions about the various traffic processes involved, present status of switch performance and the QoS needed by a chosen set of services.

To be able to define objective values for connection processing some assumptions must be made that limit the signalling traffic of the users. Otherwise, it may be possible for users with intelligent B-ISDN terminals to overload the switches with complex call attempts and thus degrade call processing performance. Such assumptions are listed in [13], e.g. it should not be possible to make a new call attempt before the last one is completed.

The values for connection rejection probabilities indicated in Figure 9 are only tentative values and must be compared with service requirements. Input on this matter is required. The example is for an STM-4 link. The rejection probabilities will depend on the requested bandwidth as a fraction of the total available bandwidth on the link. It will be difficult to give reasonable objective values for high bandwidth connections requiring 5 – 10 % of total available bandwidth and beyond, because this will result in a very poor network utilisation. For such connections one should normally make pre-reservations.

The objective values shall be understood as long term values within "busy periods" and for periods of time where the control plane is available, i.e. we are talking about long term connection rejection probability. For the definition of short term connection rejection probability and control plane availability, see Section 8. The term "normal load" refers to dimensioning load.

The worst-case objective for the end-to-end connection establishment delay Δ may be specified as:

Mean of $\Delta < \Pi$ and 95 %
- tile of $\Delta < \Psi$

for given values of Π and Ψ, in any "busy hour" period, both for normal (i.e. dimensioning) load and for overload.

The objective values of the other connection processing parameters may be specified in a similar way.

In Appendix 2 of [13] some delay figures for the processing of signalling messages of a broadband exchange are given. These delays can be seen as having two components, the outgoing link delay T_{od} and the cross office handling time T_{ch}. The component T_{od} is the queuing delay for signalling traffic and depends on the traffic load, the bitrate of the outgoing link and the MSU length. The component T_{ch} depends on the implementation of the broadband exchange. The proposed values for T_{ch} are given in Table 4. These values are preliminary and should be reviewed carefully.

With such figures as a basis provisional end-to-end delay objectives can be calculated by adding mean values. The end-to-end 95 percentile values can be calculated using the following formula [2]:

$$T_{95} = T_m + \sqrt{\sum (t_{95,i} - t_i)^2}$$

where

T_m = mean end-to-end delay

t_i = mean delay in exchange i

$t_{95,i}$ = 95 percentile of delay in exchange i.

The end-to-end values depend very much on the structure of the signalling network. The number of nodes in this network can be reduced by use of VP network. By use of associated signalling an end-to-end set-up delay of 2 – 3 seconds can be indicated for a European reference connection (see Figure 4). A reduction of the number of intermediate nodes can substantially decrease this figure.

8 Long term network performance characterisation

In [11] the network performance, from the service provider's viewpoint, is related to a set of network characteristics that can be defined, measured and controlled to achieve a satisfactory level of service quality. The service quality as perceived by the user, depends on a variety of parameters.

The network dependability (see Table 1) is usually expressed in qualitative terms and its quantification is mainly done through the availability performance. The long term network performance parameters specified below will thus focus mainly on the availability aspects of network performance related to information transfer and call processing (user plane and control plane, respectively).

8.1 Control plane availability performance criteria

Parameters characterising the call processing performance of B-ISDN are elaborated in Section 7. A call may involve one or more connections and these connections may be point-to-point (type 1) or point-to-multipoint (type 2).

Reference configurations are shown in Figure 4 and indicate the measurement points at which call processing performance for significant events are observed and relevant parameter values are determined. The more detailed location of the measurement points in the switches are detailed in Figure 8. The complete reference model should also include the time intervals at which these parameters should be tested against the specified performance objectives, as well as the loading conditions under which the objectives are to be met.

The *short term connection rejection probability* is defined [4] as the estimated end-to-end connection rejection probability over an interval of fixed duration *T*. The control plane is considered unavailable during this time interval if the estimated short term connection rejection probability over this interval exceeds a given threshold P_{max} (e.g. 10 %). The *control plane unavailability* (network inaccessibility) is defined as the long term proportion of such time intervals for which the threshold is exceeded.

The value of *T* can be service dependent. A typical value could be *T* = 10 seconds. We propose that the decision of unavailability should be based on measurements in one access node (incoming and outgoing SVCC measurement point at the access side), i.e. the threshold is for end-to-end rejections of connection attempts originating in the local network belonging to a given local access switch. We also propose that connection attempts always belong to the intervals where the response or time-out occurs.

8.2 User plane availability estimation

The error related Information Transfer Performance parameters, i.e. cell error ratio, cell loss ratio, cell misinsertion ratio, severely errored block ratio, and the delay related parameters, i.e. cell transfer delay, cell delay variation, may be measured periodically in the in-service or the out-of-service mode.

From the information transfer viewpoint, the availability (or unavailability) of an ATM connection is defined as the fraction of time for an ATM connection (during which the network provider has agreed to make available this connection) that the connection is able (or not able) to support the information transfer. In order to decide whether an ATM connection has entered into the unavailable state, a criterion should be defined, derived from information transfer parameters and applicable on the ATM connection when the user transmits or does not transmit cells. Such a criterion is the occurrence of an unacceptable second, i.e. a second during which user cells are transmitted with degraded performance.

There are ongoing work both in ETSI [4], [5] and in ITU-T SG13 [18].

The definitions of *availability of semi-permanent connection portions*, as now given by ETSI and ITU-T, are given below. These definitions apply only to compliant connections and the definitions are very much dependent on the definitions of SECBR (which are not yet stable) and the definition of a SES_{ATM} outcome.

A SES_{ATM} outcome occurs when either

- An interruption has occurred that prevents cells from being transferred during the one second period of time (defect physical layer or ATM layer, interruptions due to users excluded)
- Cell transfer performance parameters computed over the one second period of time exceed given threshold values (CLR and SECBR are candidate parameters).

It has been proposed in ITU that a minimum of one Performance Monitoring (PM) cell should be sent per second. The first bullet may then be identified by observing the flow of PM cells.

The second bullet may be more explicit given as a threshold of X % SECBs during a second. ITU-T has suggested the following values:

I CLR > 1/1024, or

II SECBR > 1/32,

during the one second interval. I is only valid for QoS classes with CLR requirement < 10^{-5}. For QoS classes with unspecified CLR objective for the aggregated cell stream this has to based on the CLP = 0 cell stream only. What to do when cells are received, but not enough cells to estimate CLR and SECBR is yet under study.

The performance at the ATM layer should be monitored by a performance monitoring OAM flow with consecutive numbering of these OAM cells. If such an OAM cell is missing blocks are assumed to be missing and severely errored.

The below specified parameters for semi-permanent connections apply to connection portions delimited by measurement points (MPT, MPI) as defined in [15] and described in Section 3.

A period of unavailable time for one direction of a portion of a B-ISDN connection begins when the SES_{ATM} continues for a period of ten consecutive seconds. These seconds are considered to be unavailable time. A new period of available time for one direction of a portion of B-ISDN connection begins with the first second of a period of ten consecutive seconds, none of which are SES_{ATM}. These seconds are considered to be available time. A portion of a bi-directional B-ISDN connection is available if both directions are available.

The *asymptotic unavailability* (ETSI) for a B-ISDN semi-permanent connection portion is the long term probability that this portion is in an unavailability state at any given time. Of course, this value must be estimated by measuring the ratio of the total accumulated unavailable time divided by the length of the observation period for a finite observation period.

In ITU-T the *service availability* ratio is defined as the proportion of time the connection portion is in the available state as seen from the user. Interruptions not seen by the user because he is not sending cells will thus not effect this parameter, and the observation period is delimited to time intervals when the user is transmitting cells. A problem with this definition is the possibility that the user is not sending cells for ten consecutive seconds. This problem is yet to be solved.

The network availability ratio is the proportion of time the connection is in the availability state as seen by the network. Interruptions will thus effect this parameter and it should be the same parameter as defined by ETSI (i.e. if the criteria are the same).

The *outage intensity* for a B-ISDN semi-permanent connection portion is defined (ETSI) as the limit of the mean number of outages in an interval of length *dt*, divided by *dt*, as *dt* tends to 0. Of course, this intensity is estimated by measuring the number of unavailable periods during an observation period (presumably ‘long’) divided by the length of the observation period.

The proposal for worst case availability objective could be 99 – 99.5 % for a given connection portion (defined between two standardised measurement points). The observation period is under study in ITU.

The *mean time between outages* (MTBO) of an ATM connection is the long term average duration of any continuous interval during which the connection is available. A proposal for this value could be in the range 800 – 1500 hours, as a minimum for a connection portion. The observation period is under study in ITU.

ITU-T defines service MTBO by concatenating consecutive intervals of available time during which the user attempts to transmit cells.

The *retainability of a semi-permanent connection portion R(t)* is the probability that the connection portion which was available at a time 0 will not become unavailable in the time interval (0,*t*). The retainability is given as $R = e^{-\lambda t}$ where λ is the outage intensity.

For a switched connection the availability / outage intensity / MTBO parameters are substituted by the *cut-off rate* [5]. A cut-off event corresponds to either

- The occurrence of 10 consecutive SES_{ATM}, or
- A premature release (i.e. a release due to a network node that is not the correct result of a user request).

The cut-off rate for a B-ISDN switched connection portion is defined as the limit of the probability of having a cut-off within a time interval *dt*, divided by *dt*, as *dt* tends to 0. This intensity is estimated by measuring the number of cut-offs observed on a set of ATM switched connections of this type, divided by the accumulated operation time of these connections.

The *retainability of a switched connection portion R(t)* is the probability that the connection portion which was not cut off at a time 0 will not be cut off in the time interval (0,*t*). The retainability is given as $R = e^{-\lambda t}$ where λ is the cut-off rate.

9 Measurements for assessing and controlling network performance

Within the EURESCOM projects [6] and [7] and the Performance Working Group of the European ATM Pilot Network [8] more detailed specifications of measurements have been worked out. This is presented as an Abstract Measurement Suite (AMS) in an Annex to [7]. The control plane measurements have been specified within the P515 project, whereas most of the specifications of the user plane measurements have been accomplished within PWG. The specified user plane measurements are

- Source and Destination Traffic Evaluation
- CDV and PCR along a connection
- Acquisition of Cell Arrival Times
- Real Time measurement of CDV
- Neighbouring Node and Link availability
- Long Term Network Availability and Performance
- Mean Cell Transfer Delay.

User plane measurements based on these specifications have been executed within [8] and are now also used within the ACTS project JAMES (Joint ATM Experiment on European Services). They are intended for operational purposes as well as traffic characterisation for acquisition of source traffic descriptors and network performance characterisation. In [26] some results of long term network availability and performance measurements within the pan-European ATM network are given. Out-of-service measurements have been performed with 1 Mbit/s CBR test traffic in loop-back and using 6 different configurations. These are international configurations comprising 4 – 7 cross-connects over a distance ranging from 2,000 km to 6,000 km. The test durations are approximately 2 weeks each. The measurements indicate

- CER ~ 10^{-9} (with one extreme exception ~ 10^{-4} – 10^{-5})
- CLR ~ 10^{-7} – 10^{-9}
- SECBR ~ 10^{-4} – 10^{-7}
- User plane availability 97 – 99.9 %.

To gain better estimates more tests have to be run and for longer time periods. The background traffic is rather low in these cases and the performance should thus be in accordance with the best QoS class. On the other hand, these tests were run on a pan-European ATM network. Unstable network equipment and lack of experience running huge ATM networks may explain part of the performance degradation. The availability will improve with the introduction of better management systems and implementation of automatic rerouting procedures in case of failure. It is thus to be expected that performance will improve in the near future. The first opportunity to check this will be the network performance campaigns that will run on the JAMES network.

The specified control plane measurements are:

- Connection Set-up Delay
- Connection Release Delay
- CPE response time
- Connection Rejection Probability
- Connection duration time
- Add Party Delay
- Drop Party Delay
- Add Party Rejection Probability
- Connection Modification Delay
- Connection Modification Rejection Probability
- Look Ahead Response Delay
- Minimum Traffic Descriptor Rejection Probability
- Alternative Traffic Descriptor Rejection Probability.

Based on these specifications control plane measurements will take place within the ACTS project JAMES whenever signalling is implemented in the network nodes. Probably measurements can start spring 1997.

This Abstract Measurement Suite (AMS) comprises Measurement Purpose definitions and procedural descriptions of Measurements to monitor/evaluate performance of a hypothetical pan-European VC-switched ATM Pilot Network.

The adopted format allows for accurate definitions of measurements, clearly indicating which functional elements are necessary for implementation without undue implications on surrounding architecture. This approach thus supports implementation within network elements as well as in specialised external measurement equipment.

10 Conclusions

In this paper we have described some important aspects of network performance in B-ISDN networks. The paper mainly covers definitions of parameters to assess user plane and control plane performance and methodology to assess these parameters. Within the EURESCOM projects [6] and [7] and the Performance Working Group of the European ATM Pilot Network [8] more detailed specifications of measurements have been worked out. User plane measurements based on these specifications have been executed within [8] and are now also used within the ACTS project JAMES (Joint ATM Experiment on European Services). In JAMES there are also plans for doing control plane measurements based on the specifications above.

References

1 ATM Forum. *ATM Forum traffic management specification version 4.0.* ??, February 1996.

2 ETSI DE/NA4-42104. *Network performance objectives for circuit switched connection processing delay in an ISDN.*

3 ETSI draft DE/NA-42122. *Accessibility performance for B-ISDN point-to-point switched connections.* ??, September 1996.

4 ETSI draft DE/NA-042129. *Broadband Integrated Services Digital Network : availability and retainability performance for B-ISDN semi-permanent connections.* ??, September 1995.

5 ETSI DE/NA-42123. *Retainability performance for B-ISDN point-to-point ATM switched connection.* ??, September 1996.

6 EURESCOM P410 Deliverable 3. *Practical guidelines for end-to-end test campaigns over ATM networks.* Heidelberg, May 1996.

7 EURESCOM P515 Deliverable D5 Part 1. *Network performance.* Heidelberg, June 1996.

8 European ATM Pilot Network. *Performance working group final report.* 1995.

9 ITU-T Rec. E.600. *Terms and definitions of traffic engineering.* Geneva, March 1993.

10 ITU-T draft Rec. E.72x. *Network grade of service parameters and target values for B-ISDN.* COM 2-R 29, Geneva, December 1994.

11 ITU-T Rec. E.800. *Terms and definitions related to quality of service and network performance including dependability.* Geneva, 1989.

12 ITU-T Rec. I.350. *General aspects of quality of service and network performance in digital networks including ISDN.* Geneva, March 1993.

13 ITU-T draft Rec. I.35bcp. *Call processing performance for a B-ISDN.* COM 13-R 73, Geneva, May 1996.

14 ITU-T Rec. I.352. *Network performance objectives for connection processing delays in an ISDN.* Geneva, March 1993.

15 ITU-T Rec. I.353. *Reference events for defining ISDN and B-ISDN performance parameters.* Com 13-R 57, Geneva, July 1995.

16 ITU-T Rec. I.355. *ISDN 54 kbit/s connection type availability performance.* Geneva, March 1994.

17 ITU-T Rec. I.356. *B-ISDN ATM layer cell transfer performance.* Com 13-R ??, Geneva, May 1996.

18 ITU-T Rec. I.357. *B-ISDN semi-permanent connection availability.* Com 13-R ??, Geneva, May 1996.

19 ITU-T Rec. I.371. *Traffic control and congestion control in B-ISDN.* Com 13-R ??, Geneva, May 1996.

20 ITU-T Rec. Q.2724. *B-ISDN user part : look-ahead without state change for the NNI.* Com 11-R 25, Geneva, December 1993.

21 ITU-T Rec. Q.2725. *B-ISDN user part : modification procedures.* Com 11-R 27, Geneva, December 1993.

22 ITU-T Rec. Q.2762. *B-ISDN general functions of messages and signals of B-ISDN user part (B-ISUP) of signalling system No. 7 for resolution No. 1.* Com 11-R 75, Geneva, September 1994.

23 ITU-T Rec. Q.2931. *B-ISDN DSS2 : user-network interface layer 3 specification for basic call/connection control.* Com 11-R 78, Geneva, September 1994.

24 ITU-T Rec. Q.2962. *B-ISDN DSS2 : Connection characteristics negotiation.* Com 11-R 167, Geneva, March 1996.

25 ITU-T Rec. Q.2964. *B-ISDN DSS2 : basic look-ahead.* Com 11-R 35, Geneva, November 1993.

26 Louvion, J R, Piller, B. Performance measurements and traffic characterisation on the ATM pilot network. *European Transaction on Telecommunication,* 7, (5), 1996.

EURESCOM and QoS/NP related projects

BY HENK GROEN, AMARDEO SARMA, TOR JANSEN AND OLA ESPVIK

1 Introduction

The European telecommunication industry, in particular the Public Network Operators (PNOs), have an important role to play in shaping the future Information Society in providing the necessary physical infrastructure, the European Information Infrastructure (EII).

None of the existing telecommunications networks are able to fully support the whole range of EII services when taking into account all aspects of Quality of Service (QoS).

A commonly accepted principle governing the establishment of the EII is that the EII will be a seamless federation of interconnected, interoperable telecommunication networks, information processing and storage equipment and terminals. The EURESCOM model of the EII is described in Figure 1 and the main areas of concern of the EII as a whole are summarised in Figure 2. The Information and Communication Industry sector players will address the EII from their business perspective. The areas considered to be of major concern to the PNOs are coloured blue in Figure 2.

However, with increased competition and liberalisation of the telecommunications market, there is concern about how the EII is going to be financed. And even more important: With increasing costs and risks, and scarce human resources, how should the R&D and standardisation to implement the EII be organised?

These are some of the challenges the European PNOs are facing.

2 Structure of activities

With the world of telecommunications evolving at an unprecedented rate and R&D resources becoming scarce compared to the needs, the Public Network Operators (PNOs) saw the need to form an organisation for co-operating at a technical level, while they could at the same time compete in the market.

EURESCOM was founded in 1991 by 20 Public Network Operators (PNOs) from 16 European countries. In 1996 it had 22 active Shareholders from 21 European countries. The Headquarters is located in Heidelberg, Germany.

The EURESCOM objectives are:

- To enable the development of harmonised strategies via strategic studies
- To specify harmonised telecommunications networks and services
- To stimulate and carry out pre-normative and pre-competitive R&D projects
- To stimulate and technically support field trials to be carried out by PNOs
- To contribute to European and worldwide standardisation.

The EURESCOM work consists of Projects, and the characteristic and unique way in which EURESCOM works is that the Shareholders themselves perform most of the work. They initiate and carry out the Projects included in the annual EURESCOM Work Programme, and they receive the results of the Projects (see Figure 3). The EURESCOM Permanent Staff (EPS) plays the roles of the catalyst, the pusher, the server and the supervisor.

2.1 The EURESCOM General Framework

The General Framework can be seen as an intended mission statement for EURESCOM. The aims are to:

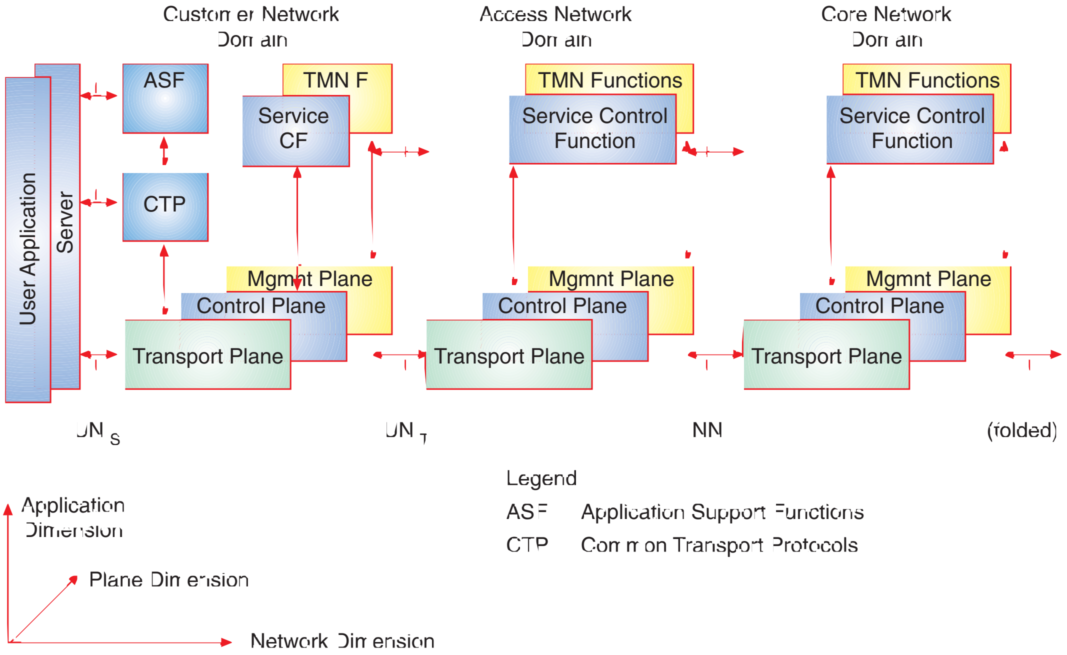


Figure 1 EURESCOM model of the EII

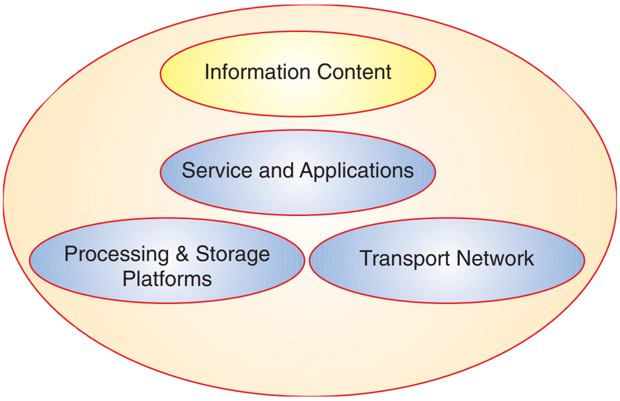


Figure 2 Main areas of concern of the EII

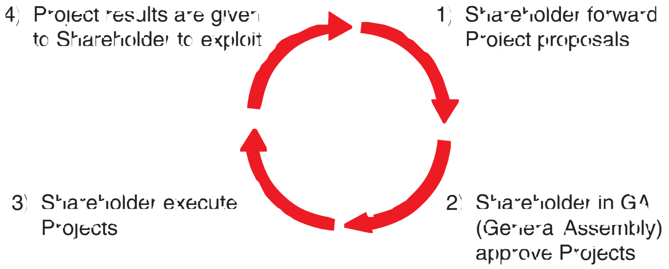

4) Project results are given to Shareholder to exploit
1) Shareholder forward Project proposals
3) Shareholder execute Projects
2) Shareholder in GA (General Assembly) approve Projects

Figure 3 the EURESCOM way of working

- Give a framework for the collaborative technical work found to be necessary in order to obtain common goals
- Give the Shareholders a document which defines the agreed scope of the EURESCOM Work Programmes
- Give the Board a framework for discussion of new work items to be included in the EURESCOM Work Programmes
- Support the EURESCOM Permanent Staff in their discussions with potential proposers on the orientation and necessary coherence of the EURESCOM Work Programmes
- Enable potential proposers to relate their proposals to the scope and the major topics of the EURESCOM Work Programmes
- Give outside bodies a means to relate the work and mission of EURESCOM to their own missions.

Over the coming years EURESCOM's goals will be focused on supporting the Shareholders in specifying, developing and testing the technological platforms for delivering and supporting the EII. Figure 4 depicts the Work Areas.

There is no complete mapping between the areas of concern of Figure 2 and the Work Areas of Figure 4. The Strategic Studies Work Area in particular and some other EURESCOM projects have a wider scope than the EII only.

The above Work Areas should be considered as labels for broadly partitioning and structuring a EURESCOM Work Programme. Some work items will naturally cross the boundaries of these areas.

2.1.1 Strategic Studies

This Work Area should provide one of many inputs that EURESCOM Shareholders will use to determine their own strategic options. It can be used with the Shareholders' agreement to prepare material on behalf of the Shareholders for use with external bodies. The work must be of a broad nature and must be of a high professional standard. One important issue for this Work Area is a strategic analysis of the evolving work on common goals such as the EII and the impact of that on proposals for new work items. The major topics within this Work Area include:

- Market and trend analysis including customer aspects
- Investigations of macro social, economic and environmental aspects of telecommunications
- Strategic aspects of the integration of networks (e.g. fixed, mobile/satellite, CATV, etc.)
- Scenarios for the development of the European Information and Communication Industry.

2.1.2 Telecommunication Services and Applications

Despite the potential competitive nature of Telecommunication Services and Applications it is clear that Shareholders may wish/need to collaborate on the pre-competitive development of some services and applications where such collaboration could lead to access to wider markets, service portability across Europe, etc. It is also an area where collaboration with other players (e.g. independent service providers, users' associations) will be necessary. In a medium term perspective there is more scope for work on advanced services and applications in the ISDN area, the broadband area and the multimedia area. In the same perspective, it can also be envisaged that nearly all transport services are highly commoditized and as such have the potential for greater European integration. The major topics within this Work Area include:

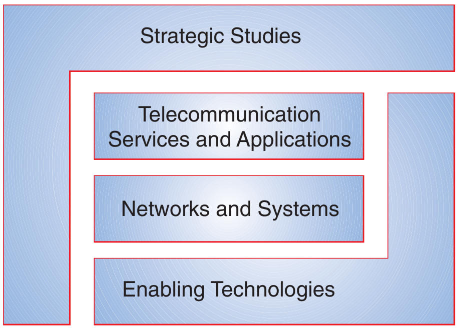


Figure 4 Work Areas of the EURESCOM General Framework

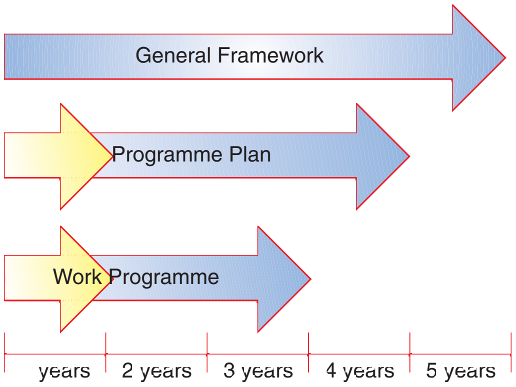


Figure 5 Timeframe of EURESCOM planning documents

- User interfaces and usability of all services
- Generic requirements for new customer equipment
- Multimedia services and applications
- Service management and Quality of Service.

2.1.3 Networks and Systems

This Work Area covers studies associated with overall telecommunication networks and systems issues. How to define the network architecture, the management (including TMN studies at the network and service layers) and control (including N-ISDN, B-ISDN, IN and information networking approaches) of the network are all key issues for this Work Area. In addition, studies on software systems that have end-to-end significance as well as other networked systems such as corporate networks would be included here. These issues certainly have a medium to longer term flavour which, however, has been subject to significant cut backs at some Shareholders and could therefore benefit from collaborative actions. Another reason for more collaborative work in this Work Area is the increasing complexity of future networks and systems. The major topics within this Work Area include:

- New network architectures and interworking
- Specification of network management and network control
- Performance and planning aspects of networks and systems
- Identification and preparation of trials to test new systems configurations and interworking between PNOs.

2.1.3 Enabling Technologies

This Work Area is very broad, spanning from some very short term new techniques or technologies (but not basic technologies) to quite speculative research. A good target may be to include techniques where the Shareholders have to improve their competence (e.g. most aspects of software design). The major topics within this Work Area include:

- Access and optical network technologies
- Systems specification, planning and procurement tools
- Software acquisition and maintenance techniques
- Voice and language processing techniques.

2.2 EURESCOM Programme Plan

Whereas the General Framework is generic and has a longer term perspective the EURESCOM Programme Plan gives an outline of what could be addressed by work items in the coming year(s). The EURESCOM Programme Plan is derived from the General Framework and contains an assessment of the existing and planned projects by means of the EURESCOM model of the EII outlined in the General Framework.

Figure 5 depicts how the three planning documents of EURESCOM complement each other:

- General Framework setting the perspectives
- Programme Plan outlining what could be addressed in the next year(s)
- Work Programme defining the work items that will be embarked upon.

2.3 Relations to other organisations

EURESCOM continues to maintain the best possible contacts with all other relevant organisations.

ETNO, ETIS

ETNO has been set up as the general policy Association for European operators. It will not carry out detailed work on subjects already covered by other operator bodies such as EURESCOM and ETIS. However, the policy decisions taken by ETNO may have an influence on the activities of other groups in which operators are involved and its positions may produce effects on the decisions to be taken by the EC Commission. EURESCOM and ETNO have established a smooth co-operation. A number of the Working Groups of ETNO are now having direct interaction with related EURESCOM Projects.

The public network operators in Europe also have an organisation to co-ordinate and co-operate the development of information systems, namely ETIS (European Telecommunications Informatics Services). Both organisations, EURESCOM and ETIS, work together on specifying common support systems for the common owners. With the continuing convergence of telecommunications and information technology there will be the need for closer relations between these organisations.

Standards Development Organisations

EURESCOM has over the years continued to interact with ETSI (the European Telecommunications Standards Institute) where it is an ordinary member. The intention of the Permanent Staff of both ETSI and EURESCOM is to co-ordinate their work programmes and benefit from each others' activities

EURESCOM follows the relevant work within the ITU (International Telecommunication Union) through the experts involved in the individual projects.

Fora and other organisations

EURESCOM is following what happens in the different new semi-standardisation organisations, called fora, such as ATM Forum, Network Management Forum, DAVIC, X-Open and OSF.

EURESCOM has been a member of the TINA (Telecommunication Information Networking Architecture) Consortium for several years.

Regular contacts with the user organisation INTUG/ECTUA have been established and information from INTUG/ECTUA is fed into relevant projects.

European Commission

Good relations have been established with the EU programme RACE and with the ACTS programme. In the future it will be even more important to have good co-operation between EURESCOM projects and EU programmes.

3 QoS aspects

The ultimate objective of all telecommunication activities within the organisation of a PNO/service provider is to produce satisfactory services to the users at a reasonable price. The essential aspect of the global evaluation of a service is the opinion of the users of the service. The result of this evaluation expresses the users' degree of satisfaction. ITU-E800 is one example of a framework for:

1 The Quality of Service concept

2 Relating Quality of Service and network performance

3 A set of performance measures.

In ITU terminology Quality of Service (QoS) is defined as "the collective effect of service performance which determine the degree of satisfaction of a user of the service". It is obvious that a service can be used only if it is provided, and it is desirable that the provider has a detailed knowledge about the quality of the offered service. From the provider's viewpoint, network performance is a concept by which network characteristics can be defined, measured and controlled to achieve a satisfactory level of service quality. It is up to the Service Provider to combine different network performance parameters in such a way that the economic requirements of the Service Provider as well as the satisfaction of the User are both fulfilled. The details of ITU-E800 is further explained in another article of this issue.

Quality issues affect the whole design and operation of networks and services and can be found in a wide range of EURESCOM projects. From the EURESCOM point of view it is very important that the quality measures should be done in a standardised way to ensure a harmonised understanding of service and its (agreed upon) quality delivered to the user. In a competitive world this gives the user a possibility for benchmarking the PNOs/Service providers, the latter having the same opportunity with their suppliers together with an unambiguous way of root cause tracing when quality is below an acceptable level. From a EURESCOM point of view it is already obvious that projects addressing design topics of networks and services are meaningless unless the QoS/NP aspects are included.

4 QoS related projects

From the very start of EURESCOM work programs QoS/NP aspects were advised to be taken into the Work program. Since then, these issues have been addressed in the following projects:

P227	Software quality assurance
P302	European Switched VC-based ATM Network Services
S26	Reliability Engineering for Future Telecommunication Networks and Services
P307	Reliability Engineering
P404	Broadcast Services for Residential Users
P412	Methodology and Tools for ISDN Network Integration and Traffic Route Testing
P460	European Multimedia Experiments in an ATM Environment (EMMA)
P504	Euresmail – a harmonised Europe wide E-mail service
P508	Evolution, migration paths and interworking with TINA
P513	Enhancements of SDH networks
P515	European switched VC based ATM network studies – stage 2
P603	Quality of Service: measurement method selection
P611	Overall Strategic Studies – IV
P616	Enhanced ATM implementation issues
P619	PNO – Suppliers Technical Interfaces
P702	Internet Protocol version 6 – new opportunities for the European PNOs
P712	Intelligent and mobile agents and their applicability to service and network management

A brief introduction to each of the projects mentioned follows:

P227 Software Quality Assurance

The Project P227 "Software Quality Assurance" defined a common set of quality procedures and indicators for the PNOs and their suppliers, methods to measure software quality indicators and a framework for technical procurement, especially for software procurement. The focus on software of telecommunications equipment for switching and transmission is vital, because it reflects the core business of PNOs.

The results produced by the Project are:

- A survey on PNOs and supplier procedures and on the state-of-the-art in the field, which has been guided by the ISO Standard 12207. It identified possible shortcomings of PNOs, such as inadequate monitoring of the supplier processes and, in general, insufficient emphasis on quality. Some recommendations are provided for tools to be used by PNOs.
- A draft recommendation of quality assurance activities (QAAs) issued in order to collect feedback within the project partner organisations and from suppliers. It lists a total of about 50 QAAs, some adopted from ISO standards, while others merely refer to them. Quality aspects to be measured are: maintainability, reliability, cost estimation, schedule (and on-time delivery) evaluation and various process aspects.
- Recommendations for Quality Assurance of Telecommunication Software, addressing the PNOs' quality departments responsible for setting up quality management systems. Procurement departments are also addressed as there are some recommendations concerning contracts.
- Technologies for Quality Assurance of Telecommunication Software, which contain the technical details and are intended for domain experts who have to perform the quality measurements.
- Recommended Practices for procurement of telecommunication Software, supplying guidelines on how to implement the technical procurement

process for software products. The guidelines are based on the ISO life-cycle model, on the EEC Directives 93/38 on open procurement and on other relevant sources. The processes covered relate to System Requirements Specification, Supplier Qualification, Contract Preparation, Supplier Surveillance and Co-operation and Acceptance Test and First Office Application. This document guarantees that state-of-the-art methods (stand 1994) are used if the recommendations are implemented. The guidelines were preceded by the development of a general model for the procurement process of systems for telecommunication applications. This model was then specialised for software and turned into guidelines. The guidelines identify a set of steps and activities to be carried out and describe information that must be exchanged between an acquirer and a supplier. References are made to relevant standards and rules. A checklist, containing all the steps of the procurement process, is also appended to the report. This result can also be seen as educational.

P302 European Switched VC-based ATM Network Services

The P302 Project followed EURESCOM Project P105, which successfully specified a first Europe-wide ATM (VP-based) network. This specification was used by a number of European PNOs for the well-known European ATM Pilot Network, interconnecting individual ATM networks. Project P302 extended the specifications to cover VC based ATM networks.

The project took as its basis all the available standards such as for transmission, ATM layer, AAL, signalling, etc., and added specifications for interworking with other networks, network performance, network management, traffic control, resource allocation, functional specification of VC switches, architecture and implementation scenarios as well as functions for the support of connection less services.

Signalling and its approval by ITU (and the ATM Forum) was always a critical factor for the project. As previously, it was assumed that during 1994, ITU would conclude on the so-called Capability Set 2 Step 1 set of recommendation, and therefore the project was planned in two phases (the second phase now being operated as project P515). In reality, approval of this signalling standards was delayed until late spring 1995. Therefore, the following specifications are basically assuming functionality supported by CS1 and only in few cases where CS 2 Step1 was stable at the time of study its functions were used. The results of the Project include quite mature specifications and in selected areas substantial input for contributions to standardisation bodies.

For a number of different services (e.g. multimedia services), features like customer control of the call and fast set-up are essential. This implies the introduction of signalling in the ATM network, the interconnection of users by means of switched virtual channel connections and the introduction in the network of VC switches. The results include a relation of the basic assumptions that have been considered for the first on demand Pan-European VC switched network. The areas of study cover network architecture, signalling, management, resource allocation, performance, routing and addressing, network elements description, interworking and connectionless service.

S26 Reliability Engineering for Future Telecommunication Networks and Services

The objective of EU-S26 was to identify possible faults and failures that threaten the reliability of future telecommunication networks and services, to review possible counter measures to cope with faults and to indicate directions of further research within EURESCOM to deal with these threats.

Reliability Engineering has traditionally been the technological discipline to develop, apply and maintain systems in order to achieve dependability. The dependability of a system is that property of a system that allows reliance to be placed justifiably on the service it delivers. In other words, a crucial aspect of Network Performance and finally QoS to the users.

In the process of telecommunication service provisioning, three major components can be distinguished: the physical network for data transport, service and network management, and service implementation and execution. The EU-S26 project was carried out in three phases. First, the major reliability threats to each of the above mentioned three major components of the infrastructure was identified. Second, a study was conducted to identify methods to procure and validate the dependability of future networks and services. Finally, special attention was paid to fault estimation and the application of such methods to control the quality of service (QoS). Directions for further research was a major part of the S26 investigation.

P307 Reliability Engineering

On the background of the work having been done in S26 P307 set out to

- Establish an overall functional understanding of the telecommunication network as a basis for dependability analysis, relating the dependability of the network functions and network elements to the end-user's QoS. Anticipating a future with a variety of hitherto unknown services, network- and technological solutions, an import. ant part of the work was devoted to establishing a service-independent description of an end-user service failure as well as a root-cause analysis method of service failures – together with a plan of important issues for further study.
- Gain a deeper understanding of the increasing influence that software has upon telecommunication systems and network dependability.
- Provide guidelines for PNOs to prevent and handle Signalling System #7 failures in order to prevent severe incidents.
- Establish a European Reliability and Quality Measurement System (E-RQMS) and European In Process Quality Measurement (E-IPQM) that are based on RQMS and IPQM as developed by Bellcore in the USA. A very successful result of this work was the formation of the EIRUS (European IPQM an RQMS Users) organisation supported by all major PNOs and Telecommunication suppliers in Europe. The EIRUS organisation is further described in another article of this issue.

P404 Broadband Services for Residential Users

P404 has been carried out in three phases. The first phase focused on service selection and service usability. This resulted in a report identifying broadband services for residential users and investigating future demand for these services. The conclusion from this work was that there are three classes of services that are most promising: videophony, tele-games, and tele-shopping. The second phase of P404 focused on the study on network, terminal, and server functionality needed to support the services mentioned above. In addition, some target values for specific QoS parameters were given.

The third phase has been an experimental phase in which example implementations of the selected services were studied in order to obtain insight into the kind of traffic they generate, as well as to verify and elaborate on the QoS parameters identified in the second phase of the project. With respect to the traffic characteristics of the services, the approach taken was to derive traffic models for the example service implementations based on experimental usage of the services. These models were validated by means of simulations and where then used for network dimensioning by means of network simulation packages. Such a methodology for network dimensioning was followed in which the combination of (limited) experimentation, modelling, and simulation form the main ingredients, while being convinced that the use of these kinds of techniques is essential for the controlled development of service networks that have to guarantee specific QoS parameters. This is opposed to the rather ad-hoc development of the Internet. Given the wide variety of access network technologies used, it was decided to limit the scope of the dimensioning to the backbone ATM network.

P412 Methodology and Tools for ISDN Network Integration and Traffic Route Testing

The introduction of present and future EURO ISDN services will require the PNOs to perform “Network Integration Testing” (NIT). This means to perform testing activities applied to the global network (Euro Network) that is made up by integrating the national networks of the different PNOs.

The “Euro Network” behaviour must be tested and monitored using a “not-only-domestic” approach and techniques. It will have to be checked, for example, that the bearer and supplementary ISDN services, as implemented in the national networks, are actually capable of inter-

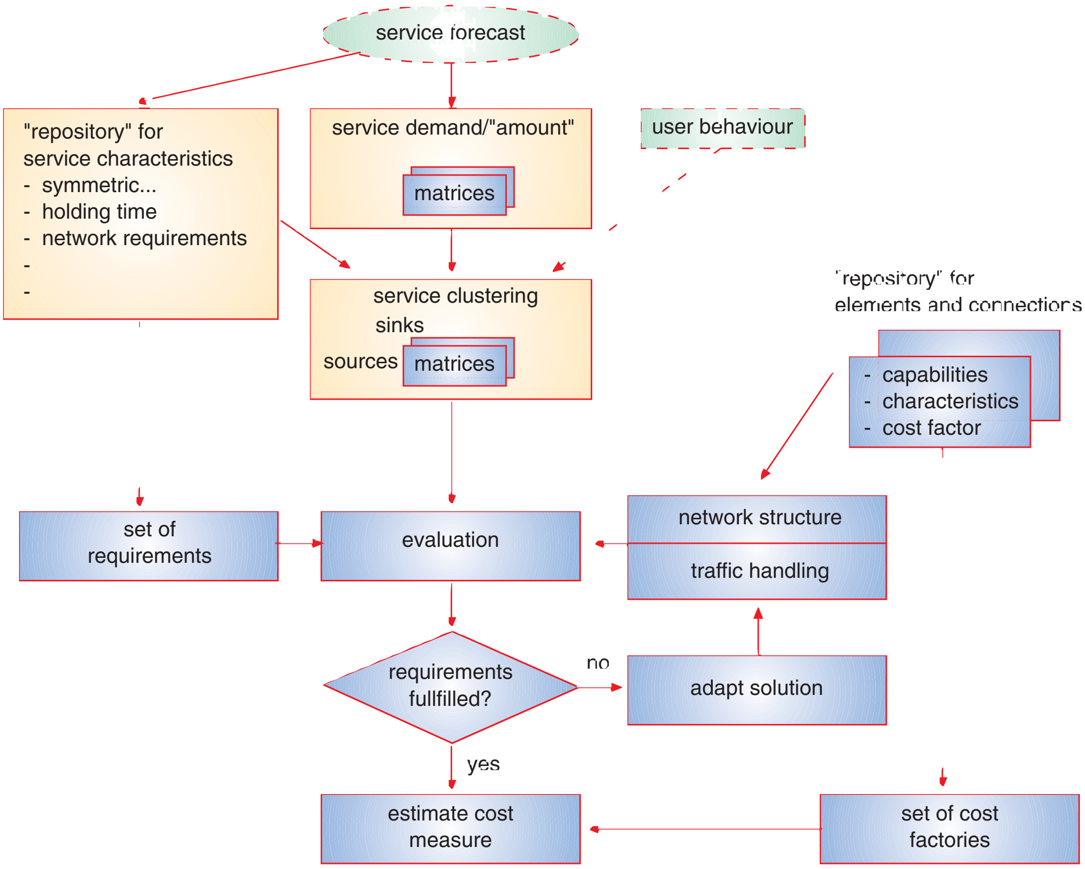


Figure 6 The network dimensioning process, illustration from P404

working in the global network, are compatible, and are provided to customers in a homogeneous way: In case of problems efficient and reliable methods & procedures for investigating the reasons of the possible failures must be available.

The QoS related result provides abstract test specifications for ISDN Network Integration for both end-to-end and node-to-node testing.

The results are interesting for people of ISDN test labs, in charge of setting up new links or monitoring existing international links and for people responsible for the acquisition of testing equipment. The results are publicly available, and should be interesting for test equipment manufacturers and network operators outside Europe who want to establish ISDN links with European PNOs.

Traffic route testing enables PNOs to measure QoS/NP for national and/or international traffic routes. This could either be carried out manually by the network operator or automatically using a traffic route test system. Such a system would generate random traffic similar to that of the subscribers and record QoS/NP for each call. This information could be used to estimate the end-to-end network performance on a number of traffic routes.

Current traffic route test systems can only test national PSTN/ISDN traffic routes. The aim of this project is to specify how these individual systems could be interconnected to provide an *international* traffic route tester.

For a national traffic route tester the call pattern is generated by a single central unit which informs the end units when to set up / clear down calls. For an International traffic route tester this functionality would need to be implemented by a common International Central Unit (ICU) which generates a call matrix which is transferred to each National Central Unit (NCU). The NCU would then inform the end units to set up / clear down international calls.

The specification for an International traffic route tester is broken down into a number of different areas which are detailed below:

- Generic network configuration for International Traffic Route testing
- A list of ISDN bearer and supplementary services to be tested
- The set of QoS/NP parameters to be measured (consistent with existing ITU standards)
- The format of the network topology database at the ICU
- The format of the call matrix generated by the ICU
- The process used for generating the test calls, including synchronisation of system components and scheduling of test calls
- The file format for information transferred between the various system components
- A description of how to carry out an error trace function on an International Traffic Route
- Physical interfaces and protocols used to transfer information between the ICU and the NCUs
- The set of technical standards to which the traffic route tester MUST conform.

This result is aimed at network operators who may already perform traffic route testing for national ISDN traffic routes but wish to extend this to international routes.

P460 European Multimedia Experiments in an ATM Environment (EMMA)

The project took the following engineering approach: In the first step of the project multimedia capable workstations have been installed and interconnected to the national ATM network at the premises of all partners. In the second step a multimedia conferencing tool supporting collaborative work by sharing of applications and telepointing has been implemented at all sites. The resulting tests and experiments yielded results with respect to the relationship of the traffic and the perceived QoS of audio/video communication and interactivity. Finally, in the last step of the project additional multimedia applications creating different traffic characteristics have been tested and measured within several subgroups of the EMMA consortium.

European-wide experiments were used as a testbed for the evaluation of the suitability of the European ATM Pilot Network to carry multimedia applications/teleservices, and to investigate traffic models. The work concentrated on measurements and evaluation at the end-user level.

The following issues were investigated:

- Accessible points of observation within the end-system and within the local ATM switch parameters to be monitored and measured

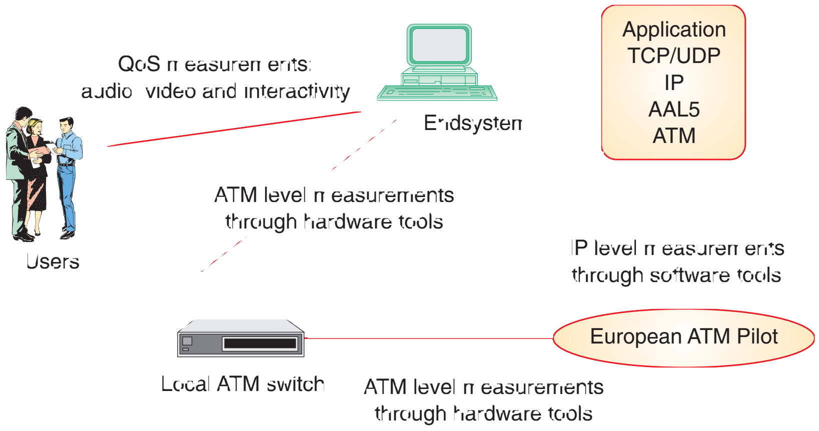


Figure 7 Identified accessible points in the P460 experimental set-up

- Special hardware equipment or software tools, suitable to perform measurements at the communication levels selected for measurements
- Procedures for gathering traffic data.

During all the experiments, personal evaluations were registered, from a user point of view, to make possible a subjective evaluation and a general overview of the facilities of the applications.

P504 Euresmail – a harmonised Europe wide E-mail service

P504 produced at first a *survey* of the E-messaging situation; and then, taking the *user requirements* into account, specified *Euresmail,* based on simple and concrete solutions based on existing systems in place, integrating the two E-mail key players, namely, Internet and X400, both from a *technical and organisational point of view*.

The specification is based on the following functional and quality/performance *requirements:*

Provided messaging services

Efficiency of gateways (interconnectivity/compatibility)

Notification aspects (delay, meaning)

Diversity of addressing scheme

Need for priority levels and non-probabilistic delay

Need for directory

Conversion capabilities

Automatic handling/processing of attached files

Message size limitation

Message traceability

Support of national character sets

Full compliance with standard functions

End-user related issues

Need of an alert when a message is deposited in the mailbox

Ergonomy of user interface

Security features

Direct access to the E-mail system from the application

Network and service access

Access from various locations (ubiquity)

Transmission speed

Easy dial up access.

P508 Evolution, migration paths and interworking with TINA

The TINA Consortium is progressing towards the specification of a distributed processing platform for the telecom industry and an object-oriented service and management architecture for the provisioning of different classes of services.

Range of available speeds per main service, illustration from P504

Main Services	Available speeds	
ADMD X400	9.6 – 64 kbps	X25, ISDN
	up to 10 Mbps	LAN connection
	64 kbps	
	2 Mbps	ISO/IP and TCP/IP
	9.6 to 256 kbps	
	1.2 to 64 kbps	
	14.4 kbps	dial-up
E-mail + mailbox service + file transfer between user and mbx	1.2 kb – 19.2 (X28)	
	up to 10 Mbps	LAN connection
	300 – 14400 bps	PSTN
	9.6 – 19.2 kbps	ISDN
	max 23 kbps	modem connection
	max 64 kbps	X25
	64 kbps	ISDN
	up to 2 Mbps	
	1.2 – 64 kbps	
Information service access point	300 – 14400 bps	PSTN
	9.6 – 19.2 kbps	ISDN
	9.6 – 19.2 kbps	X25
	up to 10 Mbps	LAN
GSM short message service	300 – 14400 bps	PSTN
	9.6 – 19.2 kbps	ISDN
World-wide access to the Internet	up to 14.4 kbps	
	up to 200 kbps	
	9.6 – 14.4 kbps, V42,MNP2-4 err.detec., V42bis MNP5 data compr.	
	300 – 28.8 kbps	modem
	19.2 + 28.8	async. ll
	19.2 + 64 + 128	sync.ll
	64 – 128	frame relay
	64	ISDN
	9.5 to 512	
	9.5 to 64	
Store and forward fax service + mail manager + info box	not provided	

In the preparation of future operational plans by the PNOs and the options for using TINA concepts and products in their network development, it is of paramount importance to define an evolution strategy for European public networks to progressively include TINA concepts and exploit their expected advantages. It has to be highlighted that the integration of existing legacy systems with TINA technology and services is critical for the successful adoption of TINA.

This deliverable aims at defining a strategy for implementing migration and interworking solutions, i.e. determining how TINA can substitute, extend or be added to the installed base of legacy systems and eventually interoperate with future TINA systems that will be deployed in a Europe-wide scale. TINA solutions have been applied to different network contexts, e.g. IN, B-ISDN, Internet and TMN, in order to support different classes of services, and migration at the service layer has also been addressed.

The initial starting point for the introduction of TINA strongly depends on identifying the commercial needs for improving technological solutions in a particular network infrastructure.

The focus of the work is on developing migration and interworking solutions for applying TINA to existing systems. Detailed specifications of Adaptation and Inter-working Units for implementation of these solutions are contained in the results.

P513 Enhancements of SDH networks

Activities related to the evolution of SDH (Synchronous Digital Hierarchy) networks have been dealt with in several EURESCOM Projects. P107 has given the basis for setting up a European-wide communication network (METRAN) based on the SDH technology.

P409 addressed a number of aspects related to the implementation of PNOs' SDH networks in order to achieve a high performance level for paths going across a European SDH network: network protection architectures, protection interworking, network restoration, network performance aspects, in particular availability calculation for different network architectures, network synchronisation aspects and identification of parameters to be monitored for evaluating the synchronisation quality of a path going across a European SDH network.

A number of study items have been identified by P409 that, due to lack of resources, were to be undertaken by P513; the objectives related to QoS/NP are to identify strategies for an integrated protection of a telecommunications network where a number of different layers with their own protection schemes co-operate to improve service availability and, as a consequence, to provide criteria for network optimisation and planning taking into account these advanced protection strategies.

A second objective of this project was to investigate how the use of radio links (terrestrial or satellite) can influence the QoS/NP in an SDH network. Running measurements on SDH radio links allows verification of whether the present definition of performance parameters, mainly based on the use of optical fibre systems, can also fit the error distribution found in radio systems.

A third objective of the project was to provide guidelines to network operators for planning and managing a synchronisation network for SDH at the pan-European level. These required extensive modelling and simulation activities as well as measurements on the existing network.

P515 European Switched VC based ATM Network Studies – stage 2

The aim of this project was to complete and update the specifications of an implementable switched ATM VC network supporting advanced service features, based on the ITU-T signalling Capability Set 2 step 1. A first set of specifications was prepared by Project P302.

The results of P302 provided guidelines to the introduction of a VC switched network based on CS 2.1 and beyond and thereby address in particular the special requirements that a network must fulfil to provide new services. They can be used by people at PNOs in departments implementing ATM networks involved in planning and dimensioning ATM Networks, specifying switches, introducing Management systems and introducing connectionless services in B-ISDN.

Some of the P515 results are:

- The impact of new services and applications (e.g. broadband virtual private network, and video on demand) together with the introduction of new connecting types (e.g. cross connect VCC and multipoint connections) on the network architecture. This completes the initial work on interworking issues done in P302 and includes a detailed analysis of service interworking and private-public B-ISDN interworking. Also included are specifications of the management related to the Control Plane, especially concerning the VC switch network elements, covering aspects related to point to multipoint and broadband virtual private networks.
- Guidelines for a specification of network elements, for example for combined equipment (VP/VC cross connect and VC switch). Valuable information related to the characteristics of the ATM equipment has been compiled. This information was obtained from replies to a questionnaire sent to a wide set of relevant manufacturers.
- Network performance requirements and the requirements on network elements derived from these network requirements. These incorporate the specification of the measurements to assess and control network performance, which includes identifying requirements on network elements derived from the need of built-in measurement facilities. Concerning resource allocation, the measurements needed to assess and control the relevant network performance parameters, both in the User, Control and Management planes have been specified. Aspects such as QoS classes, use of CLP bit, managing of CDVT within the network and studies about the ABR support within the P515 scope have been also analysed.

P603 Quality of Service: measurement method selection

From a user point of view it is considered important that the various PNOs/Service Providers present the quality aspects of their services in a standardised and understandable way making sure that the quality measurements are also made in a standardised way. This is the very foundation for enabling the users to compare the various services and corresponding prices offered.

Communicating very closely with the user organisation INTUG, P603 has started a process towards:

- Proposing a standardised measurement methods for the quality of Europe-wide voice services offered over different networks: ISDN, POTN, GSM, ATM, ...
- Producing a range of figures for QoS parameters for these services as far as the measurements allow this.

The results of the project should be used by PNOs to develop quality control and measurement methods in telecommunication services, especially for European-wide voice services, which could facilitate discussions with customers on QoS.

P611 Strategic Studies IV

The Information and Communication industry viewed in its broad concept is currently in a period of major change and uncertainty which will accelerate over the next decade. The industry's character and parameters at the end of this period are expected to bear only a modest resemblance to the present situation. Changes over all elements of the industry will be driven by myriad forces mainly arising from the increasing competition in the market place, the technological advances in electronics, information technology and telecommunications, the explosion in telecommunications applications possibilities, the convergence of previously separated media, and the potential for information technology and telecommunications to deeply penetrate and influence all elements of society.

The Shareholders of EURESCOM need to continually assess how these changes will come about and how they will impact on their businesses, if they are to survive and prosper in this new demanding and threatening environment. Understanding these changes and their impact on the Shareholders is an important part of EURESCOM's Strategic Studies in 611, so that EURESCOM Shareholders are effectively briefed and supported in their own strategic assessment activities on telecommunications related to the various elements of Society like:

- The existing situation including world-wide state-of-the-art and the activities related to the European programmes, standardisation related activities and issues, implementation related projects
- Opportunities for substitution or replacement of existing situations by means of telecommunications, dealing with such matters as market figures, segment size and revenue potential
- Perceived future development in the particular element of society, how it will change in the future, and the significant factors which will drive such changes
- How telecommunications will be influenced by changes in the elements of society and how telecommunications could influence these changes
- The synergy between telecommunications developments and the developments within the particular element of society
- Technological and conceptual aspects
- Telecommunications applications definition needs
- Standardisation needs
- Existing and future relevant projects and activities in the area
- Potential for EURESCOM and PNO actions.

The main objectives of P611 is to

- Undertake a number of short strategic studies on selected topics of mutual interest to the EURESCOM Shareholders
- Maintain and progress a longer vision and scenarios for the development of European Telecommunications
- Provide a study on the post-1998 R&D options for the European PNOs
- Provide inputs to the definition of future EURESCOM Work Programmes; give a better understanding of the advantages and potential of co-operation among the PNOs and how to develop such co-operation in a practical way.

In this context we find Task 3.11, Quality of Service and Network Performance, the objectives of which are:

- To identify goals for strategic work on QoS in EURESCOM
- Outline project(s) and project co-ordination to reach the goals.

P616 Enhanced ATM implementation issues

Project P616, which started in November 1996 and will be completed by February 1998, aims to contribute to the progress and stabilisation of the definition of an advanced ATM VC switched network. P616 addresses specific network architecture aspects, congestion control in a VC switched network, charging and accounting parameters in a VC switched network as well as Transport and Network layer protocols over ATM. All these areas have still open points that are inadequately defined in standards.

There are four distinct areas of work in P616: "Network architecture aspects", "Congestion control", "Charging and accounting parameters" and "Transport and Network layer protocols over ATM".

In "Network architecture aspects", a high demanding service to run on ATM will be first identified. Such a service, for example a broadband video conference, will place stringent network requirements that need to be solved. User application requirements will be determined in terms of bandwidth, and QoS for the new broadband service will be identified. Specific network architectural issues will be dealt with, such as multipoint aspects for the support of services and the implementation of the lower layers of the management plane.

In "Congestion control", the main objective is the provision of a general congestion avoidance strategy for an ATM switched network. The problems of congestion detection and the applicability of the techniques used in classical networks at the call layer to ATM networks will be central for this item. Finally, network dimensioning principles will be established.

In the "Charging and accounting parameters" area, charging and accounting parameters in a switched ATM network will be identified. Methods for parameter measurements will be determined and the application of the parameters to the different ATM transfer capabilities will be established. Requirements on the network elements for the implementation of the charging and accounting capabilities will be determined.

In "Transport and network layer protocols over ATM", the most relevant scenarios for the ATM/IP convergence will be identified, including the issues of Multiprotocol over ATM and Integrated IP-ATM switching. Furthermore, appropriate solutions for the integration of data and real time services in an ATM network will be discussed.

P619 PNO-Suppliers: Technical Interfaces

The QoS experiences by the end-user is the final result of a long chain of quality actions having taken place throughout the whole life cycle of all hardware and software systems that make the service possible. The Service Provider/PNO's interface with the suppliers is manyfold and important. A good technical communication forms the basis on which Quality Assurance and procurement activities are based, in order to achieve company's objectives. The purpose of P619 is to analyse the Technical Interfaces of the Quality Assurance environment, identifying on which levels these interfaces can be improved. The work of P619 is to:

- Define as much as possible detailed harmonised requirements, recommendations, and procedures within specific phases of telecommunications hardware and software product life cycle, to allow the setting up of an efficient and effective set of Technical Interfaces between PNO and Suppliers of telecommunications systems (hardware and software) and services, in accordance with the results identified in P227 and P307.
- Define guidelines to perform Identification of Needs and Requirement Specification, for innovative, off-the shelf or "to be enhanced" products, in accordance to:
  - European Directives (to enhance competition)
  - Standards (e.g. the ETSI concept of Implementation Conformance Statements – ICS – implementing a conformance testing methodology and framework which provides pro forma statements and profile requirements lists related to OSI specifications)
  - Ongoing harmonisation activities (Ephos, Euromethod, ...)

  by guaranteeing the required transparency toward the supplier, and preserving the flexibility required by PNOs in their planning.
- Define and try out a harmonised assessment instrument for telecommunication system (hardware and software) suppliers. As for software supplier, particular care will be devoted to experience made with SPICE (Software Process Improvement and Capability dEtermination).

P702 Internet Protocol version 6 – new opportunities for the European PNOs

The current world-wide Internet networks are based on the OSI-layer 3 protocol IPv4 (Internet Protocol version 4). Drawbacks of these networks are:

- Traffic management not possible (currently more than 66000 networks involved)
- No guaranteed end-to-end bandwidth (Store and Forward transport, CBR services not guaranteed)
- Limited security and limited address space
- Undefined QoS/NP control.

Notwithstanding these drawbacks the usage is increasing and – from a technical point of view – almost all known telecommunication services/applications can be offered over this network:

- E-messaging services
- Telephony services
- Audio services
- Video (still picture and moving picture) services
- Retrieval and conference services.

The next version of IPv4 is on its way (IPv6) and is now being tested, in addition to protocols for the reservation of network resources (RSVP). This version and these protocols will then have solved the following problems of the current version 4:

- "Unlimited" address space (665.570.793.348.866.943.898.599 addresses per square metre of earth surface)
- Security (authentication, integrity and confidentiality)
- Network performance/QoS (traffic priority).

The aim of P702 is to investigate network scenarios and specify building blocks, which combine the best of two worlds: the *Store and Forward IP world* and the *Circuit Switched Telephony world*. In particular, the building blocks defined should be seen to add additional and complementary capability to the existing leased line based Internet infrastructure. The functional architecture and the functions specified should be applicable to any public switched network and should support the offering of the bandwidth on demand to the end-user.

P712 Intelligent and mobile agents and their applicability to service and network management

This project is currently in the set-up phase and is likely to be kicked off in March 1997.

Today, we are moving towards highly customised multimedia and mobile services and to new concepts for effective network and service management. In addition, specific techniques are needed to attain a high degree of software flexibility and dynamic configurability at the network and administrative nodes.

Software agents are programmes that are not confined to a particular network or computing node, but can migrate and replicate themselves in the network. They do this with the aim of finding specific services and applications, thereby checking constraints and observing physical and computing interfaces. Such flexible concepts are very useful for many management procedures, such as for service provisioning that takes user, terminal and network capabilities into account, quality monitoring, or fault management.

The emerging, relatively new concept of "Intelligent Agents" and the supporting technologies, combinations of relatively well understood technologies like "distributed AI or rule-based systems" and "dynamic and interpreted languages" show a potential for enabling such a new generation of management systems. A significant research body already exists. However, what is largely unproved is the applicability of this technology to the telecommunications industry.

The main objective of P712 is to assess the maturity and implications of "Intelligent and Mobile Agents" concepts and technology and their applicability to service and network management.

Results to be expected are:

- Intelligent and Mobile Agent Technology: current state and mid-term evolution
- Prototype of an agent based system
- Results of experimentation.

5 Summary and the future challenges

Telecommunications are becoming an increasingly more important part of society's infrastructure. The impact of a network wide failure (reduction of QoS/NP) will be immense in terms of lost business for end-users, the social impact, and last but not least, the operator (PNO) of the network affected.

The objective of all EURESCOM work has been to harmonise areas of common interest between the Shareholders to make the network and services as efficient as possible. QoS/NP is indeed an important issue, mainly because QoS/NP versus price is the area where the user/customer has a direct opportunity to evaluate the service received and – in the competitive world – make the final decision about what PNO/service provider to choose. Up to 1994 our projects addressing QoS/NP generally dealt exclusively with focused areas of Network Performance – mainly Dependability. From P307 and onwards an active work was started to bring about a common understanding of QoS as presented to the user (P603), but also to understand the importance of having a harmonised set of technical interfaces between the PNOs and their suppliers (P307). The latter work has resulted in a European adaptation of the Bellcore systems RQMS and IPQM now gaining very constructive interest within both the PNOs and the major supplier of telecommunication equipment. A European organisation – a user group for quality measurements (EIRUS) – has been established, and already most of the European PNOs are full or pending members, together with major suppliers. Very important to harmonisation is having quality measurement methods agreed upon together with a well understood terminology. Here the first initiatives have been taken in the ongoing projects P603 and P611, the latter aiming at integrating QoS issues into the EURESCOM strategic studies of the future network and service development.

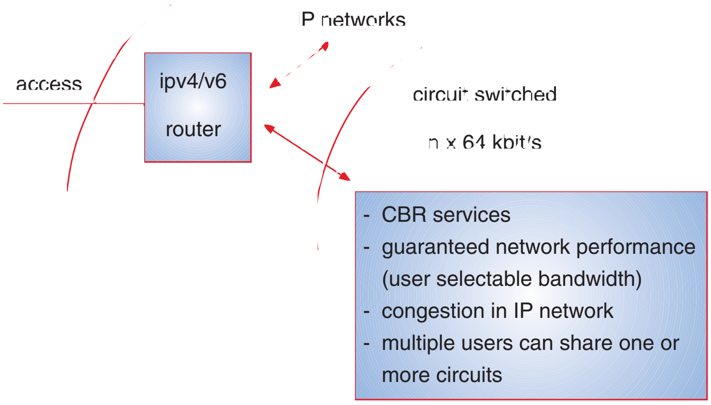

IP networks
access
ipv4/v6
router
circuit switched
n x 64 kbit/s
- CBR services
- guaranteed network performance (user selectable bandwidth)
- congestion in IP network
- multiple users can share one or more circuits

Figure 8 User's choice of network type, an illustration from P702

And major challenges are still to be met. The coming generation of public networks will have a number of new, and unproved, architectural and technological features for service handling, O&M and transport. Mainly because of the increased use of software functions strong and dynamic dependencies across the network will exist, making these new networks vulnerable for failures affecting a large number of their users/subscribers and the network functionality. The theoretical basis, methodology and tools for handling these problems are generally not available but truly a major issue for further study, hopefully resulting in an efficient treatment of unacceptable QoS/NP in the new networks.