Foreword

The work of this thesis began in January 2003 when I made contact with my teaching supervisor; Dr. Terje Jensen with Telenor R&D. He proposed the task of modeling actors in telecommunication on my initiative, which after some discussion soon developed into focusing more on the content and value of service level agreements (SLA).

My background comes mainly from studies at Oslo University College, Department of Electrical and Electronics Engineering, where I finished a smaller thesis in 2002. That thesis was modeled as a handbook for Next Generation Networks (NGN); giving new beginners an introduction to NGN's concepts and purposes.

Continuing my education at the University of Oslo, Department of Informatics in the communication systems branch of study, I now present this thesis as part of my final master's degree. I have filled my studies the last two years with a broad specter of courses, including programming, social aspects of technology and science, qualitative research methods and management. The final master's thesis is a result of 1.5 years hard work, and builds upon my knowledge acquired from writing my thesis at Oslo University College, in addition to newer course studies.

Some of the university courses I have taken since January 2003 have understated my knowledge within the thesis' subjects. Other courses have been a part of my final master's degree (siv.ing.) to broaden my education. The work with these courses has been done in parallel with the thesis work, but the last half year's work has comprised only on answering the thesis problems. The work has been interesting, and I believe my range of knowledge has broadened.

I would in relation to this like to thank my supervisor; Terje Jensen, for valuable comments, contributions and help for the thesis' presentation. His comments have been inspiring as well as motivating in relation to my work. I also owe my internal supervisor; Tor Skeie with the University of Oslo, gratitude for support throughout the working process.

29. June 2004,

With best wishes of pleasant and interesting reading,

Nina Sørsdal

Abstract

Because of high market competition between services in today's multi-provider environment, means to regulate quality of service (QoS) is required to guarantee for providers', users' and other actors' rights when purchasing/selling a service. There are already a multiple of existing telecom services available to users, and with implementation of next generation networks (NGN), or similar future network propositions, the number of services and different providers are expected to increase rapidly. More than one actor may be involved in delivering a service, and the needs for regulation of QoS between actors involved in delivering a service are therefore even more present in NGN.

Creating a model with effective traffic and network handling methods for packet based networks, using tools such as MPLS, DiffServ, routing mechanisms and SLAs is one of the future goals. This thesis concentrates on methods of how such a model can be realized, using different tools.

One proposal of regulating QoS is to implement service level agreements (SLA) as standard, regular agreements being used whenever an exchange of service resources is made. An SLA should include service description, user's and provider's rights in terms of delivery, faults, service degradation, monitoring, pricing etc. Other quality of service terms which are service specific are often included in attachments or as individual QoS agreements. SLAs can also be used by customers to compare similar services, that way improving the competing environment.

Standardizing SLAs has been a main focus for many standardization organizations; how can the SLAs guarantee quality of service, as opposed to free market competition and legal regulation. A naïve perspective may be to say that the laws in Norway and free market competition are sufficient to regulate the multi-provider environment, but there are reasons why it is not so. Service Level Agreements are used to ensure that all relationships in an actor network environment, be it service providers, network operators, customers etc, operate in the "correct" way; ensuring quality of service, traffic engineering, and economic and legal issues. This is why the content of SLAs is one of the "hottest" topics among big actors in the telecommunication sector.

Table of Contents

FOREWORD 1

ABSTRACT 2

TABLE OF CONTENTS 3

LIST OF FIGURES AND TABLES 5

1. INTRODUCTION 6

1.1: GENERAL 6
1.2: PROBLEM STATEMENT 6
1.3: SCOPE 7

2. PROBLEM STATEMENT 8

2.1: THE ASSIGNMENT TEXT (ORIGINAL) 8
2.2: MY INTERPRETATION/FOCUS 8
2.3: MOTIVATION 8
2.4: HOW TO GUARANTEE QOS IN A MULTI-PROVIDER ENVIRONMENT? 9
2.5: DO WE NEED SLAS? 10

3. KEY TERMS AND DEFINITIONS 12

3.1: DEFINITIONS 12
3.2: ACTOR VS. ROLE 13
3.3: CIRCUIT VS. PACKET SWITCHING 14
3.3.1: Circuit switching 14
3.3.2: Packet switching 15
3.3.3: Comparison of switching techniques 16
3.4: WHAT IS A NETWORK? 17
3.4.1: Autonomous System (AS) 17
3.4.2: Virtual Private Network (VPN) 17
3.4.3: Internet 18
3.5: REAL-TIME TRAFFIC VS. NON-REAL-TIME 18
3.6: WHAT IS CPE? 18

4. NETWORK AND SYSTEM MODELS 20

4.1: CURRENT WORK 20
4.1.1: Next Generation Networks (NGN) 20
4.1.1.1: Layering 21
4.1.2: Global Information Infrastructure (GII) 22
4.2: NETWORK LAYERING TODAY 22
4.2.1: TCP vs. UDP 23
4.2.3: Internet Protocol (IP) 25
4.3: ASYNCHRONOUS TRANSFER MODE (ATM) 25
4.3.1: ATM Service Categories 26
4.3.2: ATM Adaptation Layer (AAL) 27
4.4: SYNCHRONIZED DIGITAL HIERARCHY (SDH) 28
4.5: TRANSMISSION TECHNOLOGY IN THE PHYSICAL LAYER 29
4.5.1: Digital Subscriber Line (xDSL) 29
4.5.2: Cable access 30
4.5.3: Wireless access 30
4.6: OPTICAL NETWORKS 31

5. REAL-TIME OVER PACKET 32

5.1: REAL-TIME TRANSFER PROTOCOL (RTP) 32
5.1.1: RTP Control Protocol (RTCP) 33
5.2: H.323 33
5.3: SESSION INITIATION PROTOCOL (SIP) 35

6. QOS MECHANISMS 36

6.1: QUALITY OF SERVICE (QOS) 36

6.1.1: Integrated Services (IntServ) and RSVP 36
6.1.2: Differentiated Services (DiffServ) 38
6.2: ROUTING 39
6.3: MULTI-PROTOCOL LABEL SWITCHING (MPLS) 42
6.3.1: Generalized MPLS (GMPLS) 44
6.4: SECURITY ISSUES 45
6.4.1: Internet Protocol Security (IPSec) 45
6.5: TRAFFIC ENGINEERING 48
6.5.1: Traffic Management 49
6.5.2: Admission control 50
6.5.3: Policing 50
6.5.4: Traffic shaping 51

7. SERVICE LEVEL AGREEMENTS (SLA) 52

7.1: SERVICE QUALITY AGREEMENT 53
7.2: ONE STOP RESPONSIBILITY 53
7.3: WHAT TO INCLUDE IN AN SLA 54
7.4: HOW TO WRITE AN SLA 55

8. MODELING SLAS 56

8.1: SLA MODELS 56
8.1.2: How to use the SLA information 57
8.1.2: Monitoring 57
8.1.3: Measuring 58
8.2: SLA SCENARIOS 58
8.2.1: A simple case 59
8.2.2: More than one service provider – case A 60
8.2.3: More than one service provider – case B 61
8.2.4: More than one network operator 63
8.2.4.1: Mobility 63

9. PARAMETERS AND EXAMPLES 64

9.1: REAL-TIME (MULTIMEDIA) APPLICATIONS (VOIP) 64
9.1.1: The service 64
9.1.2: Constraints 64
9.1.3: Transmission between user and service provider 64
9.1.4: Scenario details 65
9.1.5: SLA between User and SP 66
9.1.6: SLA between SP and NO 71
9.2: HIGH-SPEED DATA COMMUNICATION (ATM) 75
9.2.1: The service 75
9.2.2: Constraints 75
9.2.3: Transmission between user and service provider 75
9.2.4: Scenario details 76
9.2.5: SLA between User and SP 77
9.2.6: SLA between SP and NO 79
9.3: EXAMPLES 80
9.3.1: SLA user-SP; video conference 80
9.3.2: SLA user-access provider; DSL connection 84
9.3.3: SLA access provider-NO: data communication 88
9.3.4: Comments 92

10 TODAY'S SITUATION – HOW DOES IT WORK 93

10.1: RELEVANT LAWS 93
10.2: THE "SLAS" TOWARDS ACTORS IN THE PRIVATE MARKET 94
10.3: THE "SLAS" TOWARDS ACTORS IN THE ENTERPRISE MARKET 95
10.4: REGULATING SERVICE LEVELS IN TODAY'S MARKET 96

11. REFLECTION AND ANALYZE 98

11.1: DISCUSSION 98

11.1.1: Weighting the values 104

12. CONCLUSION 105

12.1: CONCLUDING REMARKS 105
12.2: CHECK LIST 105
12.3: FURTHER STUDIES 106

A.1: REFERENCES 108

A.2: ABBREVIATIONS 111

List of Figures and tables

FIGURE 1: NGN TRANSITION 6
FIGURE 2: SCOPE 7
FIGURE 3: AN EXAMPLE OF SLA [E.860] 9
FIGURE 4: AN EXAMPLE OF AN SLA-NETWORK [E.860] 10
FIGURE 5: ILLUSTRATING DEMARCATION POINT FOR ISDN 19
FIGURE 6: DIFFERENT LEVELS OF DEMARCATION POINT 19
FIGURE 7: EVOLUTION FROM TODAY'S VERTICAL NETWORK SITUATION TO NGN'S HORIZONTAL STRUCTURE [ALCATEL TELECOMMUNICATIONS REVIEW, 1ST QUARTER 2003] 21
FIGURE 8: LAYERING 23
FIGURE 9: PROTOCOL STACK 23
FIGURE 10: STM-1 CONTAINER 28
FIGURE 11: PROTOCOL-"ZOO" [27] 32
FIGURE 12: H.323 PROTOCOL STACK [8] 34
FIGURE 13: PACKET HANDLING USING DIFFSERV 38
FIGURE 14: ILLUSTRATING IGP AND BGP 41
FIGURE 15: A NORMAL SHIM HEADER USED IN MPLS [4] 43
FIGURE 16: ESP ENCRYPTION AND AUTHENTICATION 46
FIGURE 17: SHOWING IPSEC TUNNEL MODE 47
FIGURE 18: TRAFFIC ENGINEERING PROCEDURE 49
FIGURE 19: HANDLING QOS FOR MULTIPLE CONNECTIONS 49
FIGURE 20: ILLUSTRATION OF TOKEN BUCKET 51
FIGURE 21: AN EXAMPLE OF A MULTI-PROVIDER ENVIRONMENT [E.860] 52
FIGURE 22: ILLUSTRATING ONE STOP RESPONSIBILITY [E.860] 53
FIGURE 23: SCENARIO 1; ONE USER RELATES TO ONE ACTOR WITH TWO ROLES 59
FIGURE 24: SCENARIO 2; TWO SPS, ONE NO. THE USER MAY RELATE TO SP1, SP2 OR BOTH. 60
FIGURE 25: SCENARIO 3; TWO SPS, ONE NO. THE USER RELATES TO SP1 61
FIGURE 26: COMPLEXITY INCREASES 62
FIGURE 27: SCENARIO 4; TWO SPS, TWO NOS. THE USER RELATES TO SP1. 63
FIGURE 28: VIDEO CONFERENCE SCENARIO (GENERAL) 65
FIGURE 29: VIDEO CONFERENCE SCENARIO (DETAILED) 66
FIGURE 30: VIDEO ON DEMAND/PAY PER VIEW SCENARIO (GENERAL) 76
FIGURE 31: VIDEO ON DEMAND/PAY PER VIEW SCENARIO (DETAILED) 77
FIGURE 32: CHAIN OF PROVIDERS 92
FIGURE 33: AREA OF CONSIDERATION 98
FIGURE 34: FUTURE WORK 106

TABLE 1: COMPARISON OF SWITCHING TECHNIQUES [STALLINGS 2000, P 312, TABLE 10.1] 16
TABLE 2: ATM SERVICE CATEGORIES 27
TABLE 3: SIP MESSAGES [29] 35
TABLE 4: DIFFERENT ROUTING SCHEMES 40
TABLE 5: EXAMPLE OF CONTACT-ROUTINES [18] 59
TABLE 6: REACTION AND COMPENSATION (EX. VIDEO CONFERENCE) 83
TABLE 7: REACTION PATTERNS (EX. DSL-CONNECTION) 87
TABLE 8: COMPENSATION AND REACTIONS (DATA COMMUNICATION) 90
TABLE 9: COMPARING DIFFERENT REGULATION-ENVIRONMENTS 103

1. Introduction

1.1: General

“It is possible that the entire PSTN will be replaced with a new network based on IP, which will have significant implications for equipment supply, investment requirements and the range and cost of services offered. However this transition will inevitably take several years, due to a number of factors…” This declaration is made in the final report for the European Commission, 2004 [9], and applies to the work being done by several data- and telecommunication organizations and study groups all over the world. In general the main idea is to evolve today’s vertical layered infrastructures, including separate networks for PSTN, PLMN, Internet etc, in to a horizontally layered structure, illustrated in figure 1.

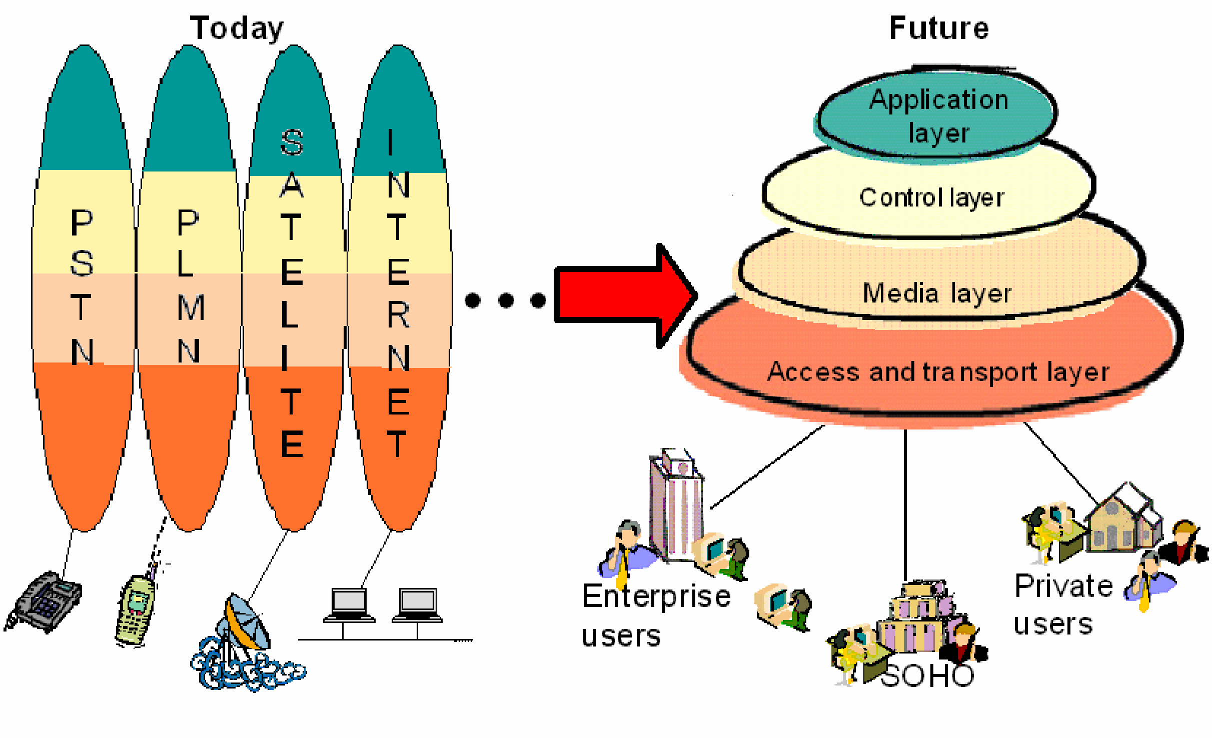


Figure 1: NGN transition

The vertical network structure, which is currently being used, requires network management, access infrastructure and specific applications for each and every system. The proposed Next Generation Network (NGN) simplifies this by creating a common access and transport layer able to handle all kinds of access types, a common application layer making applications available to all users, no matter what customer premise equipment (CPE) they use, and a control layer running an overall management situation. This structure is described in more detail in chapter 4.1.

1.2: Problem statement

The number of different services and service providers available in today’s market situation is constantly increasing due to deregulation, user demands and technologic development, among other things. Different actors in the market are therefore stressing the fact that delivering a service in a multi-provider environment is complex and needs some kind of regulation; the users require service quality and better performances (e.g. higher bit rates on their internet connections), while providers compete with each other in terms of pricing, quality and other service characteristics. The situation is likely to expand with even more services and providers with NGN realization, and some means to control the situation are therefore required.

Service Level Agreements (SLAs) are agreements which can be made between the different actors in a multi-provider environment, for instance between an end-user and its service provider, or between two service providers (see chapters 7 through 9). SLAs are written to guarantee that what is delivered is in fact what was agreed upon according to the perception of the parties involved, referring to any kind of service; applications, broadband access service, network components etc. SLAs are supposed to include regulations in terms of quality, delivery time etc, and comprise service characteristics, responsibilities and priorities of every part involved, to make the actor-to-actor relationships easier.

In a multi-provider environment it is likely that actors depend on other actors, meaning e.g. that a primary service provider delivering a service, may be dependent of service components from a sub-provider. In such a scenario it is desirable to implement an SLA between the primary provider and its sub-providers to ensure that the user does not receive degraded services due to a sub-provider's flaws. How to guarantee this kind of quality of service and what to include in such an SLA is the thesis' main concerns.

1.3: Scope

Chapter 2 will include a more detailed problem statement, while chapter 3 will supply the reader with the definitions and key terms required to understand the thesis. Chapters 4 trough 6 will include the **technical** aspects referred to in the thesis. System and protocol description, real-time, and QoS-mechanisms being used in packet switched networks are central in this section. Chapter 7 describes in detail what a service level agreement is intended for, while chapters 8 and 9 will outline a few case examples describing the theory in practice, relating to **SLA issues**. This section sets focus to some of the problems occurring in multi-provisioning service environments.

An **analysis** of today's situation follows in chapter 10, with reflection in chapter 11 and concluding arguments listed in chapter 12.

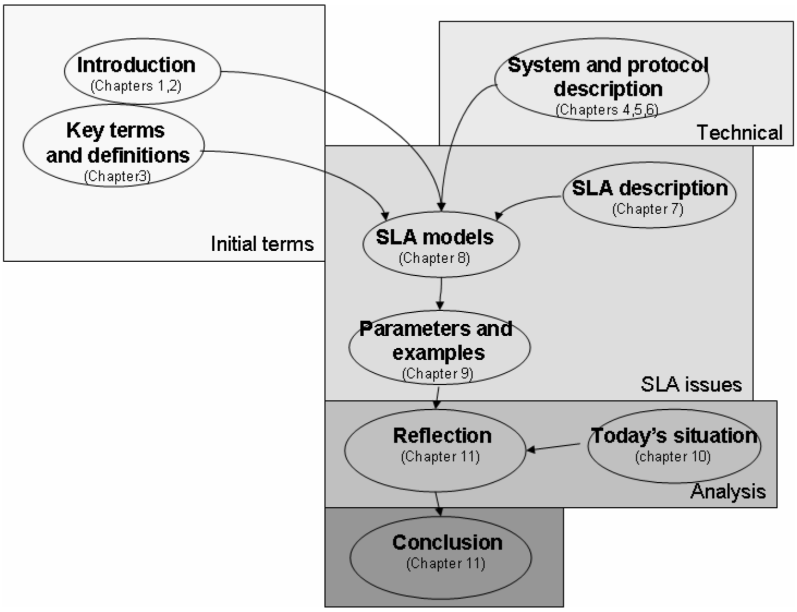


Figure 2: Scope

2. Problem Statement

The thesis' main concerns are:

- **How to guarantee quality of service?**
- **What to include in service level agreements between different actors?**

The complexity increases when looking at the foretold multi-provider environment in what is referred to as Next Generation Networks (NGN). NGN is defined quite generally (see chapter 4.1), but states mainly that all telecommunication traffic should be sent over packet switched networks, rather than several vertically layered networks using different transmission media (circuit switched fixed networks, TDM/FDM-based mobile networks etc. See figure 1). The futuristic NGN-perspective will be taken in relation to the thesis-problems.

2.1: The assignment text (original)

A number of roles connected to telecom networks have been identified in project reports and reports from standardization organizations (e.g. E.860). As results of the increasing sales in telecom services combined with regulatory demands for competition and growth of operators/providers, the number of actors involved in telecom is likely to evolve. Numerous actors may be involved in one single action, especially during session handling (i.e. a telephone call. See chapter 8). Due to this, the quality that end-users experience will depend on the quality that each actor offers. How can service quality then be guarantied? Furthermore, according to the increase of service variants and more specialized operators/providers, the complexity of these issues will soon expand.

Corresponding to this it is important to point out the difference between an actor and a role (see chapter 3.2). An actor may consist of several roles, at the same time as one role can be filled by several actors (i.e. by competition). Because of this the actor (operator or provider) can decide whether it would like to fill all the roles itself, or buy other actors' "role-results"/services. It will also be possible for actors to merely sell their services (like a product) to other actors/service providers, making the buying service providers act as users, and profit from that.

2.2: My interpretation/focus

Within a telecommunication service provider's environment there will be various actors delivering multiple similar and/or different services. This thesis will mainly concentrate on how these actors influence each other, particularly in relation to quality of service (QoS).

The perspectives are many; economics, legal and technical. The focus is set to enlighten some of the technical matters, concentrating especially on how to guarantee quality towards an end-user in a multi-provider packet switched network. This is also referred to as end-to-end quality of service.

2.3: Motivation

With the world's technology centers working to evolve public switched telephone services into what is referred to as Next Generation Network (NGN – see chapter 4.1 for details); reducing costs, overhead etc, the European Commission concludes that the work in Europe proceeds more slowly than the rest of the world [9]. Nevertheless, more and more voice traffic will be carried over packet switched networks, and in the process a wide range of services will emerge, supported in some cases by new network equipment. The challenges of smooth interworking between a multiple of providers in the same environment, implementing QoS and various traffic conditions, are defined by international standardization organizations and working groups in ITU-T (see E.860), ETSI (see ETR 003), EURESCOM, IEC etc, and is currently being researched in Norway by companies like Telenor and the Norwegian Post and Telecommunication Authority (ref. [12] and [16]).

In the internet and other inter-connected networks there are various ways to apply quality of service, and different domains using different QoS-schemes may have problems communicating and maintaining the wanted QoS-level and traffic conditions for inter-domain data flows. A common QoS framework should therefore be established, rather than making applications and routers etc relate to various interpretations and principles for QoS, depending on which services and systems are involved [16]. The content of service specifications should hold higher granularity than those implemented today, and thus, increasing the granularity of QoS-descriptions. A challenge, according to Jensen, Grgic and Espvik in [16], would be how to identify QoS-levels of granularity according to the service specification.

2.4: How to guarantee QoS in a multi-provider environment?

This thesis will mainly concentrate on quality of service and SLA-issues in the multi-provider environment predicted to arise with NGN realization, in relation to service level agreements made between the actors involved in providing services.

In general it is normally proposed some kind of "contract" between a service provider and a user, whether the user is a service provider making use of (e.g. re-selling) the service (wholesale) or an end-user. Such an agreement should comprise the terms the actors have agreed upon; service characteristics, responsibilities and priorities of every part involved. The contract is called Service Level Agreement (SLA), where the part concerning quality of service parameters is referred to as QoS Agreement. The contents of the QoS Agreement will be the issues described most thoroughly in the thesis. See chapters 7 through 9.

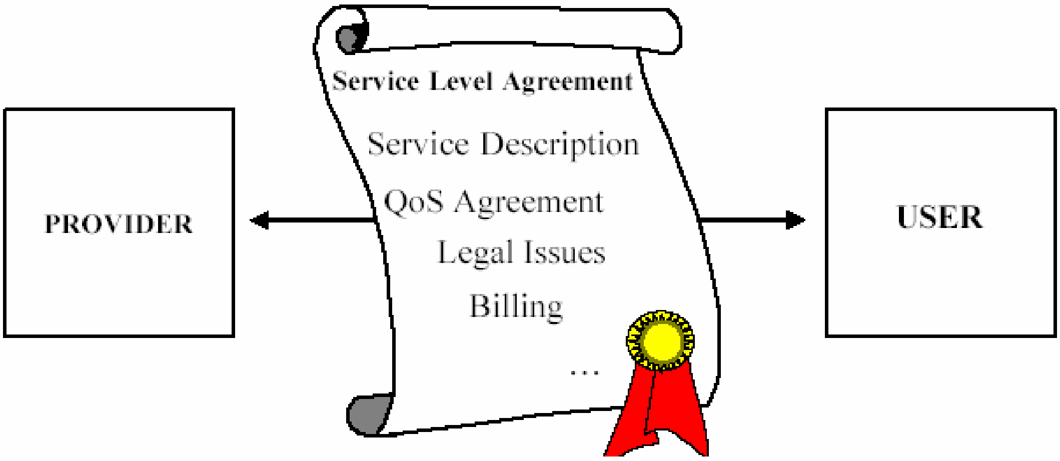


Figure 3: An example of SLA [E.860]

The figure above shows a simple example, but the situation becomes more difficult when more service providers are added, including sub-providers depending on each other, and/or include SLAs between service provider(s) and network operator(s) through which the service is distributed (see figure 4). A (primary) service provider may for instance rely on services supplied by sub-providers, meaning that the QoS delivered to the user may depend on QoS delivered by the sub-providers, as well as the QoS from the primary provider. The primary provider must therefore find a way to define and operate mechanisms for managing QoS for its users and for other relevant providers.

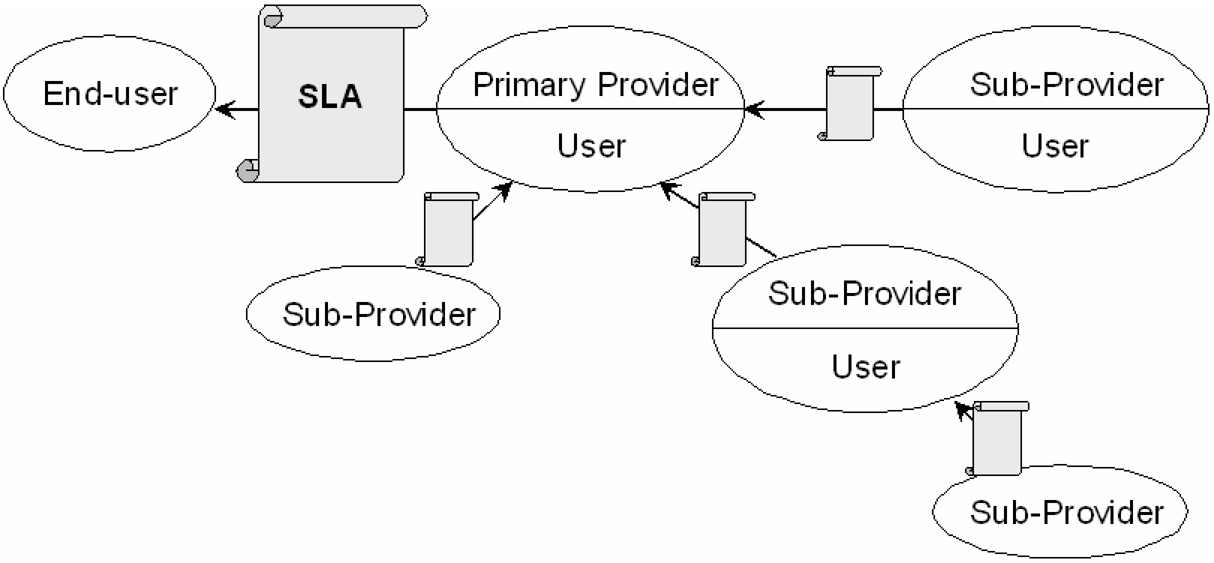


Figure 4: An example of an SLA-network [E.860]

The SLA may also include statements about performance, tariffs and billing, service delivery and compensations. See chapter 7 and 8 for more details, and chapter 9 for case examples.

2.5: Do we need SLAs?

Another discussion that will be considered in this thesis will concern whether there is a need for SLAs or not. There is most certainly some kind of SLA for all kinds of communication transactions between two or more actors. For instance a regular PSTN telephone call normally has a kind of SLA that includes service availability (SA) etc, but most people take that for granted; the QoS level is so high, that no one ever expects to pick up the receiver and not get a dialing-tone.

In most cases of regular telecom services like telephony, internet, cable-TV etc we do not even think that there would be a need to make an agreement to secure our rights, because to a private person the harm done if the service does not work as proposed is not (normally) a critical issue. If a private person is not satisfied with one service provider's service provision, he/she may decide to buy her/his service from another service provider. Is the *free market and competition* between the service providers sufficient to regulate well enough deals for the users? Or do service providers see the use of SLAs as a *competitive edge* towards other providers? This is discussed to some extent in chapters 10 and 11.

In the enterprise/business market a service failure may, in most cases, be more critical, and a change of service provider more expensive and complicated than for a private person. For instance if a stock broker does not receive stock price information at the right time, he/she might loose a lot of money, or if a sales manager does not get to send that specific e-mail to close a sale in time, he/she might loose the sale…

For a medium-sized business with 200 employees, a downtime of the computer systems (e.g. internet contact) 3 hours a month, may cost the business $16.110. Affecting larger business, service outage may impact the business by a monthly range between $80.000 and $5.6 million [19]. In such cases it may be preferred that there are some terms agreed upon, - to which the service provider commits to, to ensure that pro-active actions are taken against possible “threats” as e.g. service outage.

3. Key terms and definitions

Some of the terms used may need an explanation if the reader is not familiar with jargon used in telecommunication. This chapter will describe the general terms with which understandings the reader should be aware of.

3.1: Definitions

A **service** is a group of functions provided by an organization to a user through an interface [E.860].

A **service** is a collection of interconnected access points with a software protocol structure that enables communication [Stallings 2000, p. 499].

A **role** envelops a set of functions enabling anyone taking on the role to provide (or to use) a set of services to its environment. An **actor** could take on several roles, and/or a number of actors could take on the same set of roles, e.g. when competitors are present in the same market [P806-GI]. See further description in chapter 3.2.

An entity that delivers a service to another entity takes on the role as a **provider**, while the entity that receives the service is referred to as a **user** [E.860].

A **network provider** (operator) is an organization that provides a network for the provision of telecommunications service [ETR 003].

A **service level agreement** is a formal agreement between two or more entities (actors) that is reached after a negotiating activity with the scope to assess service characteristics, responsibilities and priorities for every part [E.860].

Quality of service is defined as “a degree of conformance of the service delivered to a user by a provider, with an agreement between them.” [P806-GI]

Quality of service is “the collective effect of service performance which determine the satisfaction of a user of the service” [E.800]. See chapter 6.

A **QoS parameter** is a variable that is used to assess QoS [P806-GI].

Network performance is the ability of a network or a network portion to provide the functions related to communication between users [E.800].

Routing information is information about the topology and the delays of the internet [Stallings 2000, p. 571]. (You can read about routing in chapter 6.2.)

A **forwarding algorithm** is the algorithm used to make a forwarding decision for a particular datagram, based on current routing information [Stallings 2000, p. 571].

Traffic engineering is improving user performance and making more efficient use of network resources by adapting the routing of traffic to the prevailing demands [1]. (See chapter 6.5.)

An **application** is a program designed to perform specific function directly for a user or, in some cases, for another application program. Applications use the services of the computer’s operating system and other supporting applications [24].

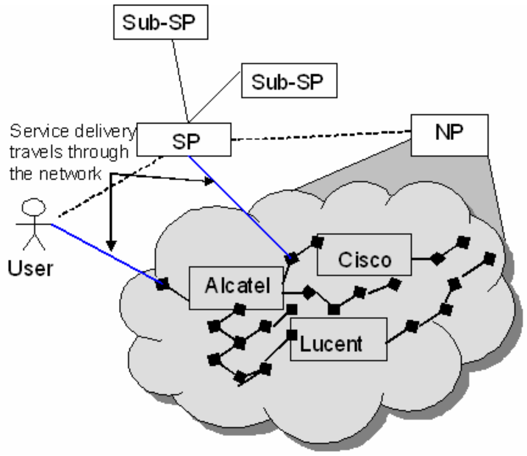
A **session** is a logical association between two or more endpoints, offering the possibility to make use of a (tele-) communication service [EG 202 009-3].

3.2: Actor vs. role

An **actor** could take on several **roles**, and/or a number of actors could take on the same set of roles, e.g. when competitors are present in the same market [P806-GI].

An entity is a generic unit involved in using/delivering a service, and is characterized by its states and transition from a state to another. During a transition an entity can execute functions and interact with other entities through its outputs. A number of entities can be composed into a new entity, and two or more entities can interact [P806-GI]. The entity delivering a service to another entity is called provider, while an entity which receives a service is referred to as user [E.860].

A business entity, e.g. a company, is named an actor when it takes actions on the market [P806-GI]. An actor can fill more than one role, and one role can be filled by several actors. Thus, an actor may be an entity filling roles such as service provider(s), network provider(s) and/or user(s), all of which takes an active role in an entity's scenario.



An example may be a user buying a service from a service provider, which in its case has supplied the service and components from two other service providers (sub-providers). The service is delivered over an infrastructure delivered from a network provider/operator, including components from Cisco, Lucent and Alcatel. In this case the user, all of the service providers (including the two sub-providers), the network provider (including management functions etc), Cisco, Lucent and Alcatel[1] are actors; - those who actively contribute to the operation of delivering a service from A to B.

EURESCOM's project P1203 identifies several roles/actors in an environment called Beyond 3rd Generation (B3G) [P1203]. B3G is a futuristic project, similar to NGN, but for mobile services; delivering mobile services to users including multimedia, real-time voice etc over IP. The roles they identify can be applied more generally and further exemplify the actor vs. role relationship. The identified roles are:

(1) *End users* who may demand ubiquitous access to applications and services, and require appropriate quality and security at reasonable costs, and understandable and user friendly interfaces on terminal equipment.

(2) *Network Operators* (NO) who most commonly have their own service delivery platform. They require optimization of network resources like efficient and flexible QoS and security handling. Network operators are also considered actors which may comprise different roles such as a fixed line access provider, a mobile provider and/or a network provider. Network providers have normally no contact with the end-user, but only provide the infrastructure for network transport services.

[1] Note that Alcatel, Cisco and Lucent are considered to manage their own network components in this example. Physical component providers are not normally considered actors because they are regularly passive (take no actions) in delivering a service.

(3) *Service Providers* (SP) who hide the complexity regarding networks and sub-provider problematic for the end-user. They may require possibilities to "fast, open service creation, validation and provisioning" [P1203], automatic service adaptation as a function of available bit rates, and secure services with QoS.

(4) *Content Providers* (CP) who deliver different kinds of content towards an end-user. Their characteristics are adapting various contents to users' requirements; depending on CPE, location and user preferences, and enables access to a large market of services through a single interface.

In addition actors such as bandwidth brokers (to manage the multiple of service and content providers), and SMS-operators etc are likely to exist in the B3G-environment. This example shows some of the complexity handled later in the thesis, regarding problems of how agreements can be controlled between roles, actors and other entities in a multi-provider environment. Relationships between different actors and roles are described in chapter 8, with more details in chapter 9.

3.3: Circuit vs. packet switching

Circuit switching and packet switching "differ in the way the nodes switch information from one link to another on the way from source to destination" [Stallings 2000, p. 278]. Circuit switching was originally designed to transmit voice in public switched telephone networks (PSTN/POTS), but has eventually developed to handle data traffic as well. Packet switching was developed to better utilize the resources in a network for bursty traffic.

3.3.1: Circuit switching

A network using circuit switching is normally implemented to handle both analog and digital data in full-duplex[2] connections. A connection between two entities is set up by establishing a dedicated path between the two, reserving the resources on the link[3] for their exclusive use during the connection. This means that during a connection no other actor may use the resources on the particularly set-up link. These types of links are normally referred to as connection oriented.

Listening
40% active use
Blablablabla
40% active use
Blablablabla
Listening
80% active usage in total

In a telephone conversation there are continuous flows of data, although this data can include "silent" coded data, meaning the periods where nobody talks. Only approximately 30-40% of the time a regular telephone conversation lasts are actively used in each direction, meaning about 20% of the time measured in both directions are so-called silent moments. However, since the resources are reserved for that specific session at all time, this may cause a waste of resources in the network; more resources are needed than necessary. Opposed to this, data flows in packet switched connections may have times when the connection is just idle, but not sending any data, thus

[2] Full-duplex means you can send data both ways on a connection at the same time (two-way communication), as opposed to half-duplex when only one part can transmit data at the same time, and the other must wait.
[3] On each physical link, a logical channel is set up for every connection. Referring to a "link" in this section really means a logical channel, typically 64kbps for an ISDN connection.

maximizing the network utilization. This is one of the reasons why packet switching is considered in relation to NGN and other future implementations for voice traffic.

3.3.2: Packet switching

Packet switching solves the problem of unused resources in a network by splitting the data flows into packets consisting of user data and control information. A packet sent through the network uses the resources needed to transmit that specific packet plus a little processing time; meaning other packets/data flows can make use of the same resources at other times.

Some of the advantages of using packet switching are:

(1) Greater link-efficiency because the node-to-node links are shared dynamically by many packets over time,

(2) Possible data-rate conversion so that connections between sources using different data-rates become possible,

(3) Packets can be accepted even though the traffic load increases by increasing delivery delay, and most importantly;

(4) Packets can be prioritized.

In packet switching there are two possible ways to switch data; either by establishing virtual circuits or by using datagram transmission. Virtual circuit is the method most related to circuit switching, because a preplanned route must be established before data can be sent, and all the packets follow the same route between the actors[4]. Although this may seem to be the same as circuit switching, it is not, because the resources on the connection is not reserved only for that connection. The packets still traverse the net by sharing the medium with other packets/data flows, but are acknowledged to belong to the same flow by a unique virtual circuit identifier.

In datagram transmission each packet is treated independently, by processing its overhead data at each node (router/switch) and switching it through the net in (possible) different routes according to traffic conditions and routing algorithms. The disadvantage is that packets may be dropped along the way and/or arrive in a different order than they were sent. The receiver may have to use some resources to set the dataflow correctly together, or even ask for retransmission.

An advantage of using virtual circuits is that it is no need for routing decisions for each packet; - it is only done once when setting up the VC, which leads to quicker transmission. On the other hand, when sending a small number of packets, the datagram transmission may be quicker, because the connection set-up is avoided. Datagram transmission may also be more flexible regarding node failure and congestion in the network, because it may take different routes between the sender and receiver and therefore spread the traffic more. Routing decisions and other traffic engineering routines are described in chapter 6.

[4] In more sophisticated network designs, the network may dynamically change the route established to a particular VC in response to changing traffic conditions like overload or failure in parts of the network [Stalling 2000, p. 317].

3.3.3: Comparison of switching techniques

The key-strength of circuit switching is that the connection becomes transparent to the user when it is first set up, and no special networking logic is needed [Stallings 2000, p. 281]. A disadvantage of using packet switching for voice data flows may be that all analog data must be converted to digital data before transmission, but this conversion is often performed anyhow, because most data switched networks today are digitalized (e.g. ISDN). Overhead bits in each packet which increases the quantity of data being sent are also needed, thus increasing the transmission time[5]. Below the key points of the different switching techniques are summed up.

Table 1: Comparison of switching techniques [Stallings 2000, p 312, table 10.1]

Circuit switching	**Datagram packet switching**	**Virtual-circuit packet switching**
Dedicated transmission path	No dedicated path	No dedicated path
Continuous transmission	Transmission of packets	Transmission of packets
Fast enough for interactive	Fast enough for interactive	Fast enough for interactive
Messages are not stored	Packets may be stored until delivered	Packets stored until delivered
The path is established for entire conversation	Route established for each packet	Route established for entire conversation
Call setup delay; negligible transmission delay	Packet transmission delay	Call setup delay; packet transmission delay
Busy signal if called party busy	Sender may be notified if packet not delivered	Sender notified of connection denial
Overload may block call setup; no delay for established calls	Overload increases packet delay	Overload may block call setup; increases packet delay
Electromechanical or computerized switching	Small switching nodes	Small switching nodes
User responsible for message loss protection	Network may be responsible for individual packets	Network may be responsible for individual packets
Usually no speed or code conversion	Speed and code conversion	Speed and code conversion

[5] The transmission time depends on the switching technology and amount of data. E.g. in ATM small packets 53 octets long are switched efficiently through the network.

Fixed bandwidth	Dynamic use of bandwidth	Dynamic use of bandwidth
No overhead bits after call setup	Overhead bits in each packet	Overhead bits in each packet

3.4: What is a network?

Network is a word used in many different circumstances. In the most general way, a network can be described as more than one unit (e.g. computers) connected together so that they can communicate.

A network located within a small range e.g. a house or a small business is referred to as a Local Area Network (LAN), whereas a network with a wider range is called a Wide Area Network. If the connection technology is wireless, it is typically called Wireless LAN (WLAN), because the range of an antenna providing wireless connection is small. In some WLANs the devices are connected ad hoc (the wireless devices communicating directly with each other), using for instance Bluetooth-technology, but may also connect to some kind of base station (router/switch…) which forwards data traffic. See chapter 4.5 for more information about transmission technology.

3.4.1: Autonomous System (AS)

In tele- and data-communication, *autonomous systems*, or so-called ASes, are often spoken of. A network system is characterized as an AS if

(1) the group of routers exchange information using the same routing protocol,

(2) the set of routers and networks are managed by a single organization, and

(3) that there is a path between any two pair of nodes [Stallings 2000, p. 572].

Routers within an AS are free to choose their own mechanisms for discovering, propagating, validating and checking the consistency of routes [Comer 2000, p. 274]. Thus, an AS may be any kind of network or networks using the same set of rules; switching technology (IP/ATM or similar – chapter 4), same QoS-schemes (e.g. MPLS and DiffServ – chapter 6) and Border Gateway Protocol (BGP) to communicate with other networks/ASes etc.

3.4.2: Virtual Private Network (VPN)

A VPN is *private* because the technology guarantees that the communication between any pair of computers in the VPN remains concealed from outsiders, and it is *virtual* because it uses the global Internet to pass traffic from one side to the other [Comer 2000, p. 391].

A VPN is typically used when a business wants to connect different systems/networks (for instance located in different countries) and secure the communication between them. If one branch office wants to communicate with another branch office, the communication may happen across the Internet, thus making it open to threats and attacks from outsiders. Especially, if the data being communicated is of confidential kind, a secure connection is highly appreciated.

3.4.3: Internet

The reader should be aware of that *an internet* is an interconnection between more than one network, and must not be confused with *the Internet*, which is the public related “database” which a user may reach from his/her computer. An internet may be a connection between several LANs belonging to one business, forming an AS, but which might not be connected to the Internet.

3.5: Real-time traffic vs. non-real-time

All data sent over a packet switched network is fragmented into packets/cells. Normally, data traffic is switched using *best-effort*, meaning the network provides the resources available at the moment and forwards it through the network in various ways. There are a number of switching techniques which guarantees throughput (see chapters 5 and 6 for details), but using best effort, a number of packets may be lost, delayed or delivered in the wrong order. Applications who handle this kind of transmission is referred to as non-real-time traffic; it can rely on re-transmissions if lost packets and/or buffer the received packets until the right order of the packet stream is established. Non-real-time traffic may typically be e-mail, file transfer etc.

Applications which require timely transmission and delivery, meaning some kind of guaranteed throughput at a certain time, are called real-time traffic [Stallings 2000, p. 540]. Traffic of this kind is dependent of low delay, low packet loss (not possible to rely on retransmissions because of timing problems), little jitter (delay variation) etc. Real-time traffic is typically IP-telephony, video streaming etc (see chapter 5). Since packet switched networks are not isynchronous; - meaning the entire system should deliver output with the exact same timing as it was generated and all paths having the same delay, additional protocol support is required when sending real-time data.

3.6: What is CPE?

Customer Premises Equipment (CPE) is the equipment needed at a customer’s site to be able to receive and make use of a service. The demarcation point is generally defined as the point where the local loop (the network transmission equipment) ends and the inside wire (which is the responsibility of the customer) begins. The exact location of the demarcation point depends on the technology, the service provider, the service provided, and the location in which the local network is located [13]. See figure 5 and 6 below.

For instance if a customer would like to connect and perform a telephone call by using ISDN, he/she would have to get hold of an ISDN compatible telephone set, an NT-box which among other things translates the signals from the subscriber loop to the network, and a connection between the different devices (S-bus). Other Terminal Equipment (TE) can be analogous telephones, fax machines, computers etc, as long as there are some kind of terminal adapter connected to the device so that it can communicate with the other devices on the S-bus. Another example is a computer connected to the internet. CPE in that scenario is a PC, a modem and a network card.

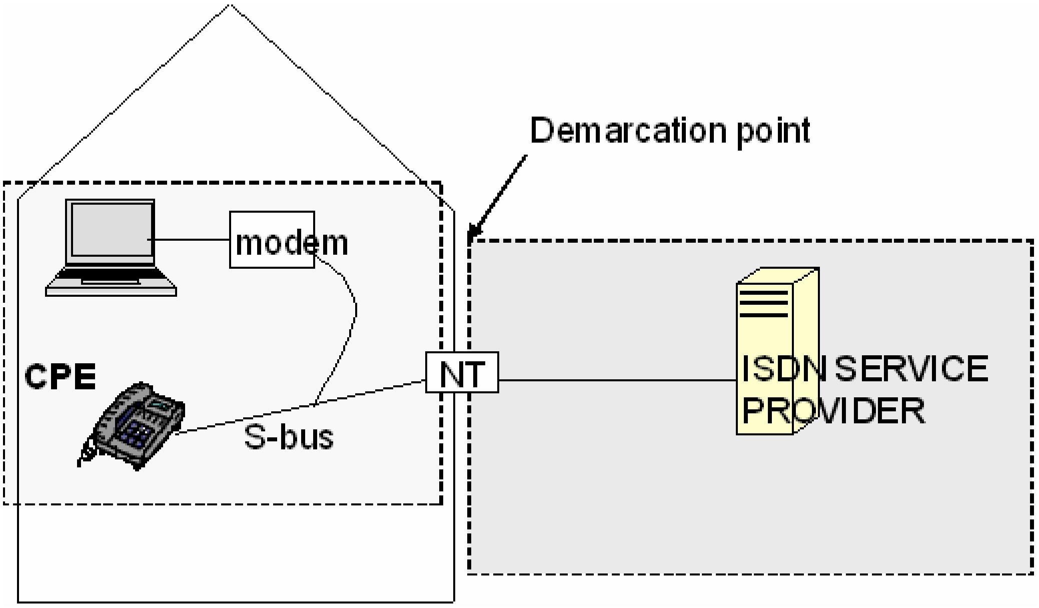


Figure 5: Illustrating demarcation point for ISDN

Demarcation points may be seen differently in relation to the horizontal layered NGN structure (see chapter 4.1.1.1), ergo different levels. One demarcation point may for example be seen from the access provider's point of view, while another point of demarcation may be at the control layer. E.g. if a user decides to buy his/her own DSL-modem and router and still purchase the DSL-service from a service provider, the access demarcation point will be at the router, while for control purposes the demarcation point will be at the DSL-modem, allowing the provider to control the router's settings etc.

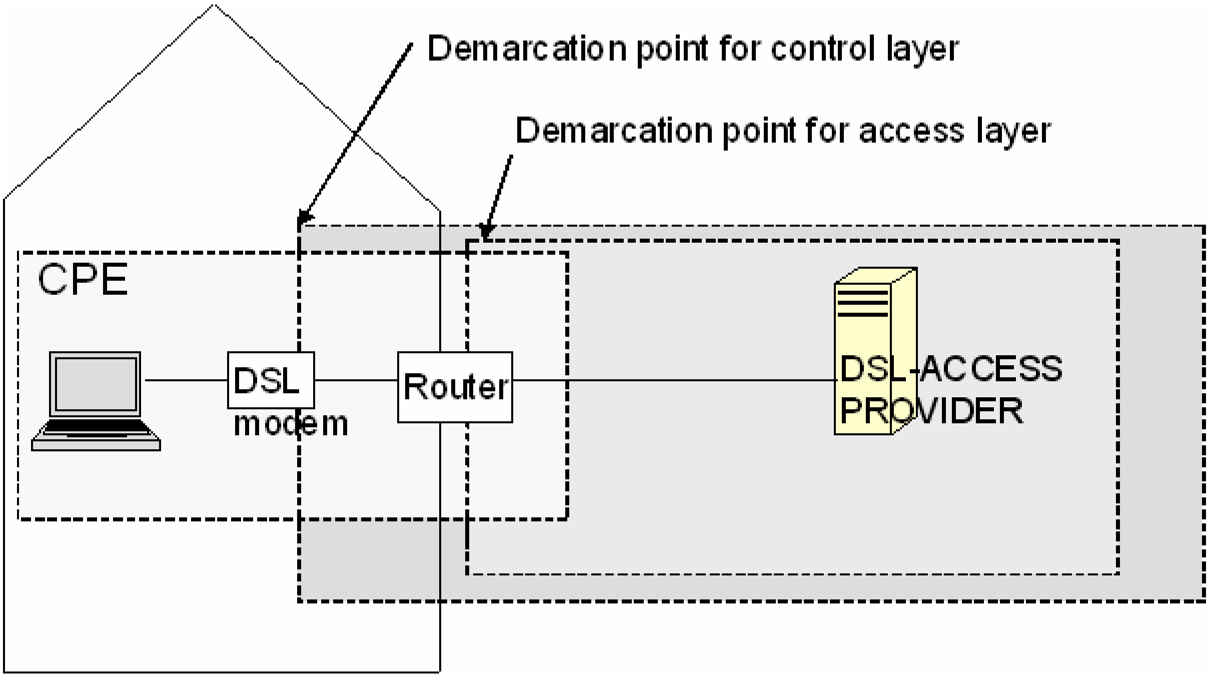

Demarcation point for control layer
Demarcation point for access layer
CPE
DSL modem
Router
DSL-ACCESS PROVIDER

Figure 6: Different levels of demarcation point

4. Network and system models

There is a number of ways to connect to a network, and below the most common access methods are described, starting with the work processes currently going on and continuing describing actual system and protocols being used today.

4.1: Current work

Network and system models are currently being researched, developed and/or standardized by different organizations. In the introduction (chapters 1 and 2) one of these systems where referred to as Next Generation Networks (NGN); an all service system model to provide all kinds of communication over a packet switched network. The definition for NGN is standardized by ETSI, but several other standardization organizations work towards the same goal, using slightly different words and/or definitions to describe it.

4.1.1: Next Generation Networks (NGN)

The development of today's communication networks in telecom in the direction of an "all IP-network", or to use a more general speech; an "all packet-based network", is referred to as Full Service Networks (FSN[6]) or Next Generation Networks (NGN).

ETSI's definition of NGN is as follows:

> NGN
>
> is a concept for **defining and deploying networks**, which, due to their formal **separation into different layers and planes** and **use of open interfaces**, offers service providers and operators a platform which can **evolve in a step-by-step manner** to **create, deploy and manage innovative services**.
>
> *Source: European Telecommunications Standards Institute (ETSI)*

The definition is quite generally written and may possibly be interpreted in various ways. What most of the service providers agree on though, is that NGN should be based on a packet switched network (most commonly ATM or IP), have open interfaces, and have a layered structure – see figure 7. This may decrease the complications when different providers should collaborate within the network.

NGN has an open interface, meaning all components in the network communicate through standardized, commonly available protocols. Components used in the infrastructure should be compatible to any kind of hardware/software from any kind of actor. This means it should be possible for different service providers to distribute services without depending on who owns/delivers the infrastructure (network operators), and who might use their services. This way it will be easier for small providers to offer one single service and profit from that without depending on undesirable agreements with bigger providers like Telenor in Norway.

Until now each network provider has had their own set of access interfaces for service providers. In Norway it is mainly Telenor who owns the PSTN infrastructure and provides access lines (copper) to more than 90% of Norwegian households. In this case it is difficult for any other service provider to offer any kind of service without

[6] The idea of Telenor's FSN is the vision of a full service network which integrates speech, data, multimedia, various forms of access and mobility. The intention is to exploit the latest developments in IP technology, as well as acquiring experience with a service integrating network concept [30].

leasing lines/infrastructure through Telenor's network, - which may cost quite a big deal of money. NGN's idea is to solve this situation by creating one common, open infrastructure, and letting minor or major actors provide whatever services they like (with some restrictions of course). Defining standard connections and interfaces means that service providers only need to develop one type of interface, using a common suite of protocols and signaling mechanisms, to access all the service features from many network (access) providers [TR-058].

The services which actors may provide can be described as everything needed in a communication network; infrastructure (routers, switches, gateways…), physical cables (copper, coax…), Customer Premises Equipment (CPE), internet services, SMS services, billing services etc. The main problems in situations occurring in multi-provider environments are, on which terms the actors should interfere and cooperate with each other.

4.1.1.1: Layering

Today the communication networks are mainly parted into separate networks in what is called vertical integration; a mobile network (PLMN), a public switched phone network (PSTN), a cable-TV network, a packet switched network (typically internet) to name a few. Each of these networks needs their own network management (NM) and overhead system, billing system, infrastructure, access interfaces, services etc. To simplify the costs due to today's circumstances, NGN combines these networks into one in a horizontal layered structure, which has at a minimum one NM-module (layer) which handles billing and different management situations, one switching layer and one control layer. The model below shows one solution of how NGN layering may be implemented.

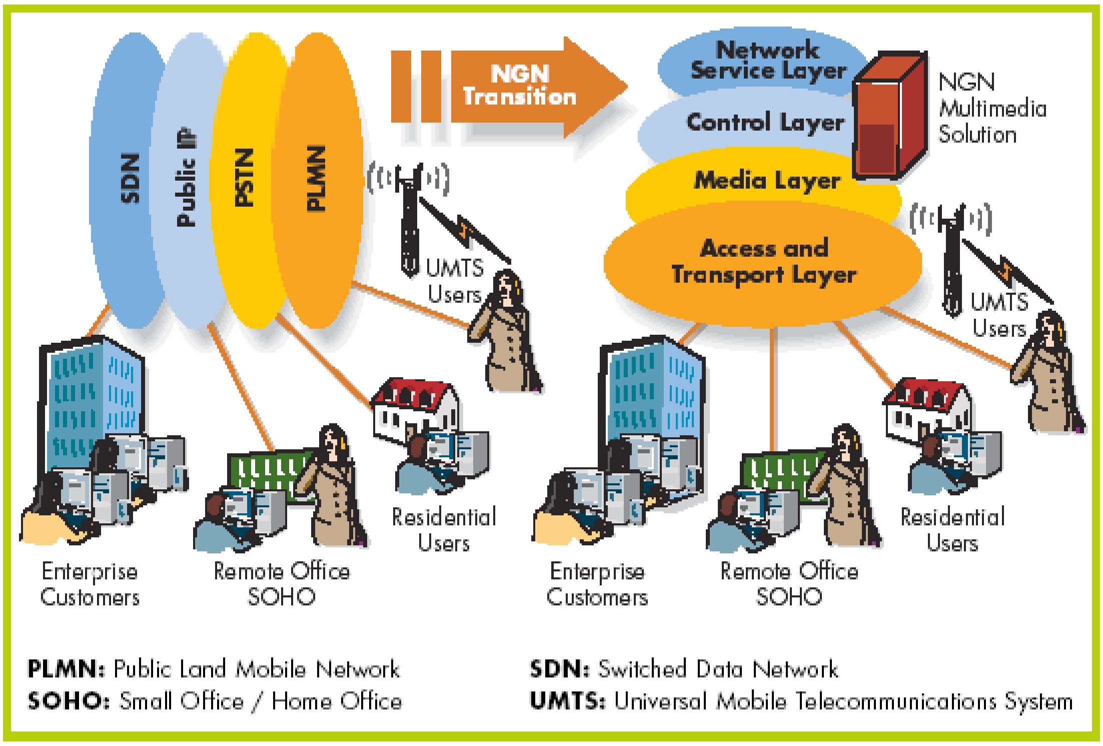


Figure 7: Evolution from today's vertical network situation to NGN's horizontal structure [Alcatel Telecommunications Review, 1st Quarter 2003]

A short description of the NGN layers may be in place.

The Access and Transport Layer is what earlier have been the different infrastructure networks (PSTN, PLMN, DataPak/X.25, Internet…) including access interfaces and transmission media. It now comprises all different transmission media and should be capable of handling any kind of access depending on whatever CPE the customer has at hand.

The Media Layer will mainly be responsible for assembling the different information flows into one, and to "translate" all different transmission media into packet switched data. In the future this layer may merge with the access and transport layer, because all original traffic will be based on packet switching.

The Control Layer will hold the main intelligence of the network. It will provide signaling to set up communication sessions, billing services etc, and hold the information of every customer's relations i.e. what extra services they subscribe to.

The Network Service Layer will hold the contracts (i.e. SLAs) between different service providers, and keep track of all the new services, IN-services etc.

4.1.2: Global Information Infrastructure (GII)

ITU-T is developing "Global Information Infrastructure" (GII), described in ITU-T's Y-series. GII is aimed to be an infrastructure which facilitates the development, implementation and interoperability of existing and future information services and applications within and across telecommunications, information technology, consumer electronics and content provision industries [Y.100]. The GII will also provide interoperability between a multiplicity of applications and different platforms through a seamless federation of interconnected computers etc including line-fed (copper pair, fiber, coax…) and wireless (satellite, mobile…) technologies.

ITU-T has named the model "global" because restrictive national and regional ways of doing business limits customer's information access and personal mobility, thus requiring global standards for information and infrastructure components, and "infrastructure" because of the development in telecommunications; computers, services, applications etc has led to new conditions, demanding different requirements from an infrastructure.

4.2: Network layering today

The open system interconnect (OSI-) model was developed and standardized by ISO in 1984, - a model for computer communication architecture, and as a framework for developing standards [Stallings 2000, p.20]. The OSI-model is organized in seven different layers, arranging the protocols with similar functions together. The TCP/IP protocol suite with only 5 layers is somewhat simpler, and has become to dominate the network standards. The layers in both models include software and hardware description of protocols that support the exchange of data, and the main layers are sketched in the figure below. The internet protocol is often placed as an own layer between the network layer and the transport layer, but since it is possible to for instance use ATM without IP in the protocol stack; the figure is sketched more in general.

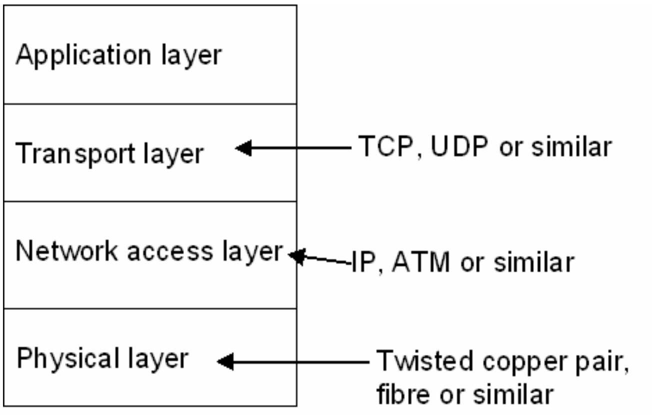


Figure 8: Layering

What is included in the different layers is shortly summarized as follows: The **Application Layer** provides communication between processes or applications on separate hosts. The **Transport Layer** provides end-to-end data transfer, establishes, manages and terminates connections when necessary, and may apply flow control and error connection (depending on the protocol used). The **Network Layer** is responsible for providing (error free) transmission, concerned with connections; establishment, management and termination, routing control, and providing a logical interface between an end system and a network. This somehow depends on the protocol being used, and is more applicable to ATM than IP. The **Physical Layer** defines characteristics of the transmission medium, signaling rate and signal encoding scheme.

The figure below shows a different perspective to the layered protocol stack. Here we for instance see that ATM is placed in the lower layer (the TCP/IP-stack's network access layer), below the IP-layer. In case of using the same layering structure as in figure 8, there would have to be two network access layers. This is because ATM can be used as a transport protocol only, e.g. for IP, or as a protocol at the network layer, controlling connection establishing/disconnection, forwarding issues etc. See chapter 4.3.

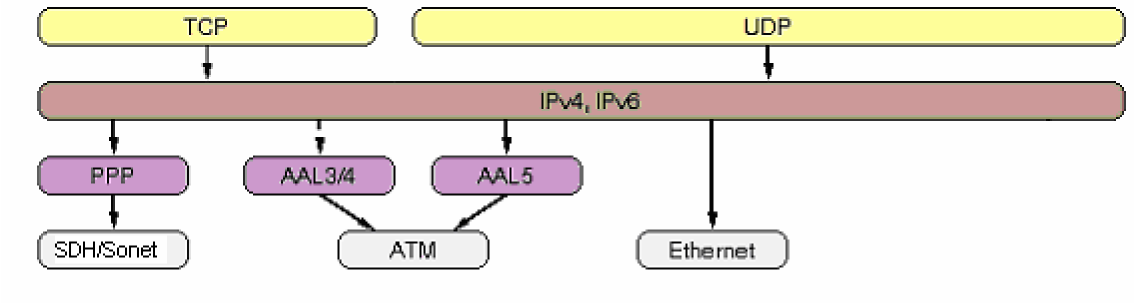


Figure 9: Protocol stack

4.2.1: TCP vs. UDP

A transport protocol provides an end-to-end transfer service that shields upper layer protocols from the details of the intervening network(s) [Stallings 2000, p.608]. The most commonly used transport protocols are Transport Control Protocol (TCP) and User Datagram Protocol (UDP). TCP is more used for common internet applications like FTP, e-mail etc, while UDP on the other hand is widely used for real-time communication (see chapter 5).

TCP is connection oriented, meaning every TCP connection is set up only if both end-points agree to set up the connection. During a connection set up, the participants agree on window size (how many segments/octets can be sent before the source receives acknowledgement) and the maximum segment size (MSS – dependent of the network MTU[7], e.g. 1500 bytes in an Ethernet network). The MSS defines how many octets that may be sent at once and should be chosen so that a packet/segment needs not be fragmented again between source and destination; - not too big. A segment should neither be too small because all the packets are sent including both an IP and a TCP header of 40 byte, meaning a small payload may lead to bad network utilization.

TCP uses a sliding window algorithm, meaning the window size may change dynamically during a session. Whenever the destination sends an ACKnowledgement for a number of segments (ACKs are not sent for every segment received!), it piggy backs a window advertisement telling the source how many octets it is able to accept, - specifying its current buffer size. When the buffer is full, incoming packets are discarded and thus demanding retransmission. Since the receiver notifies the acceptable segment size, the extra network load as described may be avoided. By applying this scheme, TCP provides end-to-end flow control, guaranteeing reliable delivery [Comer 2000, p. 220].

UDP is connection less, meaning it provides the same unreliable datagram service as IP (see chapter 4.2.3). The protocol does not provide any acknowledgement for the packets received, no feedback to control the flow of segments (packet rate), and does not order the incoming packets. UDP messages can therefore be lost, duplicated or arrive out of order. However, this does reduce the overhead of the protocol, and may be adequate in many cases.

When an application creates a datagram, and sends it by UDP, a port number must be assigned to the connection. A number of predefined port numbers are available for different kinds of services, else available port numbers are assigned by the network. It is important that the port number being used is known by both the sender and the receiver, while the port number of any received datagram is checked and validated if the port is currently in use. If not the user sends an “ICMP[8] port unreachable” error message to the network and discards the datagram, or else it queues it at the correct port. Notice that the buffer size/queue at each port is limited and that the packets may be discarded if the queue is full.

The advantage of using UDP is that it has a small header (8 byte) opposed to TCP which has a 20 byte header; it does not waste time establishing and terminating connections, and does not waste transmission capacity like retransmission etc in TCP. It is therefore most commonly used for delay sensitive applications such as real-time transmission.

[7] Maximum Transfer Unit (MTU) is the capacity over a link; how many resources available. The network MTU means the maximum resource capacity available for all the links between a source and its destination. If a packet arrives at a link with lower capacity, it must be fragmented, requiring router resources and time, which may affect the packet delivery.

[8] ICMP is a protocol used to handle errors and control messages in IP-based networks.

4.2.3: Internet Protocol (IP)

The Internet Protocol (IP) is defined as unreliable (no guaranteed delivery), best-effort, and connection less (each packet is treated independently, meaning packets from the same stream may be sent over different paths, some lost, others delivered) [Comer 2000, p.97]. IP defines a basic unit of data transfer, specifies the data format of the packets, perform routing functions; - choosing over which paths data should be sent, and characterizes a set of rules by which packets should be handled. The rules define how hosts/routers should process a packet, how and when to generate error messages and the conditions under which packets can be discarded.

Delivery failure in an IP-based network due to e.g. congestion, may cause discarded packets either because the TTL-field[9] is decremented to zero, the router's buffers are full etc. Internet Control Message Protocol (ICMP) is used to report errors and/or provide information about unexpected circumstances in the network(s), sending error and control messages as normal IP-packets. Notice that since ICMP-messages are sent by IP, these messages may also be lost and/or discarded. As a rule to prevent ICMP-messages from increasing the network resource load, ICMP-errors are not generated for faults/errors in other ICMP-messages. A router may also generate ICMP *source quench* messages to make the source reduce its rate of transmission, *redirect* messages to make the source change its routing table etc.

In today's internet IP version 4 is used, but current work is being done to implement IP version 6. A main reason for evolving IPv4 was the address lengths; which in IPv4 are 16 bits and in IPv6 128 bit, meaning IPv4 can supply 2^{16} addresses and IPv6 2^{128} addresses. More addresses were needed because all devices communicating with the internet needs a unique IP-address, and since the rate of devices connecting to the internet was, and still is, increasing, the need for more addresses is increasing as well. Another advantage using IPv6 is its built-in functions for QoS-handling. But when IPv6 will become a common implementation rather than IPv4 is still discussed, because among other things different methods have made the IPv4-address-"room" (the 2^{32} bits addresses) longer lasting, e.g. by using NATs[10], and because currently QoS-schemes are working fine with IPv4 (see chapter 6).

4.3: Asynchronous Transfer Mode (ATM)

> Asynchronous Transfer Mode is a "streamlined protocol with minimal error and flow control capabilities" [Stallings 2000, p.349], thus reducing the need for overhead when processing ATM cells and enabling higher data rates.

ATM is connection oriented and may be used both with local and wide area networks. It permits data switching at high speed (giga-terabits), but is less used because the infrastructures consist more commonly of IP-switches and routers, and costs of laying new ATM-switch infrastructure may be undesirable.

The main advantage of using ATM, except for the high-speed switching, is its ability to handle many different types of traffic simultaneously; real-time and non-real time.

[9] Time To Live is a field in the IP-header, and is used to prevent packets existing eternally in a network (internet) due to routing loops etc. TTL is decremented by at least one at each router/host every time the packet is processed. When TTL=0 the packet is discarded, and the source notified.

[10] Network Address Translation (NAT) works as a program or a piece of hardware that converts the IP address from a private address to a public address. This allows multiple users to share a single public IP address.

ATM provides minimal error and/or flow control, reducing the overhead required, and thus enabling higher data speeds.

ATM operates with a fixed cell size at 53 bytes, where 5 of them are a header and the latter 48 are payload. Because the cell size is so small, the cells require less processing and, thus, less switching time between participants. Use of small cells may also reduce queuing delay for a high-priority cell, especially because the switching is more efficiently.

ATM establishes virtual channel connections (VCCs) from one end to the other which enables exchange of variable-rate and full-duplex flows of ATM-cells. Logical connections as a VCC can be either permanent (over weeks or years) or switched (set up for every session and terminated afterwards), and must be set up with all switching terms agreed upon before a session can begin. VCCs can be set up between end users for various types of sessions, user-to-network for control signaling, and/or network-to-network for network management and routing.

To make the switching easier, the overhead less, and network management less expensive, a Virtual Path Connection (VPC) may be created in advance of VCCs. A VPC is typically a bundle of VCCs with the same conditions, between the same end-points. It means that network management actions can be applied to a small number of groups of connections instead of a large number of individual connections [Stalling 2000, p.350]. If an already existing VPC is created between end-users, the processing and connection setup for VCCs is quicker, thus adding up to the high-speed transmission.

ATM connections can be compared to PSTN telephone calls in the way it is set up as a reserved connection between the user and the receiver over what may seem as a dedicated path, but since the path is virtual, the data streams may take different ways through the network and the resources reserved for the VC may be shared. A VC does therefore not hold a link with a fixed bandwidth through the network at all times like a PSTN session would do.

All virtual connections are set up first when a number of terms are agreed with the network. The source may specify demands for quality of service, including cell loss ratio, cell delay variation etc. Other traffic parameters are negotiated for every VC and monitored by the network. (See details in chapter 6.5) Every packet flow belonging to either a VCC or a VPC is marked by a unique identifier to control which flow each packet belongs to. The identifier is assigned by the network which also informs the host in the network which identifiers are currently in use. A cell is switched through the network according to these identifiers, and needs not carry source or destination address. Each VC also supports DiffServ (see chapter 6.1) and the possibility to set different traffic parameters, which, all in all, means ATM is ideal to transmit data of various priorities (both real-time and non-real-time traffic).

4.3.1: ATM Service Categories

Because ATM networks are designed to support many different types of data traffic simultaneously, a number of service categories are defined to help the network treat data differently according to traffic requirements. Real-time applications require timeliness and low delay, while other applications require low cell loss but do not care too much if the delay varies.

The constant bit rate (CBR) service is commonly used for uncompressed audio and video information, like video conferencing, interactive video (telephone…), audio and video distribution and retrieval etc. Real-time variable bit rate (rt-VBR) is intended for time-sensitive applications requiring constrained delay and delay variations. rt-VBR transmits data at a rate that varies with time, meaning it is more flexible than CBR. Both of these service categories are used for real-time services.

The bit rate categories used for non-real-time traffic are non-real-time variable bit rate (nrt-VBR), unspecified bit rate (UBR) and available bit rate (ABR). The first category is used to apply QoS in areas of loss and delay in networks, being able to handle the expected traffic flow etc, typically used for applications with critical response-time requirements. The latter two are typically used with TCP-based applications that can tolerate variable delays and some cell losses. UBR is similar to best-effort, while ABR implements some kind of guarantees regarding throughput.

Table 2: ATM service categories

Service category	**Area of usage**
Constant bit rate (CBR)	Uncompressed audio and video information
Real-time variable bit rate (rt-VBR)	Time-sensitive applications
Non real-time variable bit rate (nrt-VBR)	Applications with critical response-time requirements
Unspecified bit rate (UBR)	Applications which tolerate variable delay and cell loss
Available bit rate (ABR)	Applications which tolerate variable delay and cell loss

4.3.2: ATM Adaptation Layer (AAL)

Computers attached to an ATM network interact with it through service-dependent ATM adaptation layers (AAL), which among other things perform error detection and correction for cells that are lost or corrupted. Whenever a connection is set up, the participants need to agree on which AAL to use; - it can not be changed during the session.

Different kinds of AAL support different types of transmission protocols, for instance Pulse-Code Modulation (PCM), voice and IP. What kind of adaptation layer is being used depends on the transmission and data type, and the different bit-rate classes described above. AAL1 is typically used for constant bit rate applications (e.g. TDM voice). AAL2 supports VBR for analog applications like video or audio that requires timing information, but does not rely on constant bit rates. AAL3/4 is no longer in use because of its complexity and advanced overhead mechanisms.

Today AAL5 is the adaptation layer most commonly used, being able to transport all sorts of data. It is a simplified version of AAL3/4, reducing its protocol processing and transmission overhead, and ensures ATM's adaptability to existing transport protocols. It provides streamlined transport facilities for higher-layer protocols.

IP uses AAL5 to send bigger amounts of data over ATM. AAL5 presents an interface that accepts and delivers large, variable length packets (1-65535 bytes). It adds an 8 byte trailer behind the payload that includes payload length, checksum etc, and divides the packet into 48 byte cells which are forwarded over ATM. In the other end AAL5 reassembles the cells and performs the CRC check. A single bit in the ATM header marks the final cell of an AAL5 block.

4.4: Synchronized Digital Hierarchy (SDH)

There are two transmission structures used to carry payload in ATM's physical layer; cell-based and SDH-based layer. Cell-based physical layer provides a continuous 53 bytes cell stream which is synchronized by using fields in the ATM-header. To hold the synchronization as long as the connection is active, empty cells are generated and sent if no payload is available at the time. The advantage of using cell-based physical layer is its simplicity when transfer and transmission functions are based on common structure [Stalling 2000, p.360].

Synchronous Digital Hierarchy (SDH) was published by ITU-T in Recommendation G.707, and is according to Wavetek's Wandel and Golterman [2] the "most suitable technology for backbones". This is because the SDH multiplexing technology and the built in functions include automatic back-up and repair to the network. Not to mention the fact that SDH uses the optical fiber network between routers or switches.

SDH is compatible with many technologies, and IP as well as ATM and other technologies may be sent over SDH. ATM is used to illustrate some of SDH's advantages, because SDH is a very common way to switch ATM-data flows.

ATM Containers (collection of data) are mapped into so-called STM-1 modules which provide a bit rate of 155Mbps. Although STM-1 is the most normal way to switch ATM data cells it is also possible for SDH to make use of higher STM-modules by using specific multiplexing techniques and thus make use of higher data rates. For instance one can combine 4 ATM streams (STM-1) to build a 622 Mbps (STM-4) interface. STM-64 can switch data at 10 Gbps speeds.

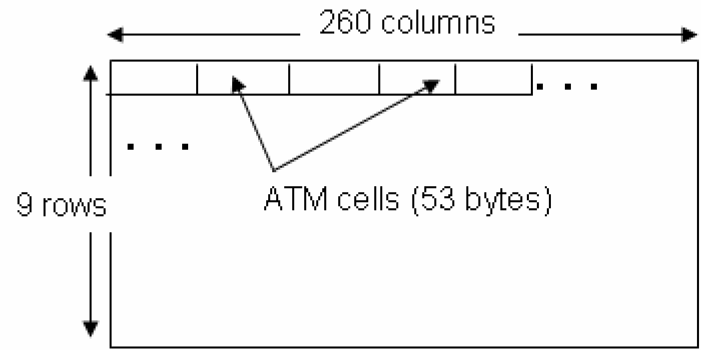


Figure 10: STM-1 Container

The advantages of using SDH are big. For instance SDH can carry both ATM and/or STM based payloads, supporting both circuit and packet switched technology. Another advantage is that several ATM streams may be combined as described above (STM-n), to build interfaces with higher bit rates than supported by the ATM (transport) layer at that particular site.

SDH networks permit different protocols to be transported simultaneously, so that logical-layer- or transport-layer-switched IP traffic and ATM traffic can be sent through the SONET[11] or SDH network on one wavelength, while time-critical and

[11] SONET is similar to SDH, but whereas SDH is used in Europe, SONET is mostly used in USA, Canada and Japan, operating on somewhat different bit rates.

unswitched traffic such as live television can be sent on other wavelengths. An SDH-based switching network also provides perfectly-synchronized and transparent transport of audio, video and all of the other types of signals needed in television broadcast contribution, thus enabling improved business processes for these industries.

DSL, outlined in chapter 4.5, is normally, but not necessarily, sent over ATM switching networks, using SDH. The alternatives to using ATM over SDH as transmission medium for DSL connections are TDM- or IP-based networks. Using Time Division Multiplexing (TDM) though, requires expensive equipment to multiplex the data streams up to a sufficient bit rate level, such as STM-64. IP-traffic is often encapsulated into ATM-cells and then sent through the network over SDH, but it is also possible to use IP over SDH without ATM-encapsulation, or IP directly over Ethernet not using SDH. See chapter 6.3.1.

4.5: Transmission technology in the physical layer

A number of different transmission technologies exist, but since most of NGN's future service will depend on high speed and wideband applications, the issues described will be broadband technology. DSL and cable access to the internet are two of the most common broadband access services being used, but wireless connections such as WLANs and Bluetooth are becoming more and more popular, especially because of IP hotspots available at airports etc.

4.5.1: Digital Subscriber Line (xDSL)

Digital Subscriber Line is a service which provides the user with broadband access to the network he/she is connected to.

Digital Subscriber Line is a transmission technology making use of the existing copper pair[12] which leads to nearly all households in Norway. Providing a possibly very high bandwidth, depending on the type of DSL used and the distance between the subscriber and the exchange office/repeater, xDSL is the most used technology for service providers with already existing copper-based infrastructure. Not only does DSL enable use of data over high bandwidth, it also enables an increased number of speech lines (VoDSL), making the technology popular among service providers and users. The speech quality experienced with ATM-based xDSL access is nearly as good as PSTN based speech.

Asynchronous DSL (ADSL) typically provides a subscriber with downstream speeds at (theoretically) 1.5-8 Mbps converting the existing twisted pair telephone lines into access paths for multimedia and high-speed communication [14]. Upstream speeds are significantly lower. ADSL enables voice and high-speed data to be sent simultaneously over the existing telephone line.

CPE at the user's site (DSL-modem etc) usually communicates with a digital carrier loop (DCL) at the local exchange office (LEX). DCL terminates the copper pair using a digital subscriber line access multiplex (DSLAM) switch and statistical multiplexing to forward/switch the data on to the transmission infrastructure network (see figure 31 in chapter 9.2.4). DSLAM enables several DSL lines to interconnect to reach the high-speed needed for the switching through the internets backbone. The DSL-

[12] Copper wires were originally the medium used for telephone connections (which they still are), but new technology such as DSL have discovered advantages for different use. The wires are referred to as copper pair, twisted pair etc.

modem incorporates forward error correction and symbol-to-symbol error correction to reduce errors caused by impulse noise and continuous noise coupled into a line. Chapter 9.2 shows an example of a service delivered by ADSL.

Very high-speed DSL (VDSL) can support downstream data transmission up to 55 Mbps, but only for very short distances (300 meters). VDSL is currently being researched and developed by several standardizations organizations, to be able to create the final drop for the full service network architecture (NGN). In addition to ultra high speed data transmissions, VDSL can for instance support provision of multiple TV-channels within apartment blocks, video conferencing, and video and data on the same line.

Symmetrical alternatives include HDSL (high data rate DSL), SDSL (symmetrical DSL) and SHDSL (symmetrical high-bit rate DSL). All provide data access up to 2.3 Mbps both upstream and downstream, although not supporting separate telephone service lines. SHDSL typically offers voice over DSL (VoDSL) instead, but the technology does still not support as good quality as experienced in PSTN.

4.5.2: Cable access

Using cable access as broadband transmission medium is a quite new technology. Originally designed to support one-way distribution of TV-signals, extra equipment was needed to support the wanted full-duplex communication channel. Usually providing higher downstream than upstream speeds, cable access is ideal for data transmission, but service providers are experimenting on speech over cable access as well.

The main difference between DSL and cable access is that DSL provides a dedicated service over a single telephone line, while cable offer a dedicated service over shared media [14]. Cable access can all-in-all provide higher data rates than ADSL (up to 30 Mbps), but because of the shared media technology, the bandwidth is shared among the users on a line, thus decreasing the bandwidth according to the number of users connected.

4.5.3: Wireless access

Wireless access has lately become more and more popular in the LAN market because it is often easier and quicker to apply in places lacking of physical infrastructure (cable/DSL…). Wireless LANs (WLAN) are therefore being used in environments where the existing cable/wire infrastructure can not be used or where it may be impossible to lay new cable. Other advantages of using WLANs are the low entry and deployment costs, the fast realization of revenue as a result of the rapid deployment, and the cost-effective network maintenance, management and operating costs. WLAN hotspots are becoming more and more normal to provide the user with wireless internet access at train stations, airports, cafés etc.

WLAN’s distribution system is typically a wired LAN backbone, such as Ethernet [Stallings 2000, p. 450], supporting servers, workstations, bridges and/or routers to link the WLAN to other networks. A control module connected in the wired LAN acts as an interface to WLAN by using bridge/router functionality. Several control modules may be required to cover wider areas.

There are three main WLAN technologies being used; infrared, spread spectrum and radio (microwave). The infrared LANs have a possible data rate ranging up to 10 Mbps, but since the waves can not penetrate walls, it is somehow limited in space.

Spread spectrum LANs may handle data rates up to 20 Mbps (1-3 Mbps using frequency hopping and 2-20 Mbps using direct sequence), ranging 100-800 ft, and microwave LANs range 10-20 Mbps, ranging 10-20ft, both using the 2.4 GHz ISM band [Stallings 2000, p.456].

A set of WLAN standards has been developed by the IEEE 802.11 committee. **IEEE 802.11** is the IEEE standard for wireless local area network interoperability. It defines a method for creating a wireless network based on Ethernet, where the physical transmission occurs through a wireless transceiver at the unlicensed 2.4GHz ISM band instead of a wire [26]. Other wireless communication standards are LMDS and Bluetooth.

Bluetooth is an open specification for seamless wireless short-range communication of data and voice between both mobile and stationary devices. It specifies for instance how mobile phones, computers and PDAs interconnect with each other, with computers, and with office or home phones. The first generation of Bluetooth permits exchange of data up to a rate of 1 Mbps [25]. It transmits and receives via a short-range radio link using a globally available frequency band; 2.4 GHz ISM band.

Local Multipoint Distribution Service (LMDS) is a broadband wireless point-to-multipoint communication system ranging up to 10 km, depending on modulation and transmission techniques, weather conditions in the area of the WLAN, the heights of base station and customer's antennas etc. LMDS is designed to provide digital two-way voice, data, Internet, and video services.

4.6: Optical networks

Optical networking was originally developed as a way to squeeze more bandwidth out of the fiber infrastructure, but service providers now see it as means to create new and flexible services, and creating or maintaining competitive advantages, in addition to reducing network costs. An all-optical network, meaning the data does not undergo any optical-electrical-optical conversions, is considered a next generation technology, and many network operators are experimenting with this.

WDM is a type of multiplexing developed for use on optical fiber[13]. It modulates each of several data streams onto a different part of the light spectrum, similar to frequencies in FDM. WDM's capability to transmit data at high bit rates (terabits per second), makes it a natural choice for future backbones.

Using *Dense* Wavelength Division Multiplexing (D-WDM) it is possible to combine multiple single-mode optical fibers, which each can manage a bit rate up to 40 Gbps per wavelength. Predicting the improvement of such technology, the possible bandwidth to reach may seem to be "indefinitely", and speeds at 400Gbps have already been measured [23]. All that is needed is an all-optical network.

One of DWDM's key advantages is that it is protocol and bit rate independent, meaning it can transmit data in IP-, ATM-, SONET-/SDH- and Ethernet-based networks, using different bit rates and handling different types of traffic over various speeds. Other advantages are of course the high switching speed, the simple "routing", etc.

GMPLS is a technology taking advantage of DWDM. See chapter 6.3.1.

[13] Note that optical networks can use other alternatives than WDM.

5. Real-time over Packet

As mentioned in previous chapters, realizing NGN means sending all kind of data communication over packet switched networks, but since packet switched networks were not originally intended to carry voice or other real-time traffic, a few problems may occur: Packet switched networks normally send data as best effort traffic, meaning packets can be lost, delayed etc. When sending real-time traffic, especially voice, over packet, timeliness is an important issue to prevent a telephone conversation from being less understandable. Today's PSTN is very well equipped regarding low delay, and since the data is circuit switched, packet loss is not an issue. But packet switching real-time data is still a trend.

A number of solutions have been proposed to solve the “problems” when sending data flows over packet switched networks. The figure below shows some of them, and how they may be implemented on IP.

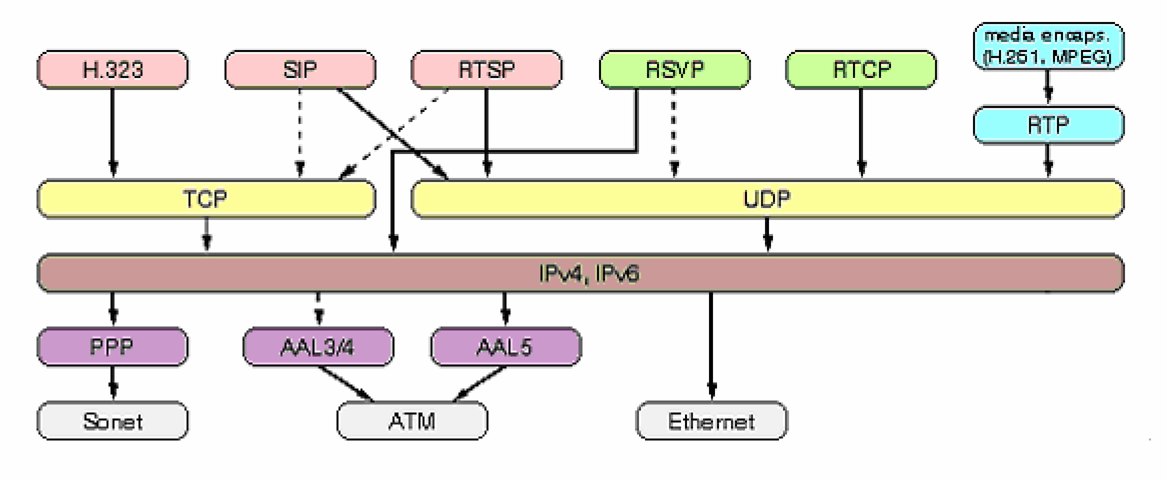


Figure 11: Protocol-"zoo" [27]

In this chapter a few of these protocols are outlined, describing how they in the best possible way can treat data, especially real-time data, and provide the same quality in a packet switched network as we experience in PSTN.

5.1: Real-time Transfer Protocol (RTP)

Common “problems” in packet switched networks are packet loss and various delay (jitter). Sending real-time data across the internet or any other packet switched network requires means to make sure the timeliness is held in one way or the other. Timeliness can be described as data presented in the exact same order as it was originally sent, with the exact same timing. According to Douglas E. Comer timeliness is more important than reliability; missing data is merely skipped [Comer 2000, p.540].

RTP is a protocol defined to transfer real-time data traffic across packet switched networks. It provides measures to prevent various delays and jitter when real-time traffic reaches the receiver, and handle problems with duplication and out-of-order delivery. As the name indicates, RTP may be seen as a transport protocol, but RTP-packets are in fact dependent of UDP encapsulation before the packets are forwarded on to a network. (Read more about UDP in chapter 4.2.1.) UDP encapsulates the payload including a specific RTP-header. In this context it is important to note that RTP does not use a specific UDP-port number, but establishes one for each session. It is therefore necessary to notify remote applications of which port number is being used.

The two main facilities in the RTP-header are a sequence number; - which is chosen at random for the first packet and incremented as more packets to the same stream are generated, and a timestamp; indicating the time when the first octet of digitized data was sampled. The header contains a few fixed fields. In addition to the timestamp and the sequence number, the payload type is worth mentioning; which is used to indicate how the routers/hosts should interpret the remainder of the header and the payload. These header-fields are followed by specific fields for whichever payload type is being sent.

RTP supports so-called “translation” (changing encodings according to application) and “mixing” (combining RTP-streams from multiple sources into one). The latter is typically used in conference calls etc, using multicasting. Instead of sending unicast messages to every participant, using severe bandwidth, RTP mixing allows the packets to be sent as one packet at each link. Multicasting typically creates a forwarding tree and sends only one packet over each branch. If there is more than one branch leading from a connection cross point, the packet is duplicated and sent on each link. If there are more than one source sending data, RTP mixing may be used to combine the data flows and send them as normal multicasting. Header fields in RTP keep track of the original sources, but use the mixer’s ID as source identifier.

5.1.1: RTP Control Protocol (RTCP)

A companion control of RTP is named RTCP and enables the participants in RTP-streams to communicate out of band. This communication consists of additional information about the data being sent and the network performance. A receiver’s report may typically consist of information about packet loss, including the sequence number of last received packet etc, allowing the sender and receiver to adapt to bandwidth changes and congestion. A “BYE”-message is used to shut down a stream.

5.2: H.323

H.323 is per definition a standard that specifies the components, protocols and procedures that provide multimedia communication services over packet networks [8]. The standard was originally developed by ITU-T to support voice transmission over LANs, but has been extended to comprise voice over packet in networks in general (e.g. internet, WAN etc).

H.323 is really a protocol stack allowing many different communication networks like ISDN, IP networks and different types of LANs to work together, although it is normally referred to as *a* protocol. It provides the necessary means for setting up voice sessions, like IP telephony, including phone registration, signaling, real-time data encoding and transfer, and control functions.

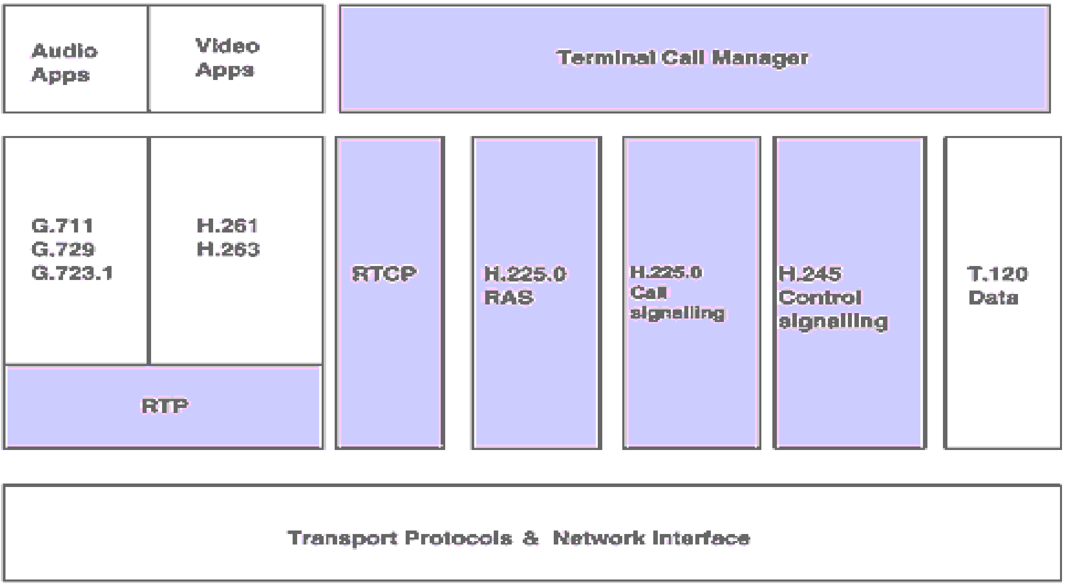


Figure 12: H.323 protocol stack [8]

All the protocols included in the stack are independent of transport and network protocols, which mean interworking with nearly any kind of network topology is possible. Primarily used as a signaling protocol, H.323 is used for session initiation and termination. When a session is initiated, a *gatekeeper* is typically used as a connection point between an H.323 terminal and the network. Using gatekeepers are optional, but recommended to networks containing IP-telephony gateways, to among other things, be able to transport E.164[14] telephone addresses. If the network consists of gatekeepers, participants have to register with a gatekeeper to be able to set up a telephone session. The gatekeeper provides services such as addressing, authorization and authentication of the terminals, billing and charging services, and call-routing services including bandwidth management [Comer 2000, p. 547].

If the session is set up between terminals of different network types (for example between an IP phone in the internet and an ISDN phone) a gateway is the intermediate connection between them; “translating” signals for call setup and release, conversion of media formats, and transfer of information between them. This type of gateway is often referred to as a media gateway (MGW) which is controlled by a media gateway control protocol (MGCP, Megaco or similar). These protocols provide detailed features of how to handle terminations[15] and communication between different network types, e.g. circuit and packet switched, and how to deal with different types of signaling, e.g. SS#7 in PSTN. Chapter 9.1 shows an example of a service delivered by H.323.

Other features you can see outlined in the figure above are audio and video codecs. These are necessary, supporting multimedia transmission, to encode the audio/video signal from the microphone/camera at the sender’s end, and to decode them in the other end. H.225 and H.245 include features similar to SS#7 in PSTN; registration, admission and status (RAS) between endpoints (terminals/gateways) and

[14] An E.164 telephone address is what we usually think as a telephone number. In Norway this is typically an 8 digit number. Number series with 5 (e.g. 05000 – Telenor customer service) or 3 digits (e.g. 110, 113…) are also considered E.164 addresses.

[15] A termination is a point of entry and/or exit of media flows relative to an MGW [RFC 2805].

gatekeepers, connection establishing (call signaling), and control signaling like opening and closing channels, flow-control and capabilities exchange.

5.3: Session Initiation Protocol (SIP)

SIP is an IETF standard and may be seen as an alternative to ITU's H.323 although it does not provide all the same features. The protocol provides signaling features, but does not specify any codecs or require the use of RTP [Comer 2000, p.548]. What SIP actually does is establishing (IP-) addresses and port numbers at which end systems can send and receive data. The intelligence and state is left at end-points, meaning the network is left rather unchanged except for routing issues.

SIP is a text-based protocol, setting up sessions using sip-addresses; any kind of URL (e-mail addresses, http, H.323, E.164 etc), for example "SIP:user@home.com". A session behaves like a client-server connection, where an end-user initiates contact with a *user agent server* within the SIP-phone. It assigns an identifier to the user (e.g. 22334455@home.com) who now can receive incoming calls. The intermediate server (between two or more SIP-phones) handles call setup and call forwarding. Some of SIP's message formats which are used for call setup and termination are summarized below:

Table 3: SIP messages [29]

INVITE initiates sessions (session description is included in message body)
re-INVITEs are used to change session state
ACK confirms session establishment – can only be used with INVITE
BYE terminates sessions
CANCEL cancels a pending INVITE
OPTIONS capability inquiry
REGISTER binds a permanent address to current location; may convey user data

To provide information about the call, SIP uses a companion protocol; Session Description Protocol (SDP) [Comer 2000, p.548]. It provides the users with information about which media type to use (codecs, sampling rate…), destination address and port number, session names and purposes etc. Using SDP is especially important to use in conference calls because participants may join and leave the conference dynamically.

SIP is less defined than H.323, but has a greater scalability making it easier to apply to big networks like the internet. H.323 is more mature, but very complex, including its own protocol stack, thus lacking of flexibility. Although H.323 uses binary coding, it may be easier to extend, but SIP is easier to read for a person since it is text based, thus more preferred among programmers and developers. IEEE foresees a coexistence of both protocols, but stresses the importance of interworking between them.

6. QoS mechanisms

In packet-based networks the most common way to transmit data is by using best effort transmission techniques, which means, in short, the data finds its way through the network by calculating shortest path between A and B at every node/router. All data is fragmented into packets (cells in ATM), and because of the best effort service, packets belonging to the same dataflow may take different paths between sender and receiver. At every node the packets traverse through the network, the packets are placed in buffer-queues and handled by the router/switch when the capacity is available, thus creating limited control whether the packets arrive in the correct sequence or at the right timeframe (various delay/jitter). Packet delay and loss is a major downside of transmitting data over packet-switched networks.

Sending real-time data or similar traffic with higher priority than best-effort (meaning separate buffer queues are implemented for different priorities), the demands to transmission quality increases. Among other things demands of bandwidth-"guarantee" will amplify. It is not acceptable that real-time traffic like Voice over Packet (VoP) is bursty and/or has other sound disturbances. In circuit-switched networks like the PSTN, the speech quality is superb according to telecom standards. Is it possible to transform a packet-switched network to give the same performance? Quality of Service is by far the best way to secure sufficient quality.

The details of how QoS may be used are outlined below. Furthermore, other quality of service mechanisms, such as traffic engineering and security issues, will be discussed.

6.1: Quality of Service (QoS)

> Quality of service is defined as "*the degree of conformance of the service delivered to a user by a provider, with an agreement between them*" [P806-GI], or as "*the collective effect of service performance which determine the satisfaction of a user of the service*" [E.800].

Historically, IP-based networks have been able to provide simple best-effort delivery service to all applications [Stallings 2000, p. 570], but the users' needs have changed; demanding better service for real-time, multimedia and multicast applications. To be able to support the variety of traffic demands, different ways of implementing Quality of Service have been researched. A QoS mechanism should be able to translate a request for service into traffic characteristics such as throughput, delay, jitter, loss and error rates. These apply especially to IP-networks, but can in general be implemented in other kinds of networks (like ATM, PSTN etc) as well.

6.1.1: Integrated Services (IntServ) and RSVP

Integrated Services implemented in network nodes makes routers able to deal with packets arriving, without concern for the type of application and/or whether the packet is part of a large dataflow or a small one. IntServ's goals are

(1) to supply efficient internet support for applications which require service guarantees,

(2) to fulfill demands of multipoint, real-time applications and large group communication,

(3) to support large-scale video conferences, and

(4) to use the existing IP for real-time data [RFC 1633].

It relies upon traditional datagram forwarding as a default, but allows sources and receivers to establish packet classification and forwarding state for each node on the path between them.

There are a few pre-proposed defined traffic classes;

- guaranteed service (absolute delay guarantees),
- predictive service (fairly reliable, delay not based on worst case scenario),
- link sharing service (resources are shared between different types of data/service),
- “pre-emptable” packet service (not-so-important packets are marked to be discarded when congestion occurs), and
- best-effort service (no guarantees), which provides the data flows with different routing conditions.

Typically, IntServ proposes to make sure that the resources needed are available when a source/host wants to send data by using a resource reservation protocol. While virtual circuit models such as B-ISDN[16] set up circuits with certain service attributes, identified by VC identifiers carried by every packet belong to a flow, IP-based networks use Resource ReSerVation Protocol (RSVP) to allocate resources to data flows (especially IP-multicast flows). The resource paths are set up when an endpoint first sends a PATH message determining the path to the destination (e.g. a multicast-address). When a reply is received, the end-user may send RESV messages towards the source requesting specific QoS terms. Each node along the path needs to determine whether it can meet the requirements (has sufficient resources) or to deny the transmission at those terms. In the latter case, the receiver may choose to send a new RESV message with lower demands.

Because RSVP flows are simplex (one-way communication only), separate paths must be set up in different directions to simulate duplex connections. They must also be periodically refreshed by the receivers, sending refresh messages upstream. This is called soft state because the connection times out. The refresh messages are e.g. generated automatically when a path must change due to link failure or congestion.

As opposed to hard state, where the connections are stable until the participants in the streaming disconnects, soft state connection may require extra resources in the router protocols to keep track of soft state connections and whether the connection reserved is active or not, e.g. extra timers. This, although, may depend on the circumstances in which the connections are setup; for instance sending short SMS-messages it could be more effective (better resource utilization) to use soft state, and “evoking” the connection when necessary, than having a hard state connection setup and disconnected.

To incorporate IntServ in the current internet, a few mechanisms must be installed, making the routers able to handle IntServ-data flows;

(1) a packet scheduler to organize buffers (being able to handle more than one queue, different priorities, equalizing jitter and/or latency variations etc),

[16] B-ISDN is a signalling protocol for ATM, similar to RSVP in IP. See ITU-T Rec. I.150 and I.221.

(2) a classifier to set the correct priority to the data flows arriving to the network,

(3) some kind of management control to make sure the resources asked for actually can be allocated properly, and

(4) some kind of reservation protocol (e.g. RSVP).

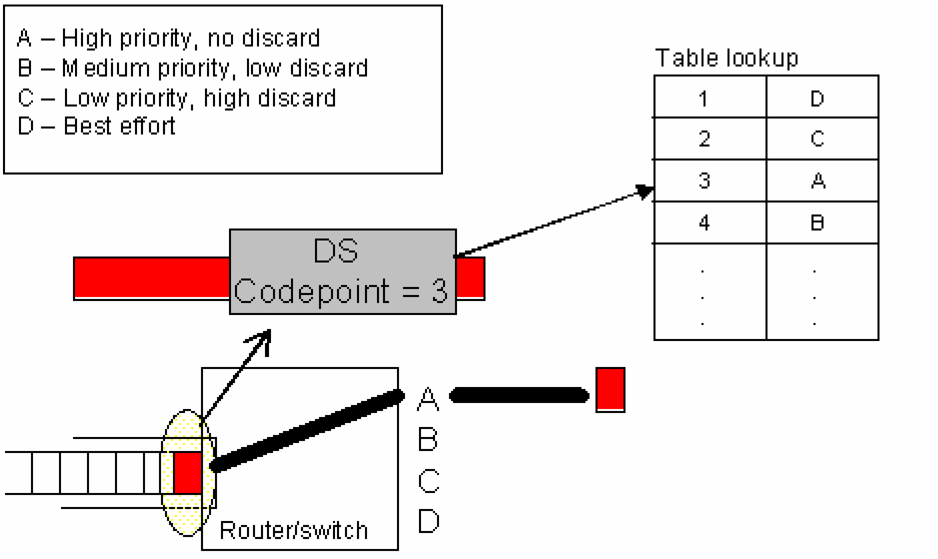
This shows that IntServ is a quite complex solution to provide QoS for different data flows.

6.1.2: Differentiated Services (DiffServ)

IntServ and RSVP are useful tools in regards to supporting QoS-bounds to data flows, but they are quite complex to deploy. They may not scale well to handle large volumes of traffic because the amount of the overhead (especially set-up signaling) required and the maintenance of state information required at the routers [Stallings 2000, p.598]. Differentiated Services is designed to provide a simpler approach which is easier to implement and requires lower overhead.

DiffServ classify packets by pre-defined priority classes (determined for instance in an SLA) and mark them to receive a particular per-hop forwarding behavior on nodes along their path. More sophisticated classification, marking, policing, and shaping operations are only implemented in network boundaries or hosts [RFC 2475]. Making use of the Type of Service (ToS) field placed in the IPv4 header and the Traffic Class field in IPv6 (hereby called the DS-field) implementing DiffServ does not require big changes in the network's packet handling.

An SLA between service provider and user may classify the priority and other QoS-demands prior to the use of DiffServ, which avoids the need to incorporate the understanding of DiffServ in applications. At routers all traffic with the same DS-field value will be buffered and treated equally, thus providing good scaling for larger networks and traffic loads. Each packet is individually dealt with, - reading the DS-field, and the routers need not save state-information about each flow.


A – High priority, no discard
B – Medium priority, low discard
C – Low priority, high discard
D – Best effort
DS
Codepoint = 3
Table lookup
Router/switch
A
B
C
D

1	D
2	C
3	A
4	B
. . .	. . .

Figure 13: Packet handling using DiffServ

All routers are required to implement at least two priority schemes; one for normal traffic and one for high-priority traffic [Comer 2000, p.100]. Packets are mapped according to the precedence field or the DS-code point field, in reference to which protocol is implemented in the IP-entity in the router. Douglas E. Comer regards the service type specification as a hint to the routing algorithm, which helps the router choose among various paths to a destination based on local policies and its knowledge of the hardware technologies available for those paths [Comer 2000, p.101]. An internet does not provide support for any particular type of service, meaning use of DiffServ-classes may not guarantee throughput for packet flows in any particular priority.

When a packet enters the network, it is classified and assigned to different traffic classes, identified by markings in the DS-field. The DS-field contains 6 bits forming a DS-code point (which priority, what QoS-limits, mark to select per-hop behavior…) and 2 bits which are currently unused. The DS-code point gives a possible number of 64 (2^6) different traffic classes. A code point set to 000000 is default and treated like best-effort traffic. Other higher-priority packets are given preference above these best-effort default packets, meaning they are processed in the router before the lower priority packets.

Most IPv4-routers are programmed to use a so-called precedence marking, using the ToS-field somewhat differently than DiffServ. The 3 first bits are used for precedence, telling the router the degree of urgency or priority to which it should handle the datagram/packet. The three next bits is a ToS-subfield, providing guidance to the IP-entity in the router/switch on selecting the next-hop. Some of these routers are compatible to DiffServ because of its legacy systems, but in case it does not, boundary routers normally change the DS-code point to match the precedence marking classes. In cases where the boundary router is not aware of what kind of compatible protocols other routers use, it may set the DS-field to the default value (000000), and hope for best-effort service.

6.2: Routing

The ideal picture of network routing is called OPT (OPTimal routing) that can direct traffic along any paths in any proportions [1]. In OPT there is no limit to what weight[17] of traffic each link can hold, but in real-life it certainly is a limit due to different traffic patterns passing by on the links. It is therefore needed to compute the weight of each link between the nodes in every network.

Requirements for routing algorithms are correctness, simplicity, robustness[18], stability, fairness, optimality[19] and efficiency [Stallings 2000, p. 315]. It means that if a routing algorithm should fulfill these requirements it must be quite complex. Data flows are normally routed by minimum-hop (number of nodes between sender and receiver), shortest path or least-cost per link. Which routing strategy being used depends on the network topology; traffic load and link cost.

[17] The weight of a link is normally calculated by measuring traffic on the link for some time, and choosing a value which describes the average traffic behaviour on that specific link. Routing tables are updated according to these values, and packets are routed by choosing the shortest path/least-cost link.

[18] Robustness is a network's ability to deliver packets via some route in the face of localized failures and overload [Stallings 2000, p. 315].

[19] The more frequent routing tables are updated, the more likely the network is to make good routing decisions, BUT transmission of update info consumes network resources, thus showing the downside of optimized routing [Stallings 2000, p. 315].

Table 4: Different routing schemes

Routing type	Description	Advantages	Disadvantages
Fixed routing	All packets from a given source to a given destination follow the same path.	Simplicity, works well in reliable networks with stable load.	Lack of flexibility, does not react to congestion or link failures.
Flooding	All packets are duplicated and sent on each link until a packet reaches destination. A unique identifier enables the designated node to discard duplicates.	All possible routes are tried; - at least one of the packets traversed the shortest path.	Generates high traffic load. May thus need high resources available.
Random routing	Packets are sent on links in a round-robin fashion, meaning each link has equal chance to transmit the traffic load.	Traffic is spread, and all links used equally, easy routing decision, does not need to establish shortest path/least cost paths.	Because of random routes, packets may travel longer paths than necessary, and the network must be able to carry higher than optimized traffic loads.
Adaptive routing	Packets are routed based on information about least-cost, failure, congestion etc in the network.	Effective routing, can aid congestion control by balancing loads.	Information about network state must be distributed among nodes, and may cause higher traffic load.

Adaptive routing is the routing scheme most used in the Internet (and other packet switched networks) today, but in a slightly different manner. A single router may use two different routing protocols simultaneously, one for communication outside its autonomous system (AS – see chapter 3), e.g. Border Gateway Protocol (BGP), and another for communication within (interior of) it's AS [Comer 2000, p. 296].

Interior Gateway Protocols (IGP) such as Open Shortest Path First (OSPF) and Intermediate System-Intermediate System (IS-IS) use adaptive routing and flooding to update what is called link-delay info to other nodes using the link-state routing protocol[20]. The BGP is used to exchange routing information between different autonomous systems.

[20] The link-state routing protocol maintains description of local link state, and distributes updates occasionally to all routers it is aware of. The update information must be acknowledged by sender.

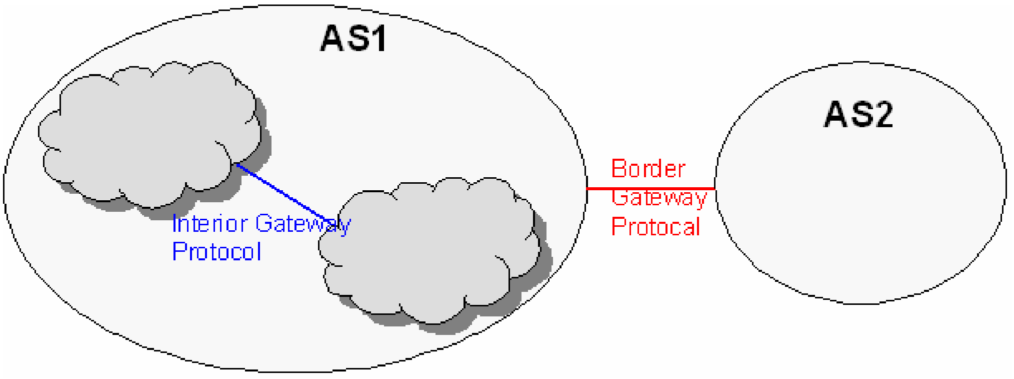


Figure 14: Illustrating IGP and BGP

OSPF and IS-IS[21] create routing tables which contain information about the shortest path from one node to another, but the algorithms do not necessarily consider the traffic conditions on each link. Instead the shortest path(s[22]) is calculated on static link weights often set (manually) by the network operator. These weights have been calculated from monitoring regular traffic in the network and seen where the needs for higher or lower bandwidth is appropriate. The problem is that the link weights often change in time variations over the hours of a day, days of week and month of year, due to link failure, congestion, bursty traffic etc. The results may show that the static set link weights not necessarily prove the correct value. It may lead to traffic congestion.

Recent standardization activities has proposed changes to OSPF to enable the protocol to set link weights according to traffic patterns in the network by incorporating the traffic load in the link-state update information, thus improving traffic engineering. By setting the link weight according to actual traffic conditions, it is guaranteed that the data flows will traverse the network in the quickest, but not necessarily shortest path available.

Fortz, Rexford and Thorup [1] propose a method to choose the link weights more closely to the optimal routing scheme, measuring and modeling network/traffic scenarios before realizing it. They claim that "good settings of the static link weights allow OSPF and IS-IS to perform almost as well as an optimal routing scheme (OPT) that has complete flexibility in selecting paths for the traffic by closer measuring and modeling the networks' performance" (traffic patterns, capacity utilization etc). Other standards propose more flexible and dynamic routing, letting the routers consider the traffic load of each link when calculating least-cost path. The latter case demands new complex process methods in the routers, and may as a consequence cause higher latency etc due to longer processing time of the packets at each node.

[21] IS-IS was designed before OSPF, but many of the IS-IS designs were adopted by OSPF, resulting in minimal difference between the two protocols.

[22] Because OSPF supports 'type of service'-routing, the routing tables may include multiple routes to a given destination, one for each type of priority or type of service [Comer 2000, p.309].

6.3: Multi-Protocol Label Switching (MPLS)

"Optimal" routing may be realized by using more flexible routing protocols such as MPLS which support explicit routes. MPLS is a switching technique developed by IETF, aimed to be

(1) layer 3[23] independent,

(2) layer 2[24] independent,

(3) independent of routing protocols such as OSPF, BGP etc,

(4) compatible with RSVP and IntServ, and

(5) work with different QoS-classification protocols [3].

MPLS is also designed so that MPLS-data flows can traverse networks which do not implement MPLS-protocol understanding, to make the technique as flexible as possible.

MPLS is used for routing, switching and forwarding packets through the next-generation network in order to meet the service demands of the network users [4]. Normally the routing decisions are only made once; at the ingress router[25]. Before the ingress router can assign a label to any data flows, the labels must be created, bound to different Forward Equivalence Classes (FEC)[26] and distributed to other nodes/routers in the network, typically using Label Distribution Protocol (LDP). A label may be uniquely assigned to a data flow either per-platform (the entire network) or per-interface. In either case, the network nodes keep an FEC-table including information how to forward the packet and which label(s) the FEC is bound to.

When a packet reaches an MPLS-environment, a normal layer 3 table lookup is conducted, finding the next-hop address. In addition, the egress router finds the FEC associated with the label. With the assigned label the packet is forwarded to the next-hop address, where the label is read and associated with a new outgoing label, including the next-hop address. The old label is used as an index into a table which specifies the next hop, and a new label. The old label is replaced with the new label, and the packet is forwarded to its next hop. It does not need to use the network-layer lookup. This process describes the advantage of MPLS; the packets need only be analyzed at the ingress and egress node, whereas all the intermediate nodes only process the packet according to the labels.

[23] Layer 3 is, referring to the OSI-stack, the network layer which is used to connect systems. It may be responsible for establishing, maintaining and terminating connections when relating to ATM, or just simple forwarding in terms of IP [Stallings 2000, p.21].

[24] Layer 2 is the data-link layer which provides for reliable transfer of information across a physical link [Stallings 2000, p. 21].

[25] The ingress router is the first router which a dataflow meets when it enters a network/AS. The outgoing router, when the packet leaves a network is typically called egress router.

[26] An FEC is a representation of a group of packets that share the same requirements for their transport. These are treated the same at each router [4].

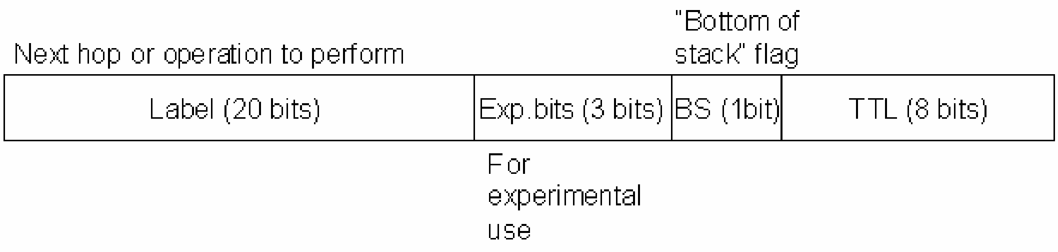


Figure 15: A normal shim header used in MPLS [4]

A normal shim header (outlined above) is typically placed between the layer 2 and layer 3 headers. In case of MPLS over ATM or Frame Relay, the label value can be extracted from the VCI/VPIs or DLCIs placed in the layer 2 header, otherwise collected from FEC-/label-tables.

The 20 bits label specifies the path a packet should traverse through the network, or which operation, or how the packet should be treated. The label(s) may specify a specific hop-by-hop route, using what is called explicit routing which is similar to source routing; listing the nodes through which the data flow should traverse. The advantage of using MPLS explicit routing over source routing is that with source routing every packet needs to carry all the addresses the packet should pass through, while in MPLS a simple label associated with the right FEC can do the same thing.

Explicit routing could be done for optimal or non-optimal routes, depending on the type of service being transmitted. It is also a difference between loose and strict explicit routing. Strict explicit routing allows the source to set the exact path the packet should follow to reach its destination. Loose explicit routing also includes a set of routers which the packet should follow, but allows multiple hops between the listed routers. The latter case is more flexible than the first because it is more compatible to network changes (failure, congestion…), and it minimizes the configuration overhead.

However, strict explicit routing applies to more stability and control requirements made by users and operators (service providers). MPLS also has the option of using constraint based routing, meaning a router may take link characteristics (bandwidth, delay etc), hop count and QoS into account when choosing routes. QoS requirements could for instance dictate which links and queuing/scheduling mechanisms to employ for the flow [4].

The 3 experimental bits are implemented for experimental and future use, for instance regarding Class of Service or Quality of Service, setting priorities etc. Using these bits to represent e.g. DiffServ code points, a number of 8 possible classes (code points) are available to treat data flows differently [5].

The BS bit is used when there is more than one label stacked. The bottommost label in a stack is set by BS=1 and the other labels are marked by 0 to show the routers how to handle the stack. In cases where more than one label is stacked (BS=0) and the next-hop address is the one of the current router, the router typically removes (*pops*) the first label which shows its first origin and switches it further on. Other possible operations are *push* (adding a new label on the top of the stack, typically used when a new control element[27] is taking action on the packet, i.e. tunneling) and

[27] A control element/plane is which kind of operation or protocol used for packet handling in a particular environment, and include unicast routing, multicast, RSVP, VPN, frame relay, ATM etc [5].

swap. The router receiving the label marked BS=1 is normally the designated end target of the packet.

The Time To Live (TTL) field is used to prevent packets from going in loops in the network. It is decremented at each node, and when the value reaches zero, the packet is discarded. Different labels in a stack may typically include different TTL-values, to guide the router (including the information in the label field) how to treat the packet.

MPLS' goal is to improve network-layer scalability, traffic engineering capabilities and price/performance. The question is whether MPLS actually lives up to the expectations. The fact is that it does not provide much faster forwarding than longest prefix match lookup forwarding [6], but it provides a simpler mechanism to forward packets which may improve on the cost/performance and time-to-market capabilities.

6.3.1: Generalized MPLS (GMPLS)

An improvement of MPLS has been developed to make the protocol more applicable, supporting time based, wavelength based and physical based switching technology. The goal achieved by developing Generalized MPLS is creating a protocol that is applicable to all service and transport traffic with full integration of all traffic types. GMPLS is also referred to as next-generation MPLS, making routing and switching more cost-efficiently and speedy.

The intention is to apply GMPLS to IP, thus enabling data flows to traverse directly over optical fiber, making use of ATM and SDH/SONET unnecessarily complex. It is possible to send GMPLS-based data streams directly over optical fiber, because of the extra MPLS-header, often referred to as the *label*, which can be interpreted by optical switches, as opposed to complex and long IP-headers which require longer processing time, thus slowing down the switching.

Although technology from the legacy systems (ATM, SDH… chapter 4) must be implemented directly in the routers and switches, the advantages are immense concerning cost-efficiency and speed. Dense Wave Division Multiplexing (DWDM), which is the method most commonly used to switch data in optical switched networks, is cost-effective and offers technical advantages like increasing the bandwidth-carrying capacity of a single fiber by creating virtual fibers, each carrying multi-gigabits of traffic per second [17].

Using optical cross connects (OXC), DWDM enables switching data streams at multi-gigabits or even terabits speed. The label stacking function “adopted” from MPLS is also an advantage, enabling interaction with devices that can only support a small label space, such as in DWDM and OXC use, where different wavelengths act as labels.

GMPLS is an extension of MPLS which implies it can be used with routers, legacy equipment like ATM-switches, SDH etc, and with newer devices such as OXCs. Using a common control plane for all types of traffic means operations and management may be simplified, thus reducing costs.

Being standardized by IETF, GMPLS includes a new Link Management Protocol (LMP), handling issues regarding link management in the optical network using photonic switches. It also includes enhancements to OSPF/IS-IS routing protocols to advertise available optical resources and to RSVP and LDP switching protocols for traffic engineering purposes that allow label switched paths to be explicitly specified across the optical core [17].

6.4: Security issues

There are also security issues to consider when sending data traffic over (public) networks, because the wrong information could easily fall into the wrong hands. William Stallings refers to the "need to secure network infrastructure from unauthorized monitoring and control of network traffic", and "securing end-to-end traffic using authentication and encryption mechanisms" [Stallings 2000, p. 677].

Various threats have been identified in public packet switched networks. The most serious ones are spoofing, where an intruder creates packets with false addresses and exploits applications which make use of address authentication and various forms of eavesdropping and packet sniffing, where attackers read transmitted information, including logon information and data base content. Other common threats are hacking, viruses and worms, denial of service (e.g. "ping of death" where the attacker sends (big) packets in indefinite loops so that the network area experiences limited capacity and severe congestion) etc.

It is important to take preventive measures against eavesdropping, illegal copying (confidentiality), unauthorized access to networks and/or terminals (authenticity), and information modification (integrity). Information in wrong hands may lead to unwanted manipulation of billing information (e.g. attacker using other user's dial-up connections for his/her own use) and blocking effects (denial of service) so that a user does not get access to various services.

6.4.1: Internet Protocol Security (IPSec)

IPSec solves some of the mentioned threats mentioned above, providing protocols and algorithms with means to authenticate and encrypt IP-packets. Given two choices; using the Authentication Header (AH) for only authenticating the packets, or using the Encapsulating Security Payload (ESP) for both authentication and encryption, IPSec gives the user a way to ensure that information does not fall into the wrong hands.

IPSec does not change parameters of the set IP-header fields, but defines separate headers added between the IP- and the TCP-/UDP-header. Working at that level means the encryption must be done for each packet, while other security protocols e.g. Secure Sockets Layer (SSL), work at higher layers and thus can encrypt an entire data stream. When SSL is being used it is up to TCP to split the stream into packets and making sure they arrive in the correct order at the other end. SSL will not be considered in detail in this chapter.

The Authentication Header is typically used where the security is less important. The ESP header is therefore more likely implemented than AH [Stallings 2000, p.678]. For instance in Virtual Private Networks (VPN) there are needs for both authentication and encryption to ensure that unauthorized users do not penetrate the VPN and that "eavesdroppers" on the internet are not able to read the information sent across the VPN-connections.

Whether the connection using IPSec is VPN or not, a Security Association (SA) must be set up between the involved parties. An SA is a one way relationship between user and receiver which identify security parameter indexes, IP destination addresses (which destinations are authorized to use the association) and security protocol identifiers defining whether the connection is using AH or ESP. This association also specifies path MTU, authentication and encryption algorithms, SA lifetime (the time the destinations agree to use the algorithm) and which protocol mode is used (transport, tunnel or wildcard, - se below) for the connection.

Several protocol modes exist in IPSec. In transport mode the user may choose either AH or ESP mode according to its needs. It is typically used for end-to-end communication using protocols such as TCP, UDP, and ICMP[28] etc. In ESP mode the payload is encrypted, but not the header. Authentication is optional. In AH mode the payload and selected portions of the header (e.g. addresses) are authenticated. Importantly, since both the senders and the destination address are authenticated, a packet sent in AH mode is prevented from spoofing.

Sending packets in different ESP modes are outlined below. Note that ESP adds a trailer after the payload including various authentication data, next header value and padding. AH mode is only recognized by the AH header between the IP and the transport protocol header.

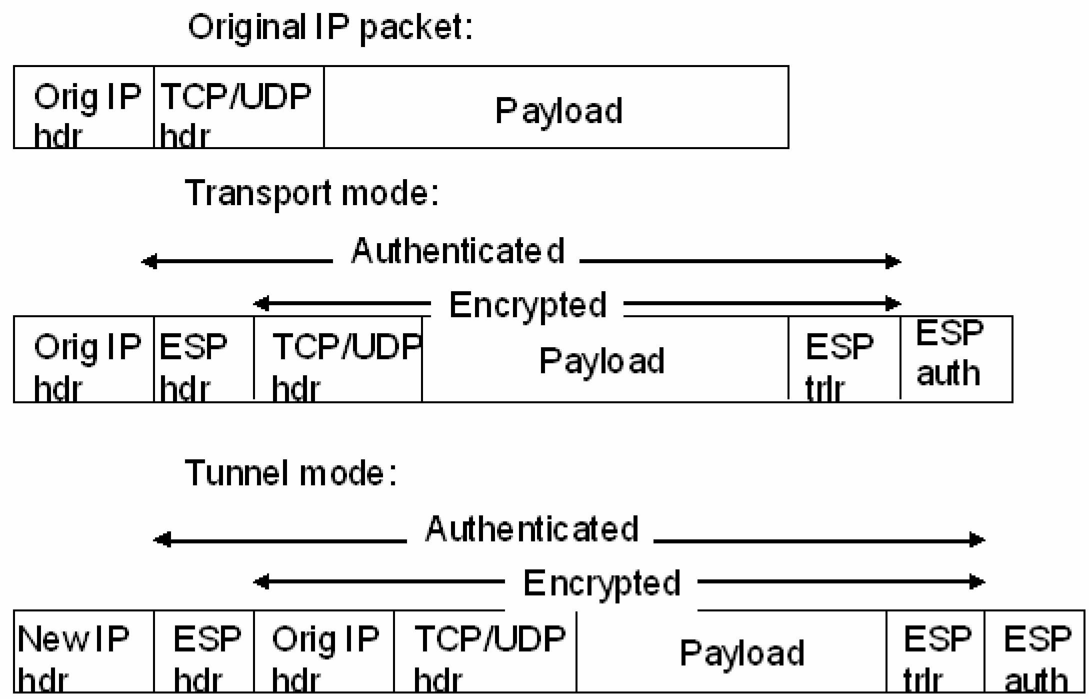


Figure 16: ESP encryption and authentication

Tunnel mode (typically VPN) allows the users to communicate on a secure connection, where unauthorized users are unable to read the header or the payload since both are encrypted and authenticated. When sending packets in tunnel mode a separate new IP header is added in front of the encrypted packet, typically including sender and destination address for which the tunnel mode is functional only. An example of tunnel mode is shown below.

[28] ICMP is a protocol used to handle errors and control messages in IP.

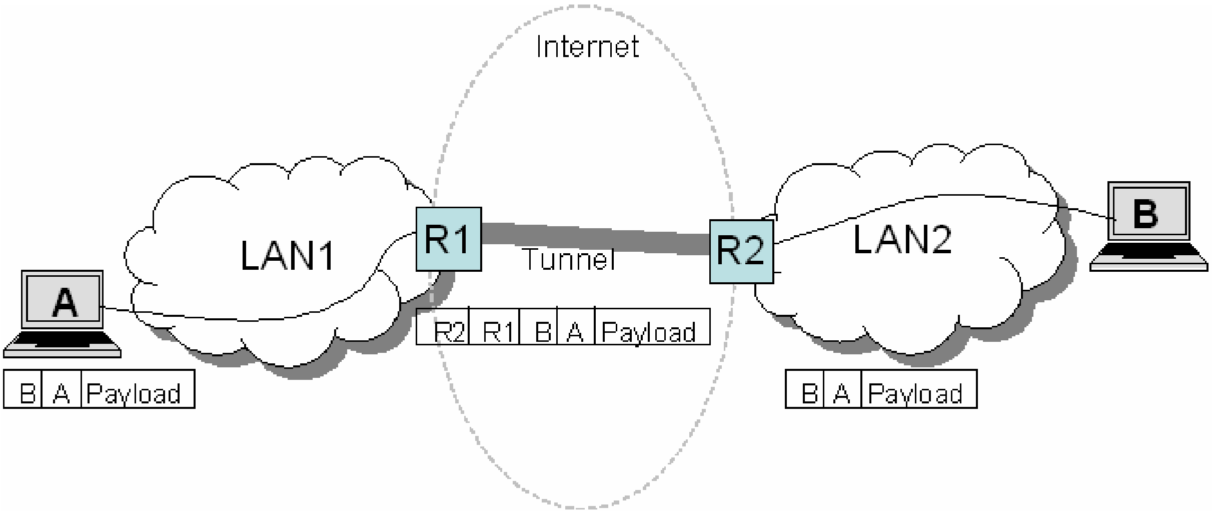


Figure 17: Showing IPSec tunnel mode

User A wants to send data to user B, but since they are connected to different networks, the data must necessarily cross the public internet. Both LAN1 and LAN 2 are secure networks, for example protected by firewalls etc, and by creating a VPN tunnel between the interconnecting routers, the line is held secure. The original packet with dest=B and sender=A is encrypted and encapsulated at R1, and a new header is added, showing the destination and sender of the tunnel only. At R2 the extra header is removed and forwarded as the original packet.

In the example it is also possible that User A encrypts and authenticates the packet before sending it, knowing its content to be confidential to all but user A and B. To be able to decrypt the packet at user B, a decryption key must be distributed. The same must be done for the process described between R1 and R2. That leads us to the third main facility in IPSec, namely the key exchange function.

The key exchange can be either manually or automatically configured. Manually configuring lets the system administrator set keys for itself and other communicating systems. Automated configuring is typically evoked on demand. This is more flexible, but requires more resources, thus inefficient for smaller installations.

Internet Key Exchange (IKE) uses a key associated with the involved parties to authenticate and establish a session key. After the exchange, the remainder of the session is cryptographically protected with the session key [7]. The IKE is parted into two phases, where the first is to authenticate the users involved. This is typically done by sharing a common “secret” (e.g. logon name and password or a pre-shared key) that can be verified by both sides, and then establishing a separate key using the Diffie-Hellman key agreement protocol[29]. This phase only happens once, while phase 2 may be repeated. In phase 2 a session key is established relying on the key shared in phase 1. Session keys may be established for each application between users, or one for all applications between a specific user pair.

Other protocols for key management are also established, like Oakley (Oakley Key Determination Protocol) and ISAKMP (Internet Security Association and Key Management Protocol). Oakley uses the Diffie-Hellman algorithm, but adds the

[29] The protocol, developed in 1976, allows two users to exchange a secret key over an insecure medium without any prior secrets using public system parameters.

means to authenticate the users. ISAKMP is similar to IKE; providing specific protocol support for negotiation of security attributes [Stallings 2000, p. 683].

6.5: Traffic Engineering

> Traffic engineering is improving user performance and making more efficient use of network resources by adapting the routing of traffic to the prevailing demands [1].

NGN is a multi-service network, meaning it supports ATM, TDM, Frame Relay etc, as well as IP traffic using e.g. MPLS. In such a network certain traffic engineering matters must be dealt with, such as adapting the routing of traffic to the network conditions, with the joint goals of good user performance and efficient use of network resources [1]. Some traffic engineering methods can be applied relating to all of today's vertical layered networks, and some apply only to specific types of networks. Throughout this chapter ATM will be used as an example, although the focus will be set on the more general functions applicable to all multi-service networks.

Considering the multiple of services provided over ATM and other multi-service networks (real-time voice/video transfer, large data flow transfer etc), tools which are used in best-effort networks to prevent congestion control are inadequate. They are not amendable to flow control (real-time data sources keep producing cells, whether the network is congested or not), can not handle the different speeds or types of bit rates used (e.g. constant bit rate vs. variable bit rate sources), or perform control due to the fact that different applications require different services (e.g. delay-sensitive service for voice/video and loss-sensitive service for other data flows). ATM-based and other multi-service networks are therefore more concerned with congestion *avoidance*, rather than congestion control.

ATM is designed to minimize the processing and transmission overhead internal to the network so that very fast cell switching and routing is possible [Stalling 2000, p. 397]. Support for a set of QoS-classes sufficient for all foreseeable network services is also required, in addition to minimizing network and end-system complexity, while maximizing network utilization. To meet these requirements ITU and ATM Forum have defined a collection of traffic and congestion control functions, determining in essence whether

(1) a new connection can be accommodated or not, and

(2) agreeing with the subscriber on the network performance parameters that will be supported.

Similar requirements are applicable to all network types, and the parameters are entered into a traffic contract (SLA – see chapter 7-9), where the subscriber agrees not to exceed the agreed traffic parameter limits, and the network provider agrees to support traffic at a certain level of performance.

The figure below illustrates the procedures being performed on traffic flows in a router for traffic engineering purposes. (Note that it does not show address lookup etc because it is not relevant to these issues.)

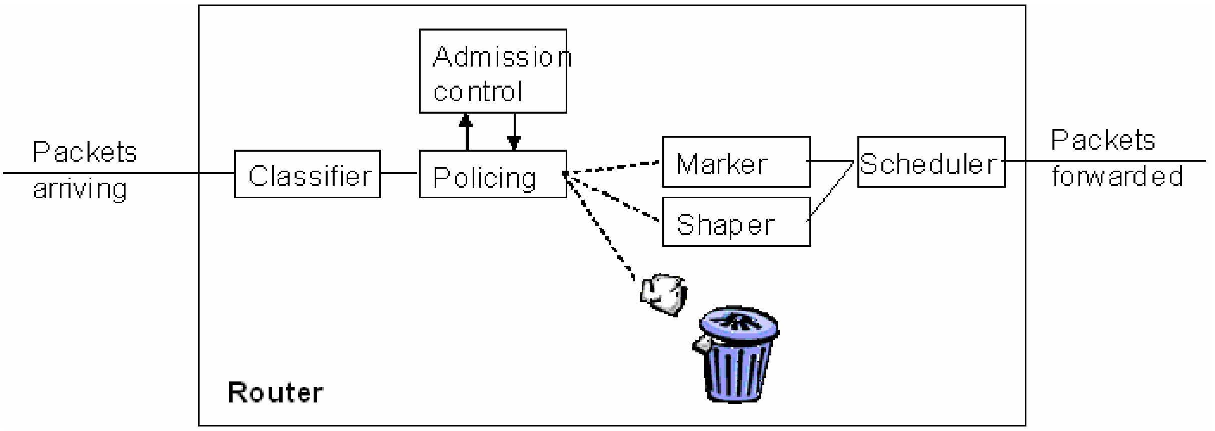


Figure 18: Traffic engineering procedure

The most important actions are described below.

6.5.1: Traffic Management

The path between a sender and its destination is often referred to as a virtual connection (VC). It is called virtual because a VC really exists of a set of links and/or several connections between each node the data streams traverse. However, since IP supports data transmission in connection-less mode as well as connection-oriented (see chapter 4.2), it is more appropriate to use the words 'traffic flow' when describing a link between sender and destination.

Several traffic flows may be handled as one if they have the same path. In ATM a collection of traffic flows are referred to as virtual path connections (VPC), but are here referred to as label switched paths (LSP) which apply to all types of packet based networks.

Managing network resources per LSP, the network resource management allocates resources so that separate traffic flows get their required performance according to service characteristics [Stallings 2000, p. 401]. The primary concerns in such a scenario are cell loss ratio, cell transfer delay and cell delay variation. Basing the resource allocation per LSP means some kind of measures must be taken to handle QoS demands for the multiple of traffic flows within an LSP. When traffic flows with different QoS requirements are bundled within an LSP, performance objectives should be agreed for the most demanding traffic flow, either by average peak rate or statistical multiplexing. Using the latter, it is preferable to bundle traffic flows with similar traffic characteristics and/or QoS requirements in the same LSP. LSPs may also be bundled within other LSPs and together with other traffic flows if necessary.

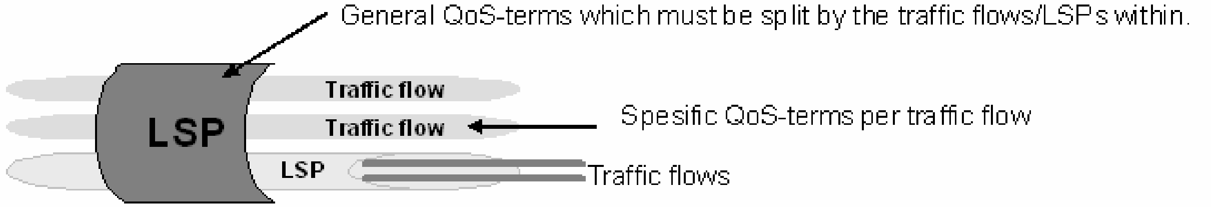

General QoS-terms which must be split by the traffic flows/LSPs within.
LSP
Traffic flow
Traffic flow
Spesific QoS-terms per traffic flow
LSP
Traffic flows

Figure 19: Handling QoS for multiple connections

6.5.2: Admission control

Providing admission control by using a Connection Admission Control (CAC) function whenever a new end-to-end connection is set up achieves lower possibility for congestion in the network. The users send a request to the network including specified traffic conditions in both directions. The traffic conditions are typically found by choosing one of several predefined QoS-classes. If there are available resources in the network requested by the user, and at the same time maintaining the agreed QoS for other connections, the request is accepted. The user and the network then form a traffic contract with the agreed traffic conditions, which the network continues to provide as long as the user meets the terms of the contract. This process must be done in real-time based on the knowledge about the traffic characteristics and QoS requirements of the new connection and of the already existing connections sharing the same network resources. Decisions are made on worst case scenarios, making sure the network can meet the traffic parameter agreements at all times.

A traffic contract typically includes peak-cell[30]-rate (PCR), cell delay variation (CDV), sustainable cell rate (SCR), and burst tolerance. The first two parameters must be specified for any connection. PCR is the maximal rate by which cells are generated by at the source, but this can be affected by CDV. CDV may cause bursty traffic, thus higher peak rate than agreed. The CAC must therefore take both parameters into consideration when allocating resources to the connection, adding a calculated tolerance limit with which the PCR may exceed. For traffic flows and/or LSPs with variable bit rates, the latter two parameters are more relevant. SCR and burst tolerance are analogous to PCR and CDV, but apply to average peak rate.

6.5.3: Policing

A Usage Parameter Control (UPC) is used to control whether the traffic characteristics in the contracts are held. It monitors every traffic flow/LSP during the active phase of transmission (e.g. a telephone session), determines if the traffic conforms to the contract, and takes necessary actions if any violations of the agreed terms occur. Where several traffic flows share an LSP, the LSP’s resources must be shared among the traffic flows. It is up to the user and the CAC whether there are enough resources in an LSP for all the traffic flows.

Deviation from a traffic agreement/contract may result from equipment failure, attempts to intentionally degrade the QoS, or be motivated by possible economic or operational advantages [11]. Actions that can be taken if the traffic agreement is violated, may be setting low priority for the non-compliant cells (marking), allowing the network to discard them if necessary (congestion), or dropping them at the point of the UPC. In accordance to networks using the negotiated cell loss priority/packet discard priority, other rules apply. (See Stallings 2000, p.406 for details.)

In networks with low utilization (e.g. during specific maintenance periods), the probability of loosing marked cells (low priority) might be considerably reduced, thus allowing the source to send data at higher values than the network has agreed to. Mechanisms for volume-based charging could therefore be applied to prevent users from taking economical advantages of such situations by intentionally exceeding the contract parameters [11].

[30] Cell ≈ packet, meaning similar terms can be defined for other packet based networks than ATM.

6.5.4: Traffic shaping

Affecting the flow of traffic traversal, traffic shaping prevents data flows coming in bursts, but rather in a continuous stream. Cell/packet delay variation is one thing that causes bursty traffic. Traffic shaping is used to smooth out a traffic flow and reduce bursty trafficking (cell clumping), thus providing fairer resource allocation and reduced average delay time [Stallings 2000, p. 406].

Traffic shaping is something that often happens in the customer premise equipment. If the policing function (UPC) is the policeman, and the charging function is the judge, then the traffic shaper is the lawyer [10]. The traffic shaper uses information about the policing and charging functions in order to change the traffic characteristics of the customer's stream to get the lowest charge or the smallest cell-loss, etc.

One way to perform traffic shaping is using the UPC algorithm *token bucket*. It is sketched below, showing the simple principle; accepting cells arriving in bursts with various delays, and forwarding them with the same interval, providing the network with a constant cell rate.

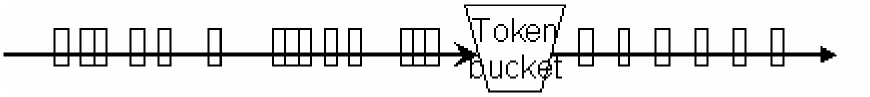


Figure 20: Illustration of token bucket

A disadvantage of this kind of traffic shaping is that it requires considerable amounts of buffer capacity, and may in time increase the end-to-end delay in the network. This is especially critical to some real-time applications.

7. Service Level Agreements (SLA)

A **service level agreement** is a formal agreement between two or more actors that is reached after a negotiating activity with the scope to assess service characteristics, responsibilities and priorities for every part. An SLA may include statements about performance, tariffing and billing, service delivery and compensation [E.860].

A so-called Service Level Agreement (SLA) is an agreement made between two or more actors/entities to guarantee certain conditions about service characteristics, responsibilities and priorities for every part involved, regardless of what kind of network the data flows shall traverse. It should cover most of the QoS mechanisms described in chapter 6, in addition to legal and economic terms.

A number of recommendations of how to use an SLA and what to include in it are made. Most referred to is ITU-T's Recommendation E.860 and ETSI's EG 202 009-3, but the Norwegian Post and Telecommunication Authority has lately released a proposal standard [12] of how to set up an SLA and what it should include. This is a suggestion which is likely to be acknowledged by actors in Norway, as a template to form other SLAs.

"An SLA is a living document and has to be updated regularly" [EG 202 009-3], meaning parameters etc agreed upon in the SLA may have to be changed or regulated. According to ITU-T Rec. E.860 the scope of an SLA is to state responsibilities of each provider and to assure QoS required from a customer(s). This is especially important in a multi-provider environment, where each entity has a number of actors to relate to.

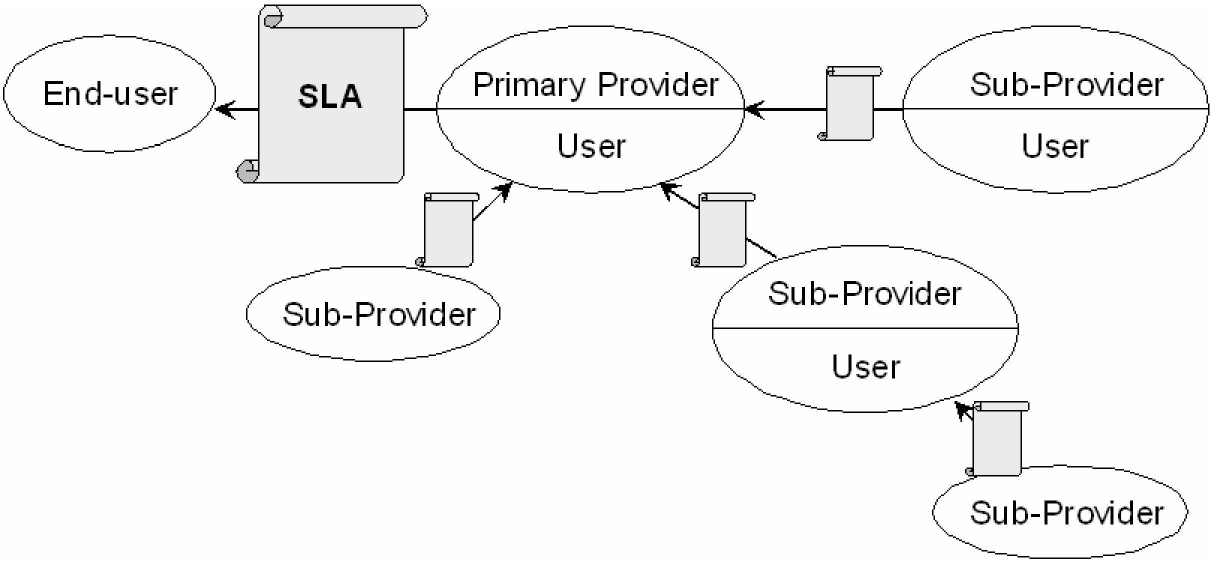


Figure 21: An example of a multi-provider environment [E.860]

It is important to standardize QoS terms and definitions in an SLA to avoid confusion introduced by contrasting terms and definitions, and to maintain the consistency between different groups involved in developing telecommunication standards [E.860]. However, QoS may be specified and perceived differently – depending on the user's conditions and service types [P806-GI], and different contents in SLAs regarding the different users are therefore required.

7.1: Service Quality Agreement

ITU-T proposes in E.801 to develop specific service quality agreements between service providers and other operating/providing actors to improve service quality world wide. They claim that since providing actors are facing increasing competitive pressure and customer driven requirements, joint service quality agreements will improve the customers' satisfaction.

The idea is to initiate a formalized program to monitor, measure and set targets that are intended to satisfy the end user and other customers. Mutually agreed action plans should be developed to improve a target that is below the expected level of performance [E.801].

A service quality agreement is very similar to an SLA, and the term SLA is therefore used throughout the thesis, even if the agreement is between two or more actors (providers), or if it is between a provider and a user.

7.2: One stop responsibility

As illustrated in figure 21, a multi-provider environment may be quite complex because primary providers may have to rely on sub-providers when delivering their services to an end-user. More SLA-scenarios are outlined in chapter 8.

In order to guarantee QoS levels stated in the SLAs, it may be important to define the responsibilities the actors involved have. The problems occurring in such situations may be "simplified" by introducing a so-called *one stop responsibility*. ITU-T proposes to apply this scheme to enable the user to hold its primary service provider, and only it, as the responsible actor for the QoS being received. The service provider, in its case, holds the next sub-provider as the only responsible. This is illustrated below.

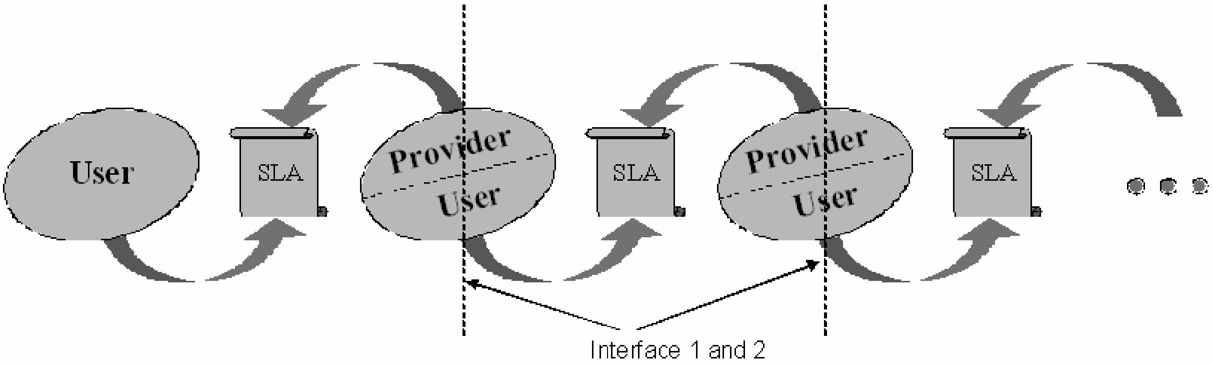


Figure 22: Illustrating one stop responsibility [E.860]

Applying this scheme recursively to all the entities involved in service provisioning, no matter what kind of service provided (infrastructure hardware, software etc), will efficiently guarantee service provision to the end user. It means the complex problem of service provisioning in a multi-provider environment is decomposed into elementary relationships between only two actors.

To explain this in regards to the figure above, you may say that any actor in the multi-provider environment can not "see" further than the interface lines. The user sees the first (primary) provider, and relates only to this, the primary provider at interface 1 relates to the user and the second (sub-) provider at interface 2, and so on.

However, an important thing to indicate using the one stop responsibility scheme is the necessity to implement QoS flexibility in the SLAs. Since the QoS elements from a sub-provider may oscillate within the agreed ranges, the QoS-elements in forwarding services may also oscillate, thus, resulting in the need to include some kind of slack to the agreed QoS-parameters in an SLA when relying on services or service elements from sub-providers. This may cause some problems. See chapter 8.2.3.

7.3: What to include in an SLA

Successful handling of the QoS involves common understanding of relevant terms and agreeing upon QoS objectives [P806-GI].

An eminent provider aims at ensuring that QoS is delivered to the user, according to the user's expectations. An SLA is therefore normally written in collaboration between the two (or more) entities involved, and acknowledges that users and service providers have responsibilities and obligations to each other [EG 202 009-3].

It is important that the conditions are expressed in a clear and convenient way with simple language for the end users, and more technical terms for the providers. Every SLA should therefore include a thorough description of the service which the SLA comprises, including

(1) scope,

(2) confidentiality (keep trade secrets, ref. market competition…),

(3) review process (defines frequency and format with which QoS information has to be exchanged) and

(4) compensations (if unreached level of quality) [E.860], and

(5) identify the critical areas of the service agreeing to a minimum level of service to provide customers satisfaction [EG 202 009-3].

A technical and a business interface are described to part the formal and the technical issues. These may include information about service delivery point, which protocol(s) should be used, measurement points, observation points, points where reaction patterns will be applied, etc [P806-GI].

The business interface (BI) located between the user and the service provider, is typically used for (re)negotiation, performance reports and reaction patterns. The Norwegian Post and Telecommunication Authority in [12] recommend the provider to mention a specific CRM or KAM to handle all customer relations over the BI due to the agreements made.

The technical interface (TI) on the other hand describes service specific exchange information, measure parameters etc regarding technical operation of the service. Including QoS parameters, traffic patterns, and reaction patterns if the terms agreed upon are not met, the SLA nearly fulfills all the necessities of a contract between user and provider. Note that the parameters stated in the SLAs are influenced by network performance and the provider's organizational structure in a more or less direct manner, and may be both direct parameters; referring to specific service elements, or indirect; referring to functions of other direct parameters.

In the technical interface it is important to define what kind of measurement scheme that should be used, and when, where and how traffic should be measured. Measurement of every interesting parameter all the time might be very expensive and can even jeopardize the network performances [EG 202 009-3]. In addition, since a service provider today normally has limited resources, monitoring every connection for every SLA is not realistic, especially providing services to the private market (it is more common in relation to business partners). It is rather normal to do random checks, perform test calls/sessions, and calculate and model traffic patterns to get the total picture of how the connections perform. Measurements should be scheduled so that it reflects the traffic variations accurately over the hours of the day, days of the week and months of the year. How and when this is done should be included in the SLA.

Many new companies, e.g. the Norwegian company BaneTele AS, provide access to GUIs on the web to allow the user(s) to follow the monitoring at all time, by graphs etc. BaneTele's "Performance Monitor"[31] provides their customers with traffic statistics and accessibility at access points towards their core and/or access network. These graphs should be a part of periodically status reports made between the user and the provider.

The most important parameter to describe a traffic relationship is Service Availability (SA) – how many % in time the service is contracted to be operational, ergo available to user. If this percentage is not met, the customer may demand some kind of reaction. Some reactions in case traffic patterns and/or QoS parameters are not fulfilled (possible service degradation), are normally automated, some are handled case individually. Reactions could be

(1) no action,

(2) to monitor the particular QoS for the effected connection,

(3) monitoring specific object where violation has been detected,

(4) traffic flow policing through traffic shaping and/or admission control,

(5) reallocating resources,

(6) warning signals ("something is wrong"-types) to user/provider when thresholds are being crossed (e.g. alarm), and/or

(7) suspending/aborting the service [E.860].

These may apply to different situations, e.g. depending of how severe the failure is. See chapters 8 and 9 for more details and examples.

7.4: How to write an SLA

It is recommended to write the SLA more or less general; applying it to a perspective of similar services and including service specific issues like QoS parameters and reaction patterns as attachments. ITU-T refers to this as a QoS-agreement[32], implying that one SLA may have many different QoS agreements. By applying the SLAs this way, the agreements become more flexible and easy to handle, especially according to different QoS-classes. A customer buying a service, e.g. data communication, may require different priorities and QoS-terms than another customer, but as all the details are left to the attachments the same SLA may be used for both customers. More details can be read in chapters 8 and 9.

[31] See www.banetele.com for more details.

[32] Note that this is similar to the service quality agreement mentioned in chapter 7.1, although the E.801 QoS-agreement is defined for contracts between service providers only, while this is defined more in general, applicable to all contracts between all types of actors.

8. Modeling SLAs

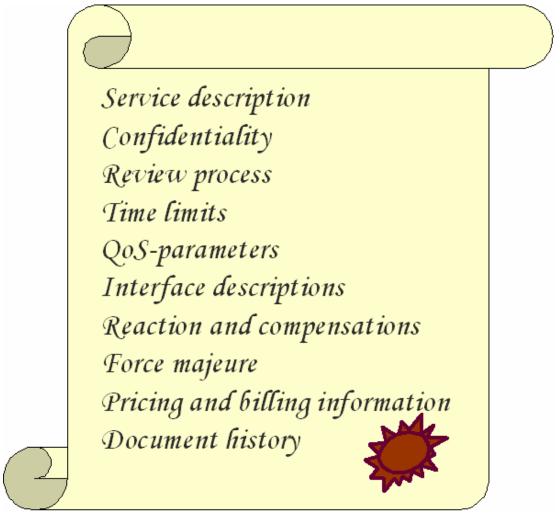
Using the theory described in the previous chapters it may be in place to make a few examples using the theory in practice. This chapter will cover the more general perspectives related to writing SLAs and to which problems may occur associated with the different scenarios, while the following chapter will include concrete examples for specific services.

8.1: SLA models

A few models are proposed by different standardization organizations to outline what an SLA should include. The model proposed by the Norwegian Post and Telecommunication Authority [12] is a good model to follow, regarding which entries are necessary to be addressed, and is very similar to the proposals made by ITU-T in Recommendation E.860 and ETSI in EG 202 009-3 and ETR 138.The main points which all the models relate to, will be referred in short in the examples in chapter 9, but the main concern will be the QoS-agreements' (the part of the SLA which comprises service specific parameters) technical demands.

An SLA should be written to form a common understanding for what the agreement comprises. This may result in higher customer satisfaction, and should help avoid misunderstandings in relation to issues regarding the SLA. The main points the agreement should include are:

- *Service definition/description*[33] – the service(s) described in detail to enlighten what the SLA includes and which issues it does not concern. It is important that this is written in clear language; so that misunderstandings do not occur; for instance in situations where the customer and the service provider expectations differ from what is really experienced. The definition should include descriptions of type of network and type of use, connection establishment, equipment provided and technical features of the various items [EG 202 009-3].



- *Confidentiality* – it is in the interest of all parties that confidential information should not be disclosed to entities which are not part of the agreement [E.860], e.g. a service provider which is a competitor in the same market.
- *Review process* – how, when and where to review the agreement's terms, including details of how to inform the parties involved about the changes.
- *Time limit* – how long before the agreement is terminated or must be renegotiated

[33] Note that the description may be changed according to the users' requirements. It is therefore (often) normal to have standard service descriptions which may be a common ground for negotiation, and change it to match the conditions asked for.

- *QoS-parameters* – described with flexible margins
- *Interface description* – TI and BI – see description in chapter 7.
- *Reactions and compensations* – what happens if the service diverges from the terms agreed upon?
- *Force majeure* – which faults or divergences the service provider and/or other involved actors are not responsible for.
- *Price and billing information* – including information if and when the service provider can adjust the prices according to market conditions and tariffs.

It is also important to keep a record of the changes made to the agreement, including what, when and who, and document every detail of the relationship between user and provider, including SLA violations, reaction patterns etc. This may show the overall relationship and how different situations are handled, and are referred in *document history*.

8.1.2: How to use the SLA information

The SLA between a user and a service provider should be written in simple language. The technical issues could be formulated so that people with less technical background may be able to understand it. This also applies to SLAs between service providers (in wholesale situations) and between service providers and network operators. If technical terms are being used in SLAs towards end-users or other actors, they should be explained and exemplified in an understandable way. E.g. response time for a TV-access service could be explained as the time from the user hits a button on his/her remote to change a channel, until the signal changes and the new channel is shown on the TV.

It is always important that all parties involved in an SLA understand the conditions agreed upon completely, so there will not be any misunderstandings or problems where the service delivered is not working as it should, and the user overlooks it. All parties are responsible to detect errors or minor faults, and to report them to the other parties involved, although this may be specified differently in an SLA. This is the reason why SLAs often are written more in general, with all the service specific conditions enclosed as attachments (e.g. QoS-agreement). The attachments may outline all the necessarily “difficult” issues, and may more easily be understood by those who have the right qualifications.

8.1.2: Monitoring

From the network operator’s and/or service provider’s point of view it may be important to monitor the traffic for specific connections, so that they know the conditions in the SLA are met. If any problems occur; for instance that the user complains about what he/she believes is a service failure, the network operator/service provider has documentation of what happened during the situation, and the problem may be solved (if there is a problem that is).

Because of the traffic monitoring, problems should be early detected and quickly dealt with, and the user may not even discover them. However, not all connections are necessarily monitored at all time, and that may cause problems in similar situations. It is therefore necessary to include details about the monitoring in the SLA so the parties involved are aware of which methods are being used and what is being measured and monitored. Details could be

(1) the identification of relevant measurement points,

(2) specification of measurement environment,

(3) description of the technique(s) for obtaining measured values,

(4) specification of the methodology to present and evaluate the results by parameters, and

(5) the method to be used for taking decision on acceptance based on the level of compliance of the measured results with the stated requirements and commitments [P806-GI].

8.1.3: Measuring

To make sure the QoS-demands are met, it is normal to check the criteria against the reference values. This is referred to as measuring. These reference values have to be identified according to the different users' requirements. QoS is a measure for "the degree of satisfaction of a user of a service" [18], and measuring may therefore be performed either objectively (in technical means by measuring physical attributes), subjectively (by surveys and subjective tests among users), or both. Subjective measurement of customers' satisfaction is widely used to assess the psychological aspects of QoS which can hardly be measured by technical means. These may include users' opinion of speech and video quality or other services, and is often used as a reference to set the objective (technical) measurements.

Another reason to monitor traffic and to use subjective measuring methods may be to acquire knowledge of people's communication habits and needs, and discover their use of communication devices and services in everyday life. This information can be used to develop better services and devices that will really be used and utilized in real-life tasks [18]. It would also help develop devices/systems to allow people with special needs, e.g. blind or deaf persons, to communicate within the same telecommunication system as others, with special input/output devices.

8.2: SLA Scenarios

Different perspectives and complexities can be taken into consideration when modeling SLAs. A few example scenarios are outlined below in addition to the most interesting problems related to them.

Note that various customers may be categorized by different priorities, depending on the service purchased and the customer-provider relationship. Providing contact information in the SLA may for instance be categorized in different time frames due to how critical the contact reason is, and/or how "important" the customer is (how much they pay…). The table below shows an example of how this can be realized in relation to fault management.

Table 5: Example of contact-routines [18]

Escalation level	XPTO management	Fault category escalation matrix			
		Critical	Major	Minor	Non-service-affecting
0 (Point of contact)	**Network Operation Centre** (NOC) 24h/24h Tel.: +0099.123.456 Fax.: +0099.123.564 noc@xpto.89	Event	Event	Event	Event
1	During office hours **Mr. Noc** Tel.: +0099.123.654	Event + 2 hours	Event + 4 hours	Next working day	Next working day
2	During office hours **Mr. Headnoc** Tel.: +0099.124.356 Fax: +0099.124.365	Event + 4 hours	Event + 6 hours		
3	Any time **Mr. Shootingtroubles** Tel.: +0099.456.123 Fax: +0099.789.123 GSM: +0099.789.000	Event + 6 hours	Event + 8 hours		

8.2.1: A simple case

The figure below shows a simple SLA-scenario. The network operator and the service provider act as the same actor, but have different roles; the service provider delivers a service and the network operator routes it through the network to the user. From the user's perspective it has only contact with one actor with which he/she has made an agreement (SLA). The SLA states which conditions are agreed upon for this specific connection/service, written in a language understandable according to the user's knowledge level.

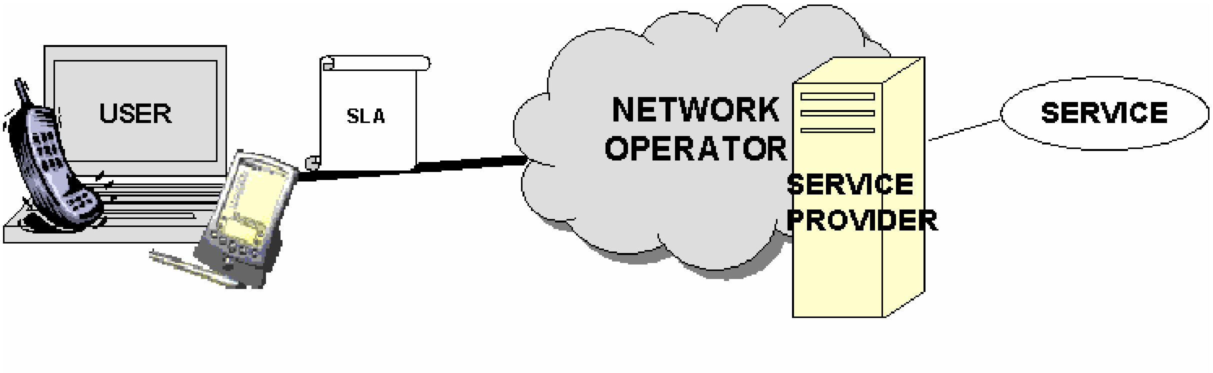


Figure 23: Scenario 1; one user relates to one actor with two roles

This scenario is quite simple and is included to make a clear example of which parameters should be included in any SLA. A similar scenario is handled in detail in chapter 9, focusing on the SLA between an end-user and the service provider. The

SLA between the network operator and the service provider is also considered, as one variant, while these are not seen as the same actor in that scenario.

8.2.2: More than one service provider – case A

The complexity increases when you add another service provider offering its services through the same network operator. If we say that the service presented by the different service providers is approximately the same, the user is allowed to choose which service provider it wishes to purchase the service from. This is more complex in real-life, because a customer normally has several service providers available; offering similar services, and he/she may invite tenders to decide which provider should be preferred. In NGN the number of service providers will increase even more, making the market competition high between providers.

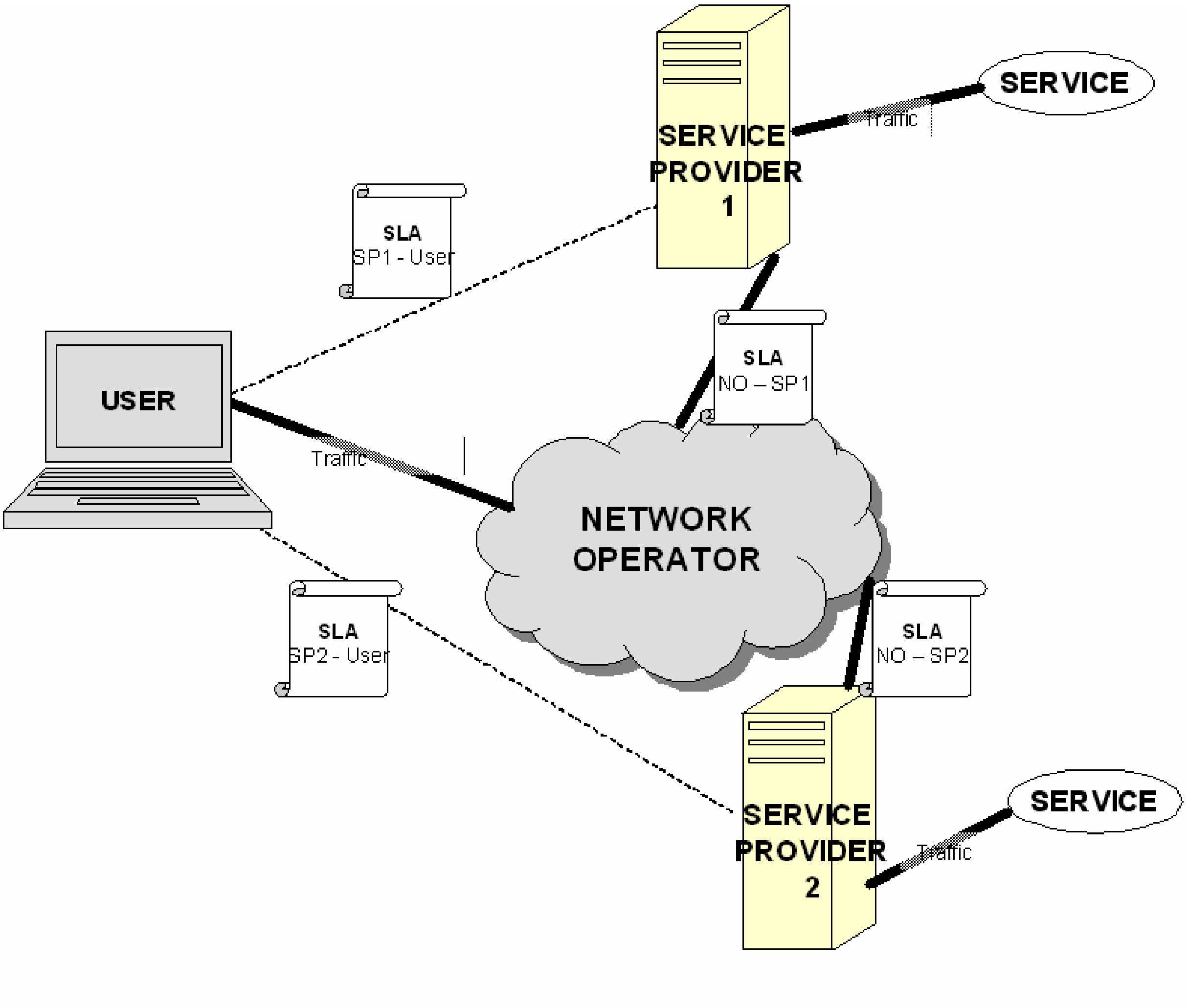


Figure 24: Scenario 2; two SPs, one NO. The user may relate to SP1, SP2 or both.

The problems arising in this scenario may be seen in two different perspectives. One perspective is if we say that, as in the first scenario (8.2.1); SP1 and the network operator is the same actor. SP2 is a separate actor providing its services through the network operator. The SLA between SP2 and the network operator is therefore the most interesting in this case, because it may vary from the more general one proposed in the previous case. It is sometimes common to prioritize the service providers within its own entity, and the costs SP2 may have to provide the service, may be higher than SP1's.

On the other hand it may be costly for the network operator to “run” SP1, and the network operator’s profit from providing its network services to SP2 may be higher than the income from SP1’s sales, thus down-prioritizing its own provider.

If we look at the possibility that SP1, SP2 and the network operator are different actors; the network operator must come to SLA-terms with both the service providers. Which provider can negotiate the best terms, is then an interesting dilemma to question.

The SLAs will not be considered in detail in the examples, because the changes are small related to what was written in relation to the first scenario examples. Although, the reader should be aware of the problems mentioned, which may imply tough negotiation between the network operator and the two service providers, which may or may not resolve in beneficial matters for the user.

8.2.3: More than one service provider – case B

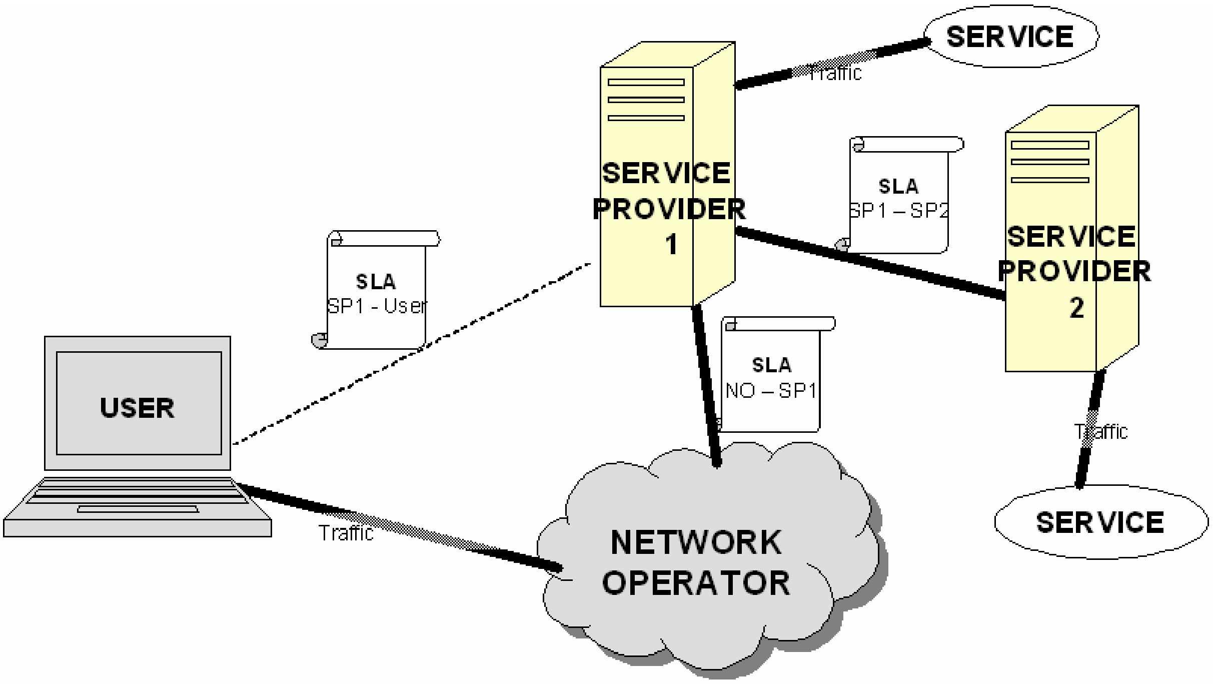


Figure 25: Scenario 3; two SPs, one NO. The user relates to SP1

The third scenario has another perspective. We see that there are still two service providers, but the relationship between them is different. They are no longer (necessarily) competitors, but “partners” in the way that SP1 provides services to the user both from SP1 and SP2 (wholesale).

The interesting perspective here is the one between the two service providers. We consider SP1 to be the user of the services purchased from SP2. SP1 combines its own service-elements with the ones purchased from SP2, and delivers it as one service in total to the user. The SLA between the two service providers must among other things reflect which is responsible if the user is not satisfied with the service; not meeting the condition agreed upon in the SP1-user SLA. Note that there may be a network(s) between the two service providers, making the issues regarding the scenario more complex.

Applying ITU-T’s one stop responsibility (see chapter 7.2); SP1 is the one responsible if the end user receives an unsatisfactory service, and must handle all problems from the SP-user SLA according to agreements made in the SP1-SP2 SLA. The problem applying this is whether or not SP1 determines good enough parameters towards the user, with the necessary "slack" and QoS flexibility the actor must relate to. Since the QoS elements from a sub-provider may oscillate within the agreed ranges, the QoS-elements in forwarding services may also oscillate.

SP1 may decide on high safety margins in its service offering, because the provider is unwilling to take the chance that SP2 holds its' part of the deal, and therefore delivering a service with rather low quality. Seeing this in a wider perspective, the "one stop responsibility" idea may not be the best solution after all.

"Delivering a service in a multi-provider environment depends on the successful operation of many components provided by other companies" [P806-GI]. When a service provider depends on several linked sub-providers, and all of them decide to implement small or big safety margins when providing services from the wholesale market, the end service provider may end up not being able to deliver the service at all due to all the previous providers' safety margins. A solution may be to change sub-providers, and try to negotiate better terms with somebody else, or try to find other arrangements. Another alternative is for the service provider to take into account the user's wishes, regarding QoS-parameters, when they (re-)negotiate terms with the sub-providers.

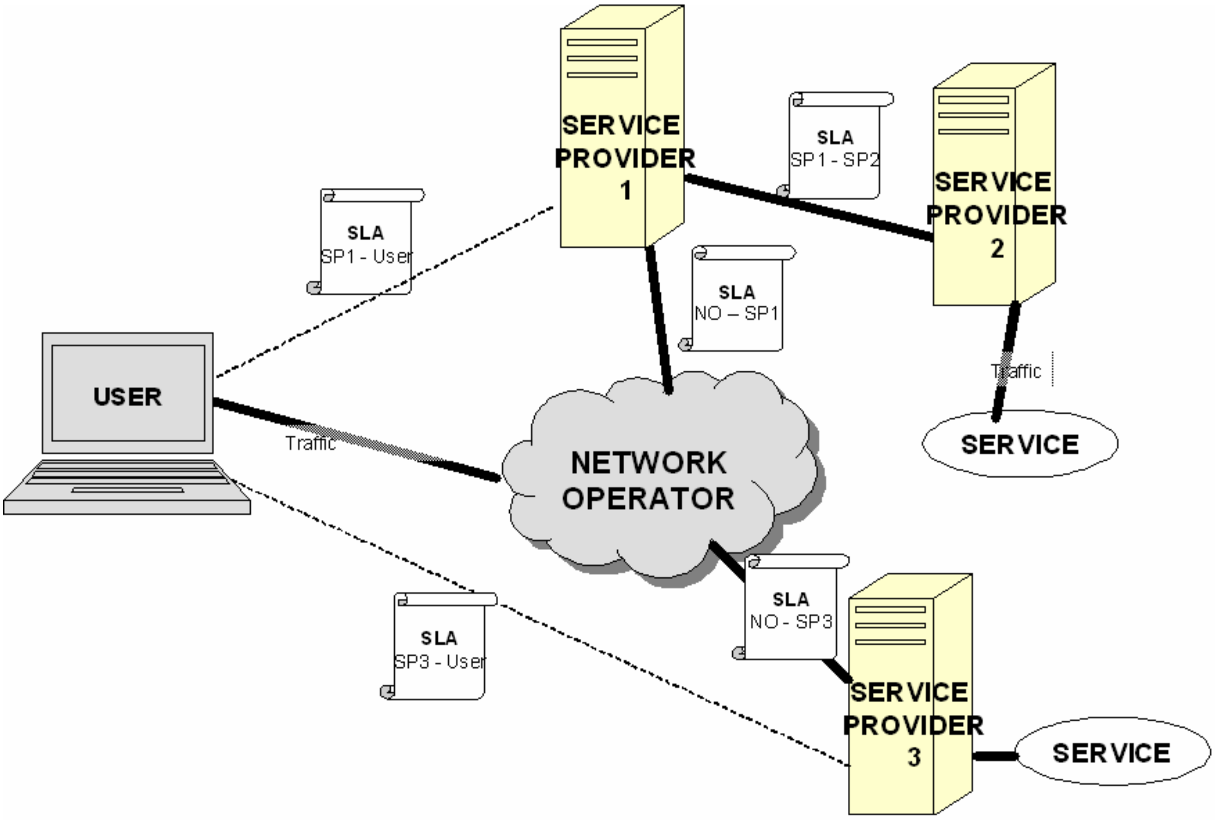


Figure 26: Complexity increases

Competition increases as more providers join in the picture, perhaps delivering its own service without depending on sub-providers. How can the service providers who depend on sub-providers (as spoken of above) compete with these? "First-hand" providers may have way higher QoS-margins to negotiate with the user and the network operator, thus creating an "unfair" advantage.

8.2.4: More than one network operator

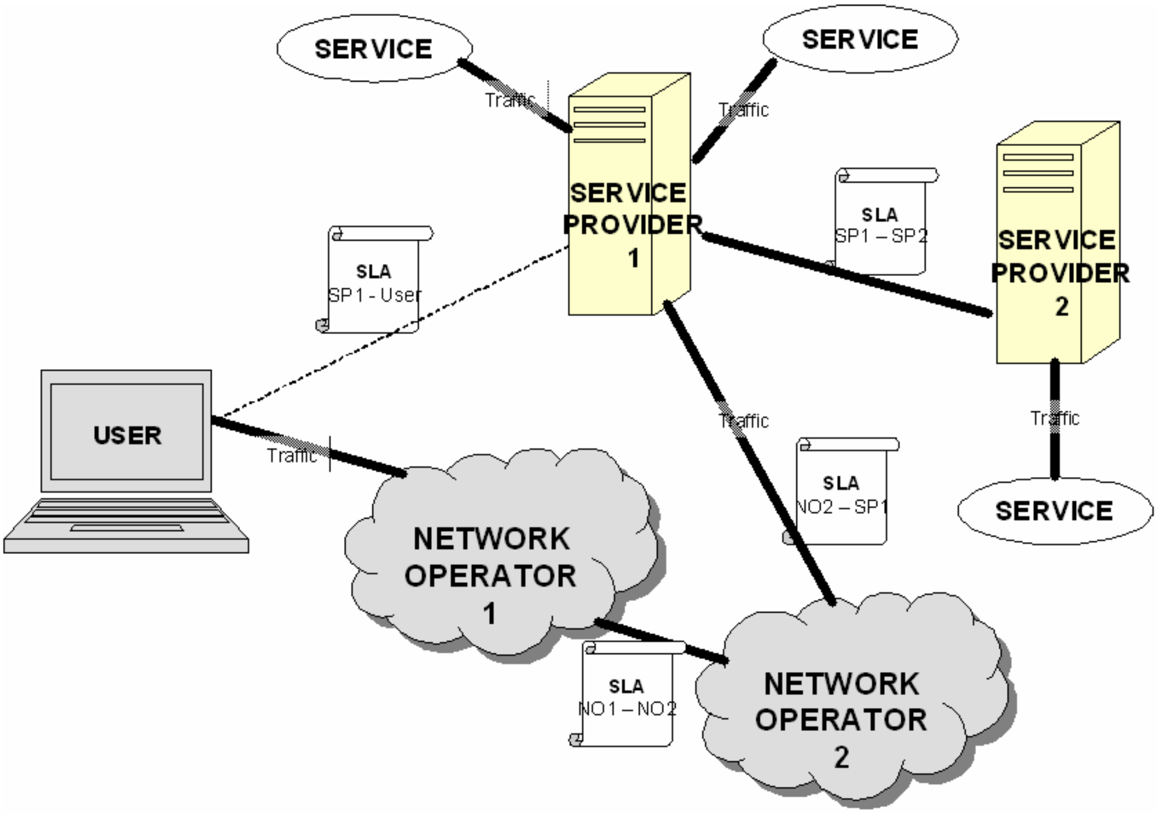


Figure 27: Scenario 4; two SPs, two NOs. The user relates to SP1.

Another point of view is when the service provider who delivers the service to the user has an SLA with a different network operator than the one the user is connected to. The service provider in this scenario has an agreement with NO2, which in its case must have an agreement with NO1 to guarantee the conditions agreed upon with SP1. How can the conditions in the SP1-user SLA be met here? Dare SP1 rely on NO2 to deliver the service safely to the user through NO1?

The NO2-SP1 SLA must include flexibility which is reflected in the NO1-NO2 SLA. The same problems as described in 8.2.3 apply here as well (ref. one stop responsibility); does the NO2 rely on NO1 and take the risks delivering the service at the same conditions, or will the operator need to take safety margins, thus providing fewer resources for the service provider's use?

8.2.4.1: Mobility

An interesting topic to discuss with this scenario is what is becoming more and more common; namely mobility. Many users have laptops which they carry around everywhere, and they may expect to access the services they subscribe to anywhere. In such a situation the service provider may not be able to communicate with the computer through which the network it has made an agreement with. The question is whether the WLAN (if we consider wireless communication, which is most common when talking about mobility) the computer is attached to has the capabilities to support the required service demands/resources. Should the SLA between a user and a service provider state location dependency, or can the SLAs be written flexible enough to justify service degradation in case the user accesses the service from a different network?

9. Parameters and examples

This chapter comprises more or less detailed description of how an SLA-agreement could look like and what it should include relating to different services. The first part outlines the possible parameters and terms of SLAs for two specific services; video conferencing and video on demand/pay per view, while the latter parts include exact examples.

9.1: Real-time (multimedia) applications (VoIP)

9.1.1: The service

The service being considered in this chapter is voice over IP (VoIP) related to video conferencing between two or more parties. A video conference or streaming is real-time (multimedia) data streaming including voice and video, and is typically performed between business partners. The SLAs may therefore contain critical conditions when seeing the perspective of the enterprise market; - if the terms agreed upon are not met or/and the video conference service is unavailable, important business contracts may not be fulfilled as needed.

The Norwegian National Bank for instance presents live press conferences over the internet, informing the nation's bankers etc about important information e.g. interest-issues. If this information is not presented at the correct time, it could cause money trouble for other banks and money instances, relying on wrong information. See also chapter 2.5 for examples of how SLA violation can cause money losses.

9.1.2: Constraints

To be able to offer a voice over real-time service, the service provider must have an agreement with a network operator which is willing to take responsibility for the demanded conditions. The user needs equipment to connect to a public network and additional CPE like video camera(s) (web cam), microphones etc, to be able to transmit the conference to other medias.

In this scenario the user is considered to be a medium business company using a private LAN within its offices. Connecting to the Internet through a proxy[34] with firewall, data transmission is monitored and security matters may be controlled. As we consider voice over IP in this context, it is expected that the user has connection to an IP-compatible network which supports real-time transmission (see chapter 5).

9.1.3: Transmission between user and service provider

In this scenario H.323 will be used as an example. H.323 is a standard that specifies the components, protocols and procedures that provide multimedia communication services over packet networks (see chapter 5.2). It requires the user to have H.323 compatible CPE, enabling registration with a gate keeper (GK) at call-setup, communicating signaling messages with a Media Gateway (MGW) by Megaco, MGCP or similar during the conference session, and disconnecting procedures after session termination.

[34] A proxy is an intermediary program that acts as both a server and a client for the purpose of making requests on behalf of other clients. Proxies are often used as client-side portals (i.e., a trusted agent that can access the Internet on the client's behalf) through the network firewall and as helper applications for handling requests via protocols not implemented by the user agent [23].

The network is a high speed all optical network using DWDM to switch data. GMPLS, described in chapter 6.3, may also be used to secure the paths the data flows traverse the networks, although the details are not interesting to how the SLA will be formulated.

Since the service being considered is video conferencing and, thus considered, a real-time application, RTP (which is regularly used with the H.323 protocol suite – chapter 5) is likely to be used. RTP is (nearly) always sent over UDP, meaning the network is unable to control packet loss, and unable to perform packet retransmissions. However, since the protocols RTP and RTCP help control these features, they are not an issue to discuss in these SLAs.

Further details about RTP can be read in chapter 5. To control the security matters, it is likely to use some kind of security protocol like IPSec, which is described in chapter 6 in addition to other QoS mechanisms.

9.1.4: Scenario details

The figure below shows the SLA-scenario considered in the example. The SLAs between the user and SP1, and the SLA between SP1 and the network operator, will be considered and possible parameters outlined.

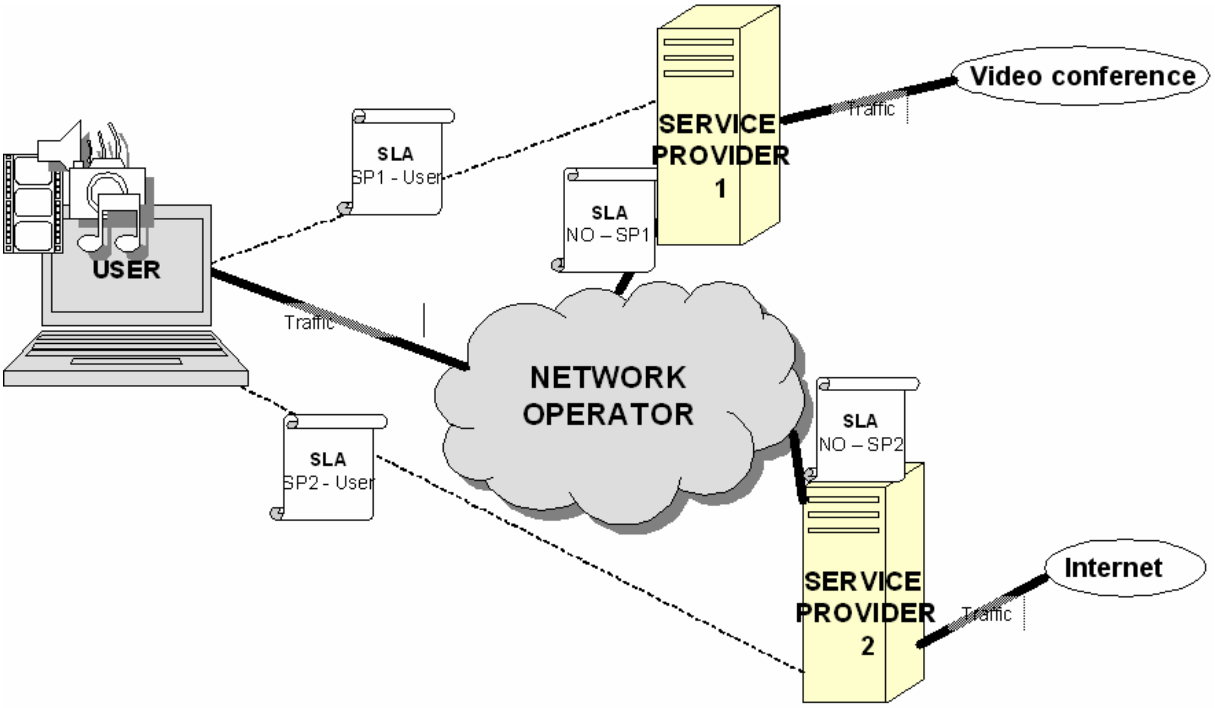


Figure 28: Video conference scenario (general)

Studying the scenario described above more carefully, we can sketch the scenario in more detail. Note that the same case may be possibly solved in several different ways (e.g. different transmission medias within the network), and that this shows the choices made in this example.

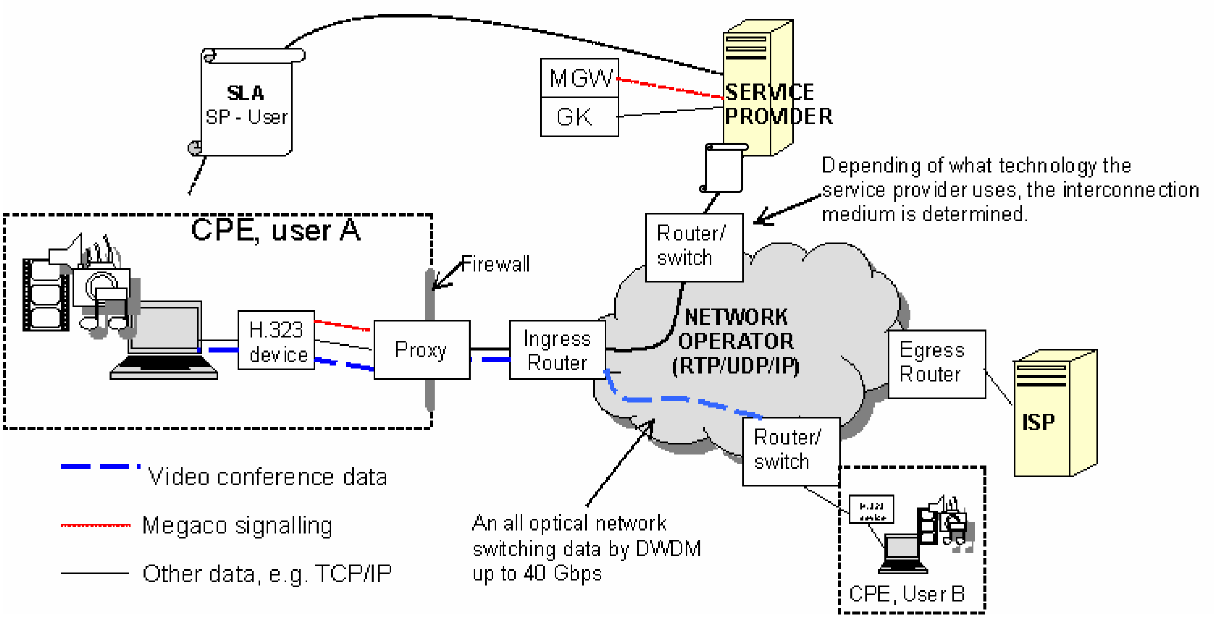


Figure 29: Video conference scenario (detailed)

An SLA is needed at literally every interconnection point in this scenario to make sure the agreements made between the actors are fulfilled. We assume that the network operator in this scenario has no directly responsibilities for other than its own interconnection devices (OXCs etc), while the service provider controls the gate keeper and the media gateway (H.323 necessary devices) accessing the network, and that the CPE is the user's responsibility. (Note that network access may be provided by a third actor.)

The interfaces being considered below are the SLAs between the service provider and the user (proxy), and between the service provider and the network operator. Similar SLAs may (necessarily) be present between the ISP and the network operator and between user B and the service provider as well, but are not the focus in the section. (Note that user B may be either a separate customer with its own agreements with SP1, or a part of a VPN connected to user A's network.) Seeing the actors' point of views, a list of conditions wanted in the SLA are outlined below.

9.1.5: SLA between User and SP

The requests made by the user may not be very technical, but nevertheless important to consider for the service provider to present, to ensure the customer's satisfaction. This section outlines a few possible parameters and explains the main points which should be included in an SLA, while there are more concrete examples with chosen values etc in chapter 9.3.

The paragraphs with indentation are meant to be examples of sentences or parameters which could be a part of an SLA in the context discussed.

Service type: Video conference

Service description: Video conference is a service delivered as a VoIP application, meaning data will be sent over an IP-based network in real-time. Video conference allows the user to arrange conferences between parties at different locations, using the video/sound equipment to send data across network(s) between the users.

The service guarantees real-time transmission, meaning continuous streaming with no flickering in picture and no sound disturbances will be experienced.

Confidentiality: Classified information or material that can be considered business secrets, must be handled confidentially; as commercially sensitive, in relation to employees of the parties involved and other parties somehow involved in the agreement's negotiation [12]. This information is exchanged between the parties involved with good faith and should, under no circumstances, be given to any other party without prior written consent from the other parties [E.801].

This may be formulated differently and depends on how the service provider sees the market competition and whether they consider their service details to be confidential or not.

Initial terms: When a customer considers acquiring a service in the first place, it is the service "life-cycle" which is reflected on. What happens before the sale is put into order (how to get in contact with the service provider), the sale itself (which terms must be agreed upon), the installation (how and when), what needs to be documented, the operation (specific conditions to make the service work as proposed, which actions must be performed by user…), the troubleshooting (what happens if there are problems/faults of the service), the billing (what and how to pay), and the evolution (technical updates etc)… The conditions below are examples of how this "life-cycle" can be mentioned in an SLA between a user and his/her service provider.

The service is implemented within x working days/hours from time of order after SLA-terms are agreed upon.

Accepted time before the service is operational through the network, measured from date of order is x working days/hours.

Service will be functional including all additional supplements from dd.mm.yyyy.

Helpdesk opening hours in case of problem, and named CRM/KAM as contact person

The name of a contact person may be left out if the customer is a private person and/or the number of employees is much smaller than the customers they are involved with, meaning the company have limited human resources and is unable to have specific contact persons for each customer. In such cases it is normally arranged for some kind of customer support, where any person answering the request from the user, should be able to deal with it.

It is always considered good service to have the name of a person which handles the deals with you/your company. In this scenario it is most likely that the customers deal with one or a few named employees with the service provider's company because it is a "business" customer. But there may be an opening for prioritization relating to contact issues, meaning a customer may pay more to have better contact conditions (see example in table 5, chapter 8).

Time, measurement and review: The duration of the agreement; how long the service will be running before the contract expires or the conditions change, must also be mentioned in the SLA. Will for instance the price be adjusted according to market tariffs? Will the customer be able to use the service until he/she tells the service provider to end it, or will the service cease to work after a period of time?

How often the SLA should be reviewed depends on who the customer is (private or enterprise), what kind of service it is, and how critical the SLA-demands are to the user. In this case, since we consider the customer to be a medium business office, the SLA requires details about the review process.

The customer may require monitoring of all the conferences they organize, at least in the beginning of the user-service provider relationship, to establish the provider's reliability, and then once a week/month or similar, depending on how often the service is used and the customer's satisfaction. The monitoring results should be included in status reports in addition to other described circumstances regarding the relationship.

The agreement itself could be reviewed once a year or more often if necessary, including renegotiation of the QoS-terms, pricing etc. Prices could be automatically adjusted according to market tariffs, or regulated as results of negotiation in a review process. The SLA should no matter what state how this is, and how it will be solved.

> The agreement exists until one of the parties ends the relationship.
>
> The agreement is renegotiated once a year; one year from SLA signed.
>
> Monitoring results will be made available to customer in reports once a month, including worst case and normal (average) scenarios.

<u>Service specific conditions</u>: After the initial terms are agreed upon; the customer may consider other conditions before acquiring the service. These conditions may be of the kind that gives the customer a basis to evaluate the service benefits according to what the other service providers offer (read: competition). The examples below are conditions made for the specified service; (real-time) video conference.

> A conference can be arranged at any time, without previously notification to provider.
>
> The number of cameras/camera angles transmitting the conference in either direction may be at a maximum xx.

The user may want to send video from different camera angles, and this condition is therefore made to limit the bandwidth needed for the transmission. If multiple microphones are used, these sound samples will be assembled and sent as one flow.

> The conference will be sent and received in real-time as a continuous stream with no flickering in the picture and no sound disturbances.

This condition gives the need for implementing maximum delay-tolerance, and that the required bandwidth to deliver the service is available through the network at any time. Acceptable "down-time" should also be agreed upon. Service Availability is measured as a percentage of time during which the service is operational [E.860].

> Service availability should be x % at any time of the day, week, year.

The percentage is typically set at 95% for telecommunication services measured over a specific period e.g. a month, but can be negotiated. The service price may depend on how low or high this is set.

Service reliability is also an issue to consider. The *availability* applies to whether the service is available when asked for. The *reliability* applies to whether the service continues to be available after initial setup or contact; meaning initiated sessions are not terminated or interrupted in any other way.

Because timeliness is more important than reliability using a real-time application, missing data is merely skipped (RTP does not support retransmission). Still, it is important to establish an agreement about the packet loss rate, meaning high priority should be set for the streaming, guaranteeing throughput and low rate packet discard opposed to other packet streams.

Other technical terms which should be handled in the SLA to fulfill the requests made are to set maximum jitter-tolerance and provide some kind of means to make sure the fragmented screening packets/cells arrive at the correct time frame (setting the right parameters for the protocols RTP, H.323 etc should provide some kind of control).

> The video conference streaming from service provider to user is secure, no chance for sniffing, spoofing or hacking.
>
> The streaming data is “clean”, - no chance of virus, worm, Trojan horse etc being transmitted.

Security can be ensured by using encryption, authentication and similar (i.e. IPSec – see chapter 6.4). The details should be outlined in attachments.

Reaction and compensation: In cases where problems and/or violations of the SLA-terms mentioned above occur, or the demands agreed upon are not met in some way, it is normal to mention some kind of action to either solve the problem or end the relationship between user and service provider. The constraints above are often named together with a predefined reaction and which resources or applications these include. These should be outlined in detail as an attachment.

There may be a few conditions important to the user when possible violations of SLA-terms occur. Important demands may be to agree on a time scheme; time event (problem) occurred, time action starts (when to react), time action end… (See the example in table 5, chapter 8).

The time between an event occurs and the reaction starts is an issue of competition between service providers, and are measured by which a customer can compare different service providers.

> Minor divergences may be solved by money compensations. (This point should be written in more detail; - depending heavily of what kind of problem is occurring, and how the time scheme is set for repair.)
>
> Service failure will be detected within x seconds/minutes/hours according to monitoring conditions.
>
> If service failure is detected the service will be fully repaired within x working days/hours from time of report (should be written in more detail depending on how severe the failure is – see table 5, chapter 8).
>
> If the conditions above are not met by any means, it is considered a contract breach, and the user has the rights to terminate the relationship to the service provider.

Force majeure: In cases where the parties are not responsible for violation of the SLA-terms, it is called force majeure. It may be nature catastrophes like earthquakes, storms etc, strikes, or other causes, in which neither of the actors involved can control.

As an example, the summer in Norway of 2003 may be mentioned. That year it was a lot of thunder and bad weather, and lightening struck many telephone lines dead. Telenor customers all over Norway complained, and some had to wait weeks before their telephone was reconnected. However, since weather conditions are outside Telenor's control, they could claim force majeure and were not responsible for e.g. the customers' economical losses or giving compensations in their advantage.

Charging/billing: The price of the service is an important condition to which the user can compare this service to similar services, offered by other service providers. In this case the price can be set as:

> xxx NOK per camera angle per conference

Pricing could also be related to the amount of time-frames or bit-quantity, meaning the user is charged for the minutes and seconds the service is used or how many bits sent during the conference. This way of pricing requires closely monitoring of all conferences, and is most likely more expensive than charging a set price for each session. The set price could also depend on how many parties participating in the conference; unicast (one-to-one) or multicast (one-to-many), relying on how the network capacity is being used.

☆

The service provider may also have some conditions which it will demand the user of the service to fulfill to be able to provide the service as agreed upon. These points will concern the customer's usage of the product rather than technical issues of how the service works.

Conditions: It is important for the service provider to meet the terms it has made with the network operator, and make sure that the customers do not abuse their available capacity. If the customer abuses the resources offered by the service provider it may lead to violation of the service provider's SLAs with the network operator, and cause more trouble than "loosing" a customer. Violation of SLAs may also lead to service degradation for other users, which in their case may be discontent and choose other ways to fulfill their service needs. The SLA should therefore include preventive measurements towards circumstances as this, and set maximum and minimum limits related to the resources available.

> The user must have connection to a network which has a minimum downlink capacity at x Mbps to allow the service provider to deliver a "perfect" service.
>
> The maximum of x simultaneously held video conference with maximum x participants per conference may be obtained without risking lack of quality.

Meaning the service provider should set a maximum number to which the customer's available bandwidth is limited by. In technical terms it means limiting the bit rate for any connection, but should be explained to the user by demonstrating decrease of quality if several conferences are performed with the same bit rate available as for one conference, meaning the quality will decrease; less pictures pr second, less pixels, or worse resolution.

9.1.6: SLA between SP and NO

To accommodate the conditions agreed upon in the SP-User SLA, the service provider also needs to have an SLA with its network operator so that it can guarantee that the conditions agreed with the user(s) are met. (This agreement may be left out if service provider and network operator is the same entity.)

Although the service being provided to the user is video conferencing, the service required between the service provider and the network operator is based on transmission and enough bandwidth to deliver the service to the service provider's customers as proposed. In other words, what the service provider really offers its customers is access provisioning through the network(s).

It can be discussed if video conferencing is a service needed to be presented by a service provider. What is really required to realize the service is guaranteed bandwidth to support real-time transmission, making it possible for the customer to make a deal directly with the network operator to fulfill his/her needs. On the other hand the network operator might not wish to have many customer-relationships, but just relate to a few service providers offering their network capacity and access for a certain price. The expenses carrying it out either way should be considered by the different actors.

In this scenario, we consider the service provider to work as an in-between entity between parties in the video conferences; making sure the set-up phase etc is done correctly. During the session, all data may be switched directly between the parties directly involved in the conference, making some kind of dedicated path, at the same time as signaling messages are handled by the service provider in relation to the network operator.

The SLA will therefore include slightly different conditions and demands than the one described above (user's perspective). Note that although the service provider has some technical knowledge, it might not know all the technical details within a network, not just because of the provider's lack of competence, but because the network operator may choose to hold back such information. It is therefore important to also here describe (technical) issues in clear understandable language.

Service type: Real-time high-speed data communication (IP)

Service description: The service gives the user the ability to transmit high amounts of data in real-time, ergo: high speed transmission. The transmission may be either unicast (one-to-one) or multicast (one-to-many).

The parts considering confidentiality and initial terms is quite similar to those mentioned in chapter 9.1.5, and will not be repeated.

Time, measurement and review: Including how often the SLA should be monitored, reviewed and renegotiated should be clearly stated. In cases such as this one, it would be likely to review the terms for instance once a month, possibly to adjust the terms to the likely expanding number of customers and thus increasing bandwidth requirements. On the other hand such an update/renegotiation should be unnecessary if the terms in the SLA are expressed well enough, with incorporated flexibility (parameter-intervals instead of set rates) to accommodate the expected expansion of customers.

To fulfill the service specific terms stated below, it is normal for the network operator to provide some kind of traffic monitoring. Measurement descriptions for monitoring should include statements for how, what, when, where and who should perform measurement procedures and test processes [E.860].

Demanding monitoring of all traffic between all users and the service provider to assure detection of failures or minor divergences of the SLA-agreed conditions may be one given condition. It could also be of interest to study how the service is being used; – to map normal and extraordinary situations/usage (studying the customers' habits). In this scenario the end user should for instance have a limit as to how many simultaneously conferences it can hold at the same time; - the capacity needed for x number of connections would be too high when supporting more than one customer.

Traffic monitoring is a costly operation, and the way traffic engineering is proposed to be performed in today's circumstances is by random checking, as opposed to controlled operations.

> Traffic is monitored at random connections to reflect the various traffic conditions of hours of the day, days of the week and months of the year. Weekly reports include worst case scenarios and normal (actual) conditions to show the service provider how its' data flows traverse the network.

Many new companies provide access to GUIs on the web to allow the user(s) to follow the monitoring at all time by graphs etc. These graphs should be a part of periodic status reports.

> Review meetings will be held monthly (if necessary) for the first six months after the first signed service contract or at any party's request. After the first half year, meeting will be held at least twice a year or on any party's request [18].

<u>Service specific conditions</u>: Traffic conditions including QoS and security may be determined from monitoring and measuring of the network's capacity and function. Measuring should be done at times when the traffic exceeds normal, to show the worst case scenario, as to show which resources are available and when resources may have to be shared among other users. The terms may be agreed upon as follows:

> The network operator provides an xx Mbps full-duplex (two-way communication) connection between any users of the service offered by the service provider.

Allowing signaling messages flowing out of band between user and service provider, while the data concerning the conference flows directly between the parties involved, the network operator needs not know which signaling protocol is being used. The service provider must in this case have enough technical knowledge to be able to "translate" signaling messages and communicate them to the network operator if necessary, or devices handling this may be purchased by a sub-provider.

> Guaranteed throughput for the connection between any user and their conference partners provide at least a bandwidth of xx Mbps.

This can somehow be applied by using either constraint based routing or (Generalized-) MPLS. Constraint based routing, described in chapter 6, allows the sender to set up specific routes (through explicit nodes) over which the packet streams should flow. These routes/paths may be predefined and marked by a label, which is what is being used in MPLS (see chapter 6.3). The routes may be chosen by various constraints such as available bandwidth, priority setting of packets, directives made by SLAs etc to ensure quick and precise delivery from A to B. Making use of MPLS allows routers to make fewer and/or quicker routing decisions, decreasing the transmission time.

> The service provider has an available bandwidth between xx and yy Gbps, within which limits it may distribute services as wished to any number of customers.

In this point the agreement should make clear how this should be put to life in practice. The service provider should not be allowed to provide one customer with the entire maximum bandwidth agreed with the network operator, but must, as long as it is possible, share its resources with other traffic in the network. Other traffic should also be allowed to use the service provider's agreed bandwidth as long as it is available; not being used by the service provider. How this is regulated should be described in detail and added as an attachment.

For the service provider to provide the user with the expected real-time transmission it is important for the service provider to agree with the network operator about several network traffic performance conditions. As mentioned above, a minimum accepted bandwidth is one of the terms. Accepted jitter, packet loss and maximum delay are others. These conditions can be formulated as proposed below:

> Low packet loss (maximum x packets pr second)
>
> Maximum jitter (x % per time frame e.g. seconds)
>
> High throughput (minimum required bandwidth/bit rate)
>
> Maximum Round-Trip delay (from 1st bit sent until last bit received in one packet) – x seconds.
>
> Availability – as opposed to service availability this point refers to the network's performance if failure, congestion etc. Are there backup routes for the traffic or similar plans for solving such problems?
>
> Accepted call-setup delay – this should be extracted from the network operator's statistics like unsuccessful call ratio measured by real traffic and generated test calls. The statistics should reflect time variations over the hours of a day, days of week and month of year. See [ETR 138] for more statistical details.

These terms can be fulfilled due to the network operator's use of RTP. The demands require the network operator to set up the RTP services according to the terms agreed upon. You can read more about RTP in chapter 5.

There are also issues about link/transmission security which should be mentioned:

> The link is secure, no danger of sniffing/eavesdropping or spoofing. If sniffing or more active threats should occur, the network operator is responsible for damage costs to the service provider (i.e. sniffing resulting in SLA-divergence between user and SP).

<u>Reactions and compensations:</u> Monitoring may lead to failure detection, and similar reaction pattern as mentioned in chapter 9.1.5 may be taken. For instance the time frame of repair should be agreed upon:

> If failures are detected, it is accepted xx working hours/days before the failure is repaired.

Notice that “possible repair time” [ETR 138] (from the network operator’s statistics) may diverge from what is agreed upon here, because service provider and user may have agreed upon quicker or later repairs for higher/lower maintenance fees (see table 5, chapter 8).

Other failures and reactions should be outlined in detail with proposed compensations and/or reactions, and added as an attachment to the agreement. In case of choosing money compensations it should be one that matters; not just “pocket money” to the involved part; e.g. 10.000 NOK may be a small fee to pay for a company which earns 100 million NOK a year. The compensations may also be turned both ways, meaning the service provider may have to compensate the network operator’s loss if it does not keep within the limits agreed upon.

Force majeure: The SLA should outline which failures it does not comprise; which failures the network operator is not responsible for (similar to those mentioned in chapter 9.1.5).

Link failures, congestion etc is not seen as force majeure, because the network operator should include back-up plans if network-failure (e.g. router/OXC on network down). The back-up plans and similar actions to prevent service degradation may be described in detail in attachments, but may also be left to the network operator to decide. It is merely important to the service provider that most of such failures will be handled, according to availability matters.

As keywords one can propose “pro-active actions” and “preventive measurements” to meet both the service provider and the network operator requirements. This means “detecting and correcting network faults before service degradation or prior to customer’s perception, thus anticipating customer’s dissatisfaction and possible financial penalties” [18].

Charging/billing: The NP and the SP may agree to pay a total prize for all the services provided, or a divided invoice according to traffic measurements, e.g. per customer etc. The latter demands closer monitoring and may be more costly than a set price for all.

☆

The network operator may also have some demands to which it wishes to make an agreement with the service provider if the service should be delivered as planned. (This may be left out if the service provider and the network operator is same entity.)

Conditions: If this part of agreement is left out, it may be possible for the service provider to use whatever resources available in the network. This may lead to total congestion to other services in the network, and may also lead to violation of the user-SP SLA because the network may not be able to provide the minimum limits set in the SLA. That is why it is important to the network operator to demand a maximum usage of network resources to avoid problems like described. Demands made by the network operator may be:

> Maximum bandwidth between the service provider and its’ customers is xx Mbps.

> Maximum bit rate per second (per customer) is xx Mbps; service provider can not use more than the bandwidth agreed upon so that the traffic in the network does not get congested.

These conditions make the SLA more flexible because the service provider needs to relate to a set of intervals like; min x Mbps, max y Mbps, rather than one specific data rate. However, this may also be some kind of gamble between NO parameters and SP parameter, in terms of how much they are willing to rely on each other. See chapter 8.2.3.

9.2: High-speed data communication (ATM)

9.2.1: The service

In this chapter the service being considered is some kind of Video on Demand/Pay per View service. VoD/PpV may have to be delivered in real-time, or very close to real-time, as the user may ask to see his favorite TV-show within a specific time frame. The user may for example ask his/her service provider to view a football-match on a TV-channel he normally does not pay to have. The service provider should then be able to transmit the match to the customer through the network, and make him pay for the time this transmission occurred (or necessarily a “one-time” price for that specific screening).

9.2.2: Constraints

To be able to offer such a service, the service provider must have an agreement with a network operator which is willing to transmit the screenings being ordered, supporting high-speed data transmission. At the user’s end it is required that he/she is connected to some kind of data communication network so that he/she can communicate in high speed with the service provider.

In this scenario we consider the user to have an ADSL connection to the network operator with a possible maximum capacity of (theoretically) 8 Mbps downlink. See chapter 4.5. The required CPE at the user’s site is a DSL modem connected to a computer, TV or similar to access the network and software to handle the high speed connection.

The DSL-service itself (access to the network) is not considered in the SLA below, because it is the condition on which the service is being delivered by, and the focus is set on the user-SP SLA. If the user does not have a high speed connection link to a network, delivering VoD/PpV is not possible, and the service can not be considered. It is assumed that the VoD/PpV service provider does not deliver the DSL access, and therefore that the user and the DSL-access provider has got an own SLA with own demands and conditions (see examples in chapter 9.3).

9.2.3: Transmission between user and service provider

Which kind of transmission medium used within the network operator, must also be considered. (Asynchronous) Digital Subscriber Line is a service which provides the user with broadband access to the network he/she is connected to. The DSL-modem at the user’s end normally connects to a PSTN line card at the local exchange office’s (LEX) end, which among other things performs Forward Error Control, protocol processing, multiplexing and mapping. It is normally placed a Digital Subscriber Line Access Multiplier (DSLAM) within the LEX which enables several DSL lines to interconnect to reach the high-speed needed for the switching through the internet's backbone.

The data stream then passes through an ATM switch before it reaches contact with the ingress router of the network. The router handles the same things as the LEX line card, especially analyzing the header of the data packets/cells being sent, to achieve the information needed to make further transfer possible.

DSL in general is normally sent over ATM switching networks using SDH[35] (see chapter 4.4), and will be the technology used in consideration to this case. Asynchronous Transfer Mode is a "streamlined protocol with minimal error and flow control capabilities" [Stallings 2000, p.349], thus reducing the need for overhead when processing ATM cells and enabling higher data rates. See chapter 4 for ATM and SDH details.

9.2.4: Scenario details

The figure below shows the SLA-scenario considered in the example. The SLAs between the user and SP1, and the SLA between SP1 and the network operator, will be considered and possible parameters outlined. The user is considered to be a private person.

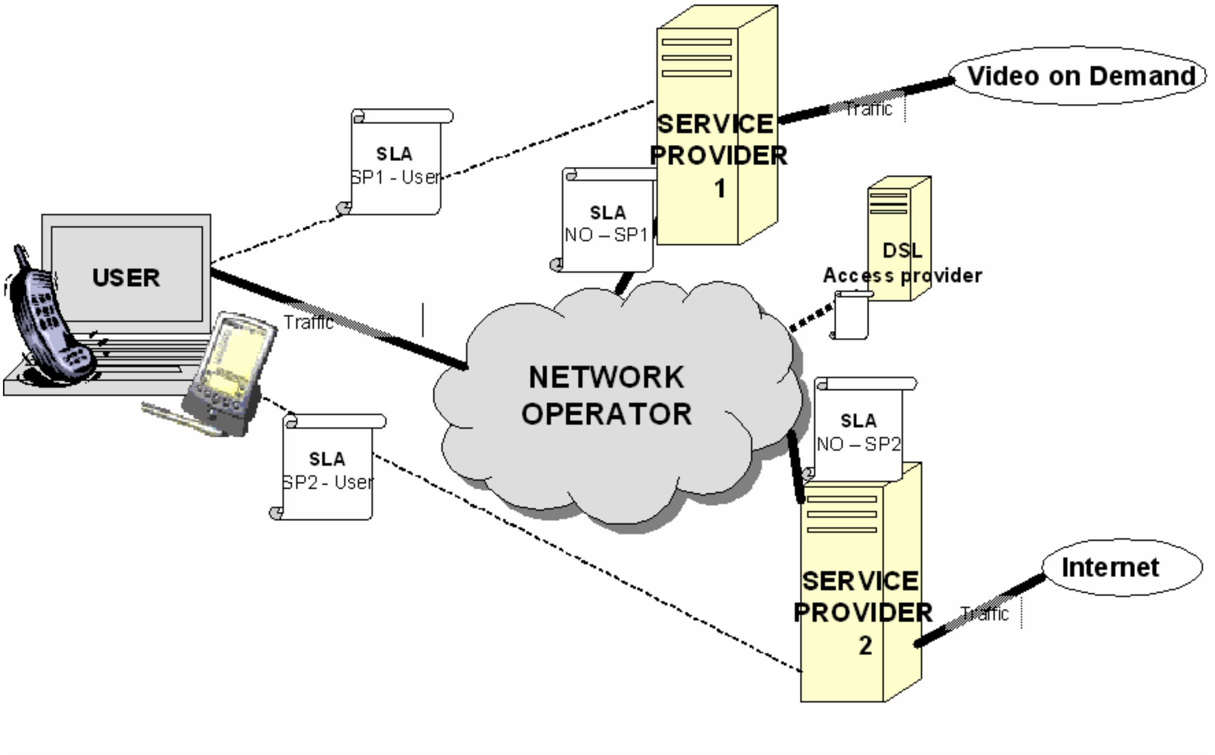


Figure 30: Video on Demand/Pay per View scenario (general)

[35] …although, in the future, other technology may be used.

Studying the scenario described above more carefully, we can sketch the scenario in more detail. Note that the same case may be possibly solved in several different ways, and that this shows the choices made in this example.

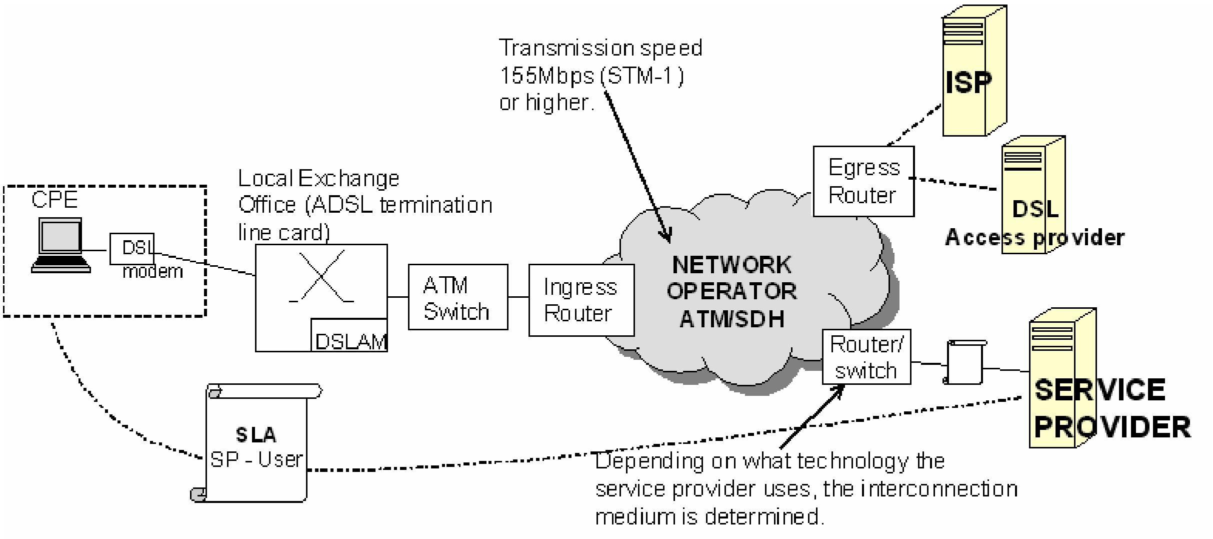


Figure 31: Video on Demand/Pay per View scenario (detailed)

An SLA is needed at literally every interconnection point in this scenario to make sure the agreements made between the actors are fulfilled. We assume that the network operator in this scenario controls the LEX, ATM switch and the routers accessing its' network, and that CPE is the user's responsibility. Note that the network access in this scenario is delivered by a DSL access provider, which is sketched in the figure, and some CPE, e.g. the DSL modem, may be the access provider's responsibility.

The interfaces considered below are the SLAs between the service provider and the user, and between the service provider and the network operator. Similar SLAs may (necessarily) be present between the ISP and the network operator and between the user and the DSL access provider as well, but are not the focus in this section. Seeing each actor's point of view a list of conditions wanted in the SLA are outlined below.

9.2.5: SLA between User and SP

The requests made by the user may not be very technical, but nevertheless important to consider for the service provider to present, to ensure the customers satisfaction. This section outlines a few possible parameters and explains the main points which should be included in an SLA.

Service type: Pay per View/ Video on Demand

Service description: Video on Demand is a service delivered through a high-speed data communication network which allows the user to order screenings on demand from the service provider, meaning he/she may be able to see the ordered screening on his/her TV/PC, or similar equipment, at the time and in the quality agreed upon.

Some of the points (confidentiality, initial terms, and time, measurement and review) considered in this SLA are quite similar to those mentioned in chapter 9.1.5, and will not be repeated. Note that since this customer is a private person, while the customer in chapter 9.1.5 was a small business partner, the review process etc will not necessarily be as detailed in this relationship. Private persons often get less guarantees and "worse" service, because they do not normally add to the company's major income the way a business partner does. For instance monitoring all connections to private users may be a waste of resources, because it is expensive to monitor many links all at once, and loosing a private customer due to service dissatisfaction often means less than loosing a big business customer.

Service specific conditions: may differ somewhat from those mentioned in chapter 9.1.5 since the example below concerns the conditions made for the specified service; VoD/PpV, although some parameters may be repeated to see them in context.

> A screening can be ordered at any time.
>
> A screening can be delivered at any time, x minutes/hours after ordering.
>
> The screening ordered will be delivered all at once, possibly in real-time if asked for. Meaning you can for instance see a football match when it is taking place without any delay.

These conditions give the need for implementing maximum delay-tolerance, and that the required bandwidth to deliver the service is available through the network at any time. Acceptable "down-time" should also be agreed upon. Service Availability is measured as a percentage of time during which the service is operational. [E.860]

> Service availability should be x % at any time of the day, week, year.
>
> The screening has DVD-quality, continuous streaming with no flickering in picture, no sound disturbances.
>
> The screening will be transferred without interruption guaranteed by x%. *(This refers to service reliability.)*

The technical terms needed to fulfill this request is to set maximum jitter-tolerance, packet-/cell-loss rate and provide some kind of means to make sure the fragmented screening packets/cells arrive at the correct time frame (use of protocols such as RTP, RSVP, H.323, SIP etc may provide some kind of control – see chapters 5 and 6).

> The streaming from service provider to user is secure, no chance for sniffing or hacking.
>
> The streaming data is "clean", - no chance of virus, worm, Trojan horse etc being transmitted.

Security can be ensured by using encryption, authentication and similar (e.g. IPSec – chapter 6.4). The details should be outlined in attachments.

The parts including reaction and compensation and force majeure are similar to those mentioned in chapter 9.1.5.

Charging/billing: In this case the prices can be set as:

> The price per VoD is dependent of the movie's popularity, its release date etc, and is specified in the currently available pricelist.
>
> Price on PpV depends on time length and content of screening (e.g. Super bowl or low-division season-game may be prized differently), also according to the currently available pricelist.

It is normal to VoD-services that the screening has a specific time-limit to which the user can hold the film. Because the user is not buying the film there must be this kind of time-limit, although the price/billing-conditions may say differently if you would like to keep a film for a longer period of time. This should also be specified in the SLA.

> A film transmitted to the customer is valid with a decryption certificate valid for 24 hrs. Pirate copy and further distribution of film is prohibited, and may be reported to the police.

9.2.6: SLA between SP and NO

To accommodate the conditions agreed upon in the SP-User SLA, the service provider also needs to have an SLA with its network operator so that it can guarantee that the conditions agreed with the user(s) are met. (This may be left out if the service provider and the network operator is the same entity.) The reader should notice the available “automatic” traffic policy functions[36] in ATM for guaranteed QoS within the network itself. See chapter 6.5 for further details.

Although the service being provided to the user is video on demand, the service required between the service provider and the network operator is based on transmission techniques, providing sufficient bandwidth to make it possible to deliver the service to the service provider’s customers.

The SLA will therefore include slightly different conditions and demands than the one described above (user’s perspective). Note that although the service provider has some technical knowledge, it might not know all the technical details within a network, not just because of the provider’s lack of competence, but because the network operator may choose to hold back such information. It is therefore important to also here describe (technical) issues in clear understandable language.

Service type: High-speed data communication (ATM)

Service description: High-speed data transmission may include both unicast and multicast. The network operator offers ATM transmission over SDH, which guarantees high speed switching of the data flows. Built-in functions in ATM enable the network operator to provide the wanted QoS-demands and monitoring.

The details should be outlined in attachments, describing how the network operator has configured its network to apply the QoS-requirements according to possibilities in ATM. Some of these may be read in chapter 6.

The parts considering confidentiality and initial terms are quite similar to those mentioned in chapter 9.1.5, and will not be repeated. Time, measurement and review processes are similar to those mentioned in chapter 9.1.6.

Service specific conditions: may differ somewhat from those mentioned in chapter 9.1.6 since the example below concerns the conditions made for the specified service; VoD/PpV. The terms may be agreed upon as follows:

> Guaranteed throughput for the connection between any user and the service provider should at least provide a bandwidth of xx Mbps, using either CBR or rt-VBR.

[36] Parameters which the traffic policy follows may be set explicitly by the network operator to meet the demands stated in the SLA.

Constant Bit-Rate (CBR) is normally used for the kind of service in this scenario; uncompressed audio and video information. Real-time Variable Bit-Rate (rt-VBR) may cause more bursty traffic because it transmits at a rate varying in time, but since real-time video compression often results in a sequence of image frames of varying sizes, requiring a uniform frame transmission rate, the data rate may vary [Stallings 2000, p.364-365]. Either is therefore sufficient depending on how the data is transmitted and which time-frame is considered. More details about ATM service categories can be read in chapter 4.3.

> The service provider has an available bandwidth between xx and yy Gbps, within which limits it may distribute services as wished to any number of customers.

In this point the agreement should make clear how this should be put to life in practice. The service provider should not be allowed to provide one customer with the entire maximum bandwidth agreed with the network operator, but must, as long as it is possible, share its resources with other traffic in the network. Other traffic should also be allowed to use the service provider's agreed bandwidth as long as it is available; not being used by the service provider. How this is regulated should be described in detail as attachment.

Other parameters and conditions are similar to those outlined in chapter 9.1.6. These are so similar because the service is basically the same, except the service in chapter 9.1.6 is IP-based, and this is ATM-based. The SLAs could be exactly the same, applying ITU's proposal of QoS agreements as attachments. The main SLA would therefore include general service description and terms, while the attachment would be specific to the service delivered.

9.3: Examples

These examples include chosen parameters, and may not necessarily be correct or applicable to similar services.

9.3.1: SLA user-SP; video conference

Telio[37] is a new company delivering pure IP-based, primary line telephony services to all Norwegian households that have, or that can have, a broadband connection. They deliver telephone services independent of what broadband connection/agreements the user has, ergo independent of who owns the infrastructure. All the customer needs is an adapter to attach to his/her broadband connection, provided by Telio, - and Telio's services are available to use.

In facts; what Telio delivers is telephony (but are working on other applications such as video telephony), but allow internet surfing on the broadband channel at the same time as using the telephone. We can therefore consider the Telio connection to be able to support video conferencing (video and voice), which is the example concentrated on in this section.

Service Level Agreement

This is a contract between Telio, hereby referred to as **provider**, and customer xx, hereby referred to as **user**.

Service type: video conference hereby referred to as the **service**.

[37] See www.telio.no

Service description: Video conferencing allows the user to perform communication between parties at different location based on speech and video transmission.

User's requirements: The user has to have an established broadband service. The provider will offer an adapter and a switch or router to which the user can connect its broadband modem and/or other customer premises equipment (CPE). All CPE is the user's responsibility as long as it is in her/his property.

The user has to choose the ~~Mini/Medium~~/Maxi subscription type to be able to use the service without lack of quality.

Provider's requirements: The provider has to fulfill all the requirements according to this agreement at all time. The provider has to provide the user with a switch/router and an adapter to connect to the user's broadband connection. The equipment is considered a loan and is the user's full responsibility as long as the agreement is valid.

Definitions: A broadband connection is a connection to the internet with the minimum speed of 1536 kbps.

A video conference is a two-way connection between two or more actors, exchanging voice and video information in real-time.

Timing: The service is operational from point of delivery.

Delivery will be done within 7 working days after accepted order.

Customer service opening hours are weekdays 07.00-21.00 and Saturdays 09.00-16.00. Sundays are closed.

Phone no: 76543, 0.32 NOK per minute

Mail: telio_cs@telio.com

24 hour service: 829 76543, 13.49 NOK per minute

The agreement may be terminated on either party's notice. The service will come to an end 2 months after first notice, and the router/switch and the adapter must be returned to the provider. If the equipment is not returned, the user is made responsible for the devices' value, and will have to compensate them to the provider.

The connection between user and provider will be monitored at a random basis to reflect the various traffic conditions of hours of the day, days of the week and months of the year. These statistics will be made available to the user in reports sent every third month.

Maintenance and upgrades are typically performed Wednesdays 01.00-04.00. Unordinary maintenance work will be notified to the user if it has duration for more than 4 hours within working hours (07.00-18.00 weekdays).

Confidentiality: The information exchanged between the user and the provider which may be considered commercially sensitive is confidential, and exchanged with good faith. This should under no circumstances be made public or by other means made available to any other party, without written consent from the other parties involved.

Pricing and billing: The user will receive invoices each month/~~quarter/year~~.

Details in the invoice may be specified on user's initiative. Contact customer service.

Subscription types:

Mini (1 telephone line): 159 NOK per month - works as a regular analogous telephone service; one telephone line and one telephone.

Medium (2 telephone lines): 239 NOK per month – similar to an ISDN-subscription; two lines with two numbers.

Maxi (3 telephone lines) : 289 NOK per month – ideal for a SOHO with two separate telephone lines, and a third line and number for e.g. a telefax.

Price per session: If more than two parties are involved in a session, an expense of 0.50 NOK per minute per extra participant accrues. *(Because of the use of extra network resources.)*

Force majeure: Situations such as service failure due to weather conditions, power failure, and other situations which are out of the provider's control, are referred to as force majeure and are not the provider's responsibility.

Appendix A: QoS terms

Service specific terms apply in addition to the ones mentioned. If they disagree of any kind, it is the terms made in this attachment which applies first.

A conference can be arranged at any time, without previously notification to provider.

Setup time < 5 sec

The number of cameras/camera angles transmitting the conference in either direction may be at a maximum 3.

The conference will be sent and received in real-time as a continuous stream with no flickering in the picture and no sound disturbances.

Service availability should be at least 92.5 % per month, measured at any time of the day, week or year.

The video conference streaming from service provider to user is secure, no chance for sniffing, spoofing or hacking.

The streaming data is "clean", - no chance of virus, worm, Trojan horse etc being transmitted.

The user must have a broadband connection which has a minimum downlink capacity at 1536 kbps to allow the provider to deliver a "perfect" service.

The maximum of 2 simultaneously held video conference may be obtained without risking lack of quality.

Appendix B: Reactions and compensations

If the QoS terms or other requirements/demands are not met as proposed in this agreement, the parties have the rights to claim some kind of reaction and/or compensation.

Service failure will be detected within 1 hour, and the parties involved will be notified within 1 day of detection.

Planned downtime resulting in service unavailability is not affected by the following rules. Note that the time used for planned maintenance etc is not considered when calculating the service availability (SA).

Table 6: Reaction and compensation (ex. video conference)

Description	Reaction/Compensation	Worst case
Payments are not received regularly	After 1 month: reminder is sent After 2 months: the matter is left to a dept-collecting agency	The relationship is terminated
The video conference is terminated during a session	Refund of 0.50 NOK	If this happens often (on regular basis), it is most likely a network fault, and it will be repaired within 4 hours of detection. If not, the user has the same rights as for service unavailability (see below).
Unable to establish conference < 12 hours	The service will be fully repaired. – No compensation	
Unable to establish conference 12-24 hours	The service will be fully repaired within 4 hours	If not; the provider should compensate the user with 2% of the monthly fee.
Unable to establish conference >24 hours	The service will be fully repaired within 4 hours	If not; the provider should compensate the user with 5% of the monthly fee per extra working day service not available, from day 2.
Unable to establish conference > one week	The service will be fully repaired within the next working day	If not; the provider should compensate the user with 20% of the monthly fee per extra working day, from day 8.

The user holds more than 2 simultaneously conferences, thus requiring more network resources.	1. time: warning is sent 2. time: warning is sent, and user suspended from service usage for 1 month 3. time: relationship terminated	If severe damage to network utilization, user may be made responsible for money losses.

If the conditions above are not met by any means, it is considered a contract breach, and the user has the rights to terminate the relationship to the provider.

9.3.2: SLA user-access provider; DSL connection

Since Telio sets a broadband connection as a condition to be able to realize its services, the user's agreement with the access provider will be considered in this section. In Telio's webpage they recommend the user to use NextGenTel's broadband DSL service.

NextGenTel[38] delivers ADSL, SHDSL and VDSL among other things. Since Telio requires a minimum bandwidth of 1.5 Mbps downlink, we choose ADSL, since it is the cheapest alternative.

Service Level Agreement

This is a contract between NextGenTel, hereby referred to as **provider**, and customer xx, hereby referred to as **user**.

Service type: ADSL broadband internet access, hereby referred to as the **service**.

Service description: Asynchronous Digital Subscriber Line is a transmission system which enables the user to achieve internet surfing at high speeds. ADSL makes use of the user's existing copper pair, and maximizes the capacity available in separate uplink and downlink connections.

User's requirements: The user has to have an existing copper pair on the site where he/she wishes to make use of the service. The provider will offer a DSL modem which can be attached to other customer premises equipment (CPE). The modem is the user's responsibility as long as it is in her/his property.

Provider's requirements: The provider has to fulfill all the requirements according to this agreement at all time. The provider has to provide the user with a DSL modem to connect to the network, and may provide installation help for certain fees. The modem is considered a loan and is the user's full responsibility as long as the agreement is valid.

Timing: The service is operational from point of delivery.

Delivery will be done within 2-5 weeks after accepted order.

Customer service opening hours are weekdays 08.00-20.00 and closed on week-ends and may be reached at phone no 55555 or e-mail: nextgentel-cs@ng.com.

[38] See www.nextgentel.no

The agreement may be terminated on either party's notice. The service will come to an end 1 month after first notice, and the borrowed modem must be returned to the provider. If the equipment is not returned, the user is made responsible for the device's value, and will have to compensate them to the provider.

Regular maintenance is performed the last Tuesday every month from 00.00 to 05.00. Irregular maintenance, upgrades etc will be notified by announcement on provider's web site.

Confidentiality: The information exchanged between the user and the provider which may be considered commercially sensitive is confidential, and exchanged with good faith. This should under no circumstances be made public or by other means made available to any other party, without written consent from the other parties involved.

Pricing and billing: The user will receive invoices each month.

Detailed invoice information may be specified on user's initiative. Contact customer service.

Establishing fees of 1098 NOK will be added to all orders. This is a one-time payment fee.

If required, the provider offers help with installation etc at the user's request for the price of 1350 NOK. Contact customer service to order.

Subscription types:

Bravo (1100/384)	498 NOK
Charlie (1600/512)	598 NOK
Delta (2200/640)	698 NOK
Echo (8032/864)	1198 NOK

(The user has to choose the Charlie subscription or higher to be able to use the Telio video conference service)

The pricing is per month, and the quantum of which the service is being used is not considered.

Legal statement: The user may by no means distribute the service further outside the user's premises without the provider's permission.

The user may by no means collect information about other users or the network by monitoring any process.

The user may not use any kind of spider virus, packet-sniffer, Trojan horse routing etc that is designed to provide means of unauthorized access to the network, or which may damage the networks utilization.

The service may not be used to modify, delete or damage any information contained on the computers of any users connected to the network.

The provider reserves the rights to modify pricing and other features related to this agreement, and add additional features or functions to the subscription, only when prior notice is made available on provider's web site or otherwise made available to the user.

Force majeure: Situations such as service failure due to weather conditions, power failure, and other situations which are out of the provider's control, are referred to as force majeure and are not the provider's responsibility. Subscription termination due to force majeure is not an issue for compensation.

Attachment 1: Quality agreement

Service specific terms apply in addition to the ones mentioned. If they disagree of any kind, it is the terms made in this attachment which applies first.

The internet will be available at all times at the speed agreed upon (SA=99%).

The internet connection is fast and continuous, meaning internet browsing should be seamless, either the pages are heavy to upload (e.g. lots of pictures etc) or not.

The browser's start page should be loaded and all elements present within 8 seconds after opening the browser.

The e-mail server's availability is 99%.

Contact with e-mail server is established within 10 seconds from any location (also remote log-on).

The user can access its subscriber details and e-mail everywhere (remote log-on).

The user may not connect more than 3 communicating devices to the modem.

Attachment 2: Reaction patterns

If the QoS terms or other requirements/demands are not met as proposed in this agreement, the parties have the rights to claim some kind of reaction and/or compensation.

All failure should be notified the customer service. Failures notified by mail will be answered within 2 hours during opening hours, or else the following workday.

The user should follow regular "netiquette" when using the service (see legal statement and point 8 & 9 below).

Table 7: Reaction patterns (ex. DSL-connection)

	Description	Reaction/Compensation	Worst case
1	Payments are not received regularly	After 1 month: first application, reminders fee accrues After 2 months: second application, incl. notification that matter is left to dept-collecting agency. Service suspension until payment received.	The relationship is terminated
2	SA < 99%	Up to 50% off of the monthly payment during present period (see details point 3+)	
3	The service is randomly available, meaning sessions are being terminated	The service will be fully repaired within 4 hours, no compensation.	If not, the rules in 4-7 apply regarding compensation.
4	The service is not available < 12 hours	The service will be fully repaired within short time. – No compensation	
5	The service is not available 12-24 hours	The service will be fully repaired within 4 hours	If not; the provider should compensate the user with 2% of the monthly fee.
6	The service is not available >24 hours	The service will be fully repaired within 4 hours	If not; the provider should compensate the user with 5% of the monthly fee per extra working day service not available, from day 2.
7	The service is not available > one week	The service will be fully repaired within the next working day	If not; the provider should compensate the user with 20% of the monthly fee per extra working day, from day 8.

8	The user tries to utilize the network's resources outside his/her limits (e.g. increased bandwidth)	1. time: warning is given 2. time: user is suspended from using the service for 1 month 3. time: relationship is terminated	If severe damage to network utilization or other actors, user may be made responsible for the provider's money losses. The matter may be notified to the police.
9	The user makes use of the service for illegal (by law) purposes, such as hacking, virus distribution etc	1. time: warning is given 2. time: user is suspended from using the service for 1 month 3. time: relationship is terminated	If severe damage to network utilization or other actors, user may be made responsible for the provider's money losses. The matter may be notified to the police.
10	The points 2-9 above happen more than 2 times in a row (at a more or less regular basis)	Compensation for up to a full month's payment. Cases are handled individually on event.	

If the conditions above are not met by any means, it is considered a contract breach, and the user has the rights to terminate the relationship to the provider.

9.3.3: SLA access provider-NO: data communication

In reality NextGenTel owns the infrastructure as well as access to the broadband service, and are able to control the aspects of network operation them selves. Here we assume the network is controlled by Telenor to be able to illustrate some of the issues described and discussed in the previous chapters. In the example it is assumed that NextGenTel has maximum 1000 users, which the terms in the agreement are negotiated by.

Service Level Agreement

This is a contract between Telenor, hereby referred to as **operator**, and NextGenTel, hereby referred to as **customer**.

Service type: high-speed data communication, hereby referred to as the **service**.

Service description: The service gives the customer the ability to transmit high amounts of data using the operator's infrastructure. The transmission may be either unicast or multicast, and real-time transmission is supported to some extent, see appendix A.

Timing: First contact by phone or mail with request for service will be answered within 2 hours (or the following working day if request is made off regular hours).

The service is operational from point of delivery and contract signed.

Delivery is made after initial agreement, no more 2-3 weeks after accepted order.

The agreement may be terminated on either party's notice. The term of notice is equally 3 months after first notification.

Regular maintenance periods are 01.00-04.00 (service window) all days. If the maintenance results in necessary service outage for more than 10 minutes, the customer will be notified. This time does not affect service (un-)available ratios or otherwise performed statistics.

Unplanned maintenance, upgrades etc can not be done without customer's consent.

Contact information: The user's contact person during the regular business hours is Mr. Operator who may be reach at phone no + 47 999 99 999 and mail john.operator@telenor.com.

Outside regular hours (18.00-06.00+weekends and holidays) customer service at 12345 (0.64 NOK per minute) can be reached for irregular business matters.

Price: The service has a set price, and the customer will be billed for 4 million NOK per year. Extra maintenance fees may accumulate if necessary.

Establish fee: 100.000 NOK

The establishing fee includes installation and a "first aid"-course; *'how to regulate your traffic resources and achieve good results using traffic engineering'*. All other support is billed by the hour: 400 NOK per hour, - provided by a consulting engineer from the provider's company. The customer's contact person; Mr. Operator, can be of some assistance regarding simple matters.

Confidentiality: The information exchanged between the user and the provider is considered commercially sensitive, confidential, and is exchanged with good faith. This should under no circumstances be made public or by other means made available to any other party, without written consent from the other parties involved.

Force majeure: Situations such as service failure due to weather conditions, power failure, and other situations which are out of the provider's control, are referred to as force majeure and are not the provider's responsibility.

Appendix A: QoS

Service specific conditions are as follows:

The provider offers a full-duplex (two-way) connection with the total available bandwidth of 4 Gbps which it may distribute to its users, and 2 Gbps for overhead, maintenance and monitoring causes. The customer may distribute access to the service to its customers/users limited by a mean bandwidth of 4 Mbps per user.

The provider holds no responsibility for any third party which may be involved.

The provider claims the right to utilize the customer's bandwidth if the traffic is lower than expected.

The provider will offer support to the customer regarding traffic engineering and resource utilization, e.g. to help them guarantee low delay and packet loss for data delivery between two or more users.

Appendix B: Monitoring

The provider claims the right to monitor the customer's relations and traffic utilization.

Traffic will be monitored and measured at all interfaces between customer and provider, ensuring the SLA terms are held. From this the provider extracts and calculates statistics which will be made available to the customer in monthly reports. On www.telenor.com/monitor the customer may logon and find more resent monitoring reports, including:

- Response time and packet loss for the provider's backbone.
- Bytes out/in in the network
- Percentage of total available capacity being used

presented in a graphically web-interface. The web-graphs are updated frequently, and the customer may choose over which period it wants to see the statistics; day, week, month or year.

The monitoring of third parties' connections will be done at random used for control of customer's access distribution.

The provider guarantees that monitoring will not interfere with traffic regulation or by any means lead to service degradation.

Appendix C: Compensations and reaction

If the SLA terms above are not fulfilled or diverge of any kind, reactions should be taken and (monetary) compensations in either party's favor may be taken.

Table 8: Compensation and reactions (data communication)

	Description	Reaction, regularly	Reaction, worst case	Comments
1	Delivery not made on agreed time	1/3 establishing fee	Order cancelled	If delivery overdue 8-13 days: 2/3 of establishing fee Overdue more than 14 days: free establishment
2	Customer uses more bandwidth than agreed.	Customer receives warning and fine: 100000 NOK	Relationship terminated	Fines may accumulate due to provider's loss/damage and/or if terms are broken repeatedly.

3	Provider gives access to less resources (bandwidth) than SLA terms	Provider must pay 100000 NOK to customer for lost income.	Relationship terminated	Since customer provides access to its resources, it may have SLAs with its users which in terms may require compensation if SLA terms to them are violated. The compensation may therefore be (re-)negotiated regarding of the customer's user's SLA terms.
4	Service unavailable (SUA*)) < 8 hours	No compensation		*) Service unavailability is calculated per calendar month
5	8 < SUA < 16 hours	60000 NOK (15‰ of annual payment)		
6	16 < SUA < 24 hours	132000 NOK (33‰ of annual payment)		
7	SUA > 24 hours	200000 NOK (50‰ of annual payment)	Relationship terminated	Compensation may accumulate if service unavailability continues for several days. Must be handled case individually.

Appendix D: Document history

All happenings in relation to this agreement will be listed chronologically below including re-negotiation phases etc.

9.3.4: Comments

These three examples (9.3.1-9.3.3) are agreements which relate to each other:

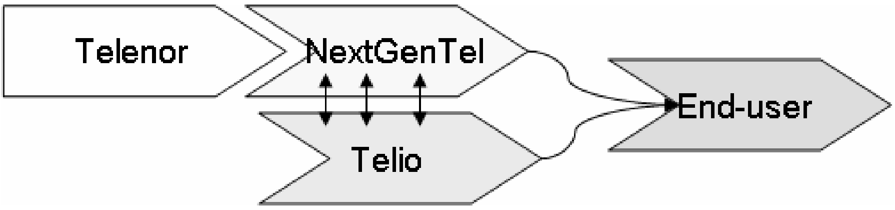


Figure 32: Chain of providers

One of the main differences between the agreements is the so-called service window; the time frame for when regular maintenance and upgrades will be executed. In the Telio-to-end-user relationship the service window is set to every Wednesday at 01.00-04.00. In the NextGenTel-to-end-user relationship it is Tuesdays once a month between 00.00 and 05.00. The service window between Telenor and NextGenTel is 01.00-04.00 every day, but should not cause service outage without explicit notification.

The service window is regularly set to periods when the service traffic (usage) is expected to be lower than usual. It is for instance not expected that video conferences will be held in the middle of the night (referring to the Telio service window), and if so is necessary, the user should schedule them so that it does not interfere with the service window Wednesday nights.

In the service window periods the user may experience lower quality than usual and perhaps total service outage. Users of for example a broadband internet service as ADSL provided by NextGenTel can thus not expect the service to function correctly (as stated in the SLA) every last Thursday of the month.

The service window for the agreement between Telenor and NextGenTel is wider than the other service periods due to the extent of the relationship. Greater costs are at stake if something should happen to be wrong in the network delivering this kind of service or on the connections between the actors involved. It is therefore necessary to arrange maintenance and network checks more often to take preventive actions against possible threats like link failure and service outage.

It may seem as though NextGenTel has the “worst” service window towards their DSL-customers; - with maintenance only once a month. This period is chosen because we assume that NextGenTel rely on their agreement with Telenor and their maintenance efforts to regulate the service. This maintenance should be sufficient to regulate NextGenTel’s services, which NextGenTel distribute further to their customers.

Another difference worth commenting is the service availabilities. Because Telio and NextGenTel have specified their SA at respectively 92.5% and 99%, it may seem as though Telio’s video conference service is less stable than NextGenTel’s ADSL service. However, since the SA-values state the minimum limit to which the service availability should hold, measured over a period (normally per month), it might as well be higher, thus better. These values are chosen to illustrate the issues of safety margins discussed in chapter 8.2.3. Because Telio’s service relies on other providers’ broadband services, Telio does not wish to take any chances as to how well those services work, and applies a safety margin of 6.5% for the service availability.

10 Today's situation – how does it work

This chapter deals with which methods are being used today to ensure good quality and relationships between users of telecommunication and data communication services and their providers, as a partial alternative to conditions considered in an SLA. See chapters 7-9.

In Norway several laws include issues about sales of goods and services; - the users' privileges if the services or goods are not fulfilled as expected. In addition all, or at least most, providers have made their own agreements concerning consumers, outlining their responsibilities according to the service being provided and their responsibilities in relation to the customers. In contrast; the user being a big business customer (e.g. Norsk Hydro), it may in its case dictate the service provider to supply its' service needs, meaning the customer sets the terms to the agreement.

10.1: Relevant laws

"Lov om kjøp 01.01.1989, last edited 01.07.2002" [20] (The sale of goods law) includes, among other things, issues which every salesperson has to relate to by Norwegian law:

- Secure delivery of goods/services between purchaser and salesman.
- Reference to responsible person/instance if the delivery is not performed at the right (agreed) time. If this is not fulfilled it is referred to as a contract breech.
- Reference to responsible person/instance if the goods or services are not delivered as predicted; with losses or damages, both previously and after the purchase, and relative compensations.
- Details about the good's type, quantity, quality and other characteristics should agree with purposes which similar goods are normally aimed for. When the goods do not fulfill these demands, or if the information about the goods is incorrect/withheld; - the goods are considered a deficient.
- Force majeure issues; - if the causes to damage or delay of a delivery or similar situations are unpredicted and not caused by any of the parties involved, they are excused from the responsibilities.
- How purchases can be changed, cancelled or compensated e.g. by price reduction or other (money-wise) compensations.
- How relationships to a third party e.g. provider should be handled.

The law also includes information about what is considered a contract breech and what should happen if it occurs according to any kinds of sales; - be it goods or services. The government has also extracted laws concerning sales specifically towards consumers [21] and "Lov om opplysningsplikt og angrerett m.v." [22] (law concerning rights to repent purchases) to handle issues similar to what is considered in an SLA.

As these laws apply to all purchases being made in Norway, the rights of the user should be quite clear. The problem is that many service providers and other salesmen try to "interpret" these laws in their favor, and write their own "laws" into agreements made between them and the customer. Especially private persons are often "tricked" by this, because they are less informed of what their rights really are. In circumstances by which this thesis has concentrated on, relying on the law is not sufficient to the customers, neither to the providers. Service specific details must be explained and agreed upon, and ensure that relationships between the actors in an agreement are well conducted.

"Ekomloven" (law for electronic communication) [32] states mainly which issues that a provider must relate to in order to deliver "good, reasonable and future-oriented electronic communication services" to all users, by "using a community's resources efficiently to arrange for competition and stimulation of business development and innovation". Some of these issues will be referred more carefully in chapter 11.

10.2: The "SLAs" towards actors in the private market

Studying some occasional "SLAs" according to telecom services being used by private consumers today, a few comments may be in place. All comprise similar issues, but some companies renounce more responsibility than others. Some of these agreements are outlined shortly below, before they are compared and debated up against each other.

Netcom is the second greatest company in Norway who provides mobile services towards consumers in the private and the enterprise market. The agreement they make with their users outline which responsibilities they have towards the user, but concentrates mainly of how the user should make use of the service. Netcom renounces a lot of responsibility by i.e. not willing to guarantee service availability;

> *"Netcom does not guarantee that the customers' service-usage will not experience interruption, nor that all sessions reach the destination or meet other transmission obstacles."*

They also renounce their responsibility for session disturbance e.g. causing breaks. Although they do compensate the consumers' direct losses due to service shortage, they may claim "no responsibility" if Netcom can claim the shortage is resulted by matters outside Netcom's jurisdiction, like network outage, lost profit due to changes in 3.party contracts etc. In all cases of disagreement, Netcom hand over the responsibility to the Norwegian Post and Telecommunication Authority.

Canal Digital is a company delivering digitalized TV-signals and internet over cable. Their "SLA" is more comprehensive on terms of which responsibilities the parties in the agreement have. They state the consumers' responsibilities in terms of CPE; the CPE must have a certain standard to make possible usage of the services. Extra CPE is supplied by the provider (e.g. modem or decoder), and is the consumer's responsibility as long as it is in his/her possession. But the provider will repair any damage, faults or shortage on the rented CPE which is out of the leaser's control. The consumer may demand compensation if the service is somehow interrupted, - if the shortage is within the provider's jurisdiction, but the up and download speed are not guaranteed to meet the exact speed stated in the agreement at all time.

Canal Digital claims that the user can not demand better sound and video quality than what can be expected due to the cable-networks' standard. They also state the rights to monitor the service to prevent and discover possible abuse. "The consumer obliges to follow regular "netiquette"." Abuse or missing payment leads to automatic ceasing access to the service.

☆

The agreements made by a service provider should be a supplement to the laws, ensuring that the users' rights are fulfilled in all matters regarding a service purchase and use. It should among other things comprise issues of how a consumer should relate to a fault or service shortage, and explain how the service should work correctly.

The agreements outlined above are examples of how the laws can be interpreted into the service provider's favor. They comprise most of the issues mentioned in the laws, but renounce obligations to their customers which are caused by faults etc that they claim are out of their hands. It is legal to do this because they mention the issues in an agreement which the consumer should read before he/she agrees to the terms and signs the deal. However, it gives the customers less rights in case a problem should occur.

The agreements do not give the customer a set of tools to secure his/her privileges, but rather claim which rights the user does not have; which responsibilities the provider has in case the user exploits the service's agreed characteristics, and margins of how high money-compensations the service provider may give in case the service shortages are their faults. These agreements can be interpreted in such ways that the user does not have any rights, but hold the full responsibility no matter what the fault should be and who is responsible for it. It seems hard for the user to benefit from the agreement at all because the provider renounces nearly all responsibility in one way or the other.

10.3: The "SLAs" towards actors in the enterprise market

The actors in the enterprise market were more reluctant to hand over their customer agreements, and a description as above can not be given. SLAs are held as business secrets which nobody is willing to share.

Generally spoken; the focus to service agreements and guaranteeing the customer's rights has become very "in" the last year or so. Many big telecommunication companies have begun to write SLA-agreements with their customers, and are requiring standards regarding what they should include and for standard values to the parameters proposed. One reason for setting standards is that it makes it easier for any customer to compare the services' advantages. Another advantage is to support the service providers with less technical knowledge, so that they can make competent SLAs as well as other companies with better competence.

Users in the enterprise market, e.g. the Norwegian National Bank, demand agreements to secure their rights, because their use of the services is critical to their business (they for example sell/buy NOK by e-mail). If the services lack of quality or what they receive differs from what they thought they bought from the provider, big (economical) damage can arise for the company (see chapter 2.5).

Similar inquiries resulted in the Norwegian Post and Telecommunication Authority's SLA proposal [12], using ISDN as an example to outline the parameters (see chapter 8). Workshops and meetings have been held to discuss this proposed standard, and some changes have been made to make it more general as an SLA template for similar services. But inquiries for similar proposals in relation to other services, e.g. data communication, IP-switching etc, were soon to come. This shows that the interest is big, and customer's rights may soon be better covered that they are today, especially for those in the enterprise market.

Statens Forvaltningstjeneste (SF- the office for public administration) has a different approach than other service providers because they provide a multiple of services to a small number of users (normally there are a few services offered to a high number of users). Even so, they realized, as other service providers, that a service agreement is the answer to many problems in the state departments, including solving misunderstandings and disagreements, establishing common understanding of the services content, and creating a common ground for communication; the actors need similar perception of what the issues to be discussed are [28]. SF had to use a lot of effort to get the state departments to understand the necessity of using service agreements, and the work of carrying out their thoughts were tough, but finally they had it forced through. Other companies, big or small, have similar processes going on, both within their companies and towards outside (potential) business customers.

10.4: Regulating service levels in today's market

As outlined above, one may conclude that there already are sufficient methods to guarantee QoS towards users of telecommunication services. There are laws which providers and the users have to relate to, and most providers have written their own service agreements including the legal issues concerned by the laws. However the interest of how to write an SLA and what such an agreement should include is increasing and could mean that the necessity of SLAs is present.

One can argue that free competition is sufficient to regulate the market. If a customer of a service is not satisfied, he/she can simply switch to a different provider with similar service characteristics and hope that they do a better job providing the quality you want. Since the number of providers in the multi-provider environment is increasing, it should not be a problem to find another actor providing similar services. A downside though, is that big money may be involved in changing providers, especially if the customer is in the enterprise market.

In Norway there is always a problem at New Years Eve, where nearly everybody tries to call or send SMS' to their friends to say "Happy New Year". Every year the mobile network breaks down; it is impossible to get through on the phone and the SMS' does not arrive until the morning after. The two big mobile network operators; Telenor and Netcom were interviewed in the newspaper in the beginning of January this year, being confronted with the network problems [31]. Neither of them claimed they had the responsibility for the break down, and claimed it was the other party's responsibility. Telenor blamed Netcom's infrastructure, and Netcom blamed Telenor's infrastructure. Where the problem lies was never established, and the same problems will most likely occur next New Years Eve as well as the next. In the operators' defense, mentioning the problem of dimensioning a network for peak balances is bad utilization of network resources, because it is so seldom the traffic reaches so high. But people mean that some preventive measures should be taken to handle such events. This example shows that switching from Telenor mobile to Netcom mobile, or any other mobile company, might not give you any better throughput for your SMS' on New Years Eve.

Another perspective can be seen from the providers' point of view. If they provide a service with bad quality and they have dissatisfied customers, they will lose customers and get a bad reputation (jungle telegraph), thus losing income. Is that a position they would like to take? In addition, competition may make the actors get involved in using "dirty tricks" to get more customers. They will exploit other providers' downsides and turn it to their own advantage.

Could a well written SLA correct these things? Would an SLA for example be able to regulate mobile traffic so that the SMS' arrived in real-time or close to real-time at New Years Eve as they are supposed to? Would a customer be willing to pay more to get his/her SMS priority over other SMS' in high traffic periods? Or are today's conditions sufficient to the customers?

In the enterprise market we have seen examples that quality of service may be more critical, e.g. in chapter 2.5. Are business customers more likely to be interested in well-defined SLA-"standards", or do most of them also rely on the law, free market competition etc which exist today?

11. Reflection and analyze

We have seen different ways to arrange for guaranteed quality of service (chapter 9), and the main focus has been set to whether today's situation's law and/or free market competition (chapter 10) is sufficient, or if the need for SLAs is present. The table 9 below sums up these issues, considering three main groups:

- Regulation in an SLA-environment
- Regulation using the laws
- Regulation (or no-regulation) with free marked competition

Figure 33 illustrates the area of consideration in this analysis. To the left you see strictly law-regulation, meaning every thing is controlled by a court or regulation office, all issues described in chapters 7-9 regarding service agreements are pre-defined, and nothing is left to the actors' own choice. You may see this as centralized regulation (in worst case; monopoly), where all disputes is solved by the court or a regulation office.

On the other side, to the right, you see totally free market competition. Here nothing is controlled; all actors may do as they "wish" (see discussion below). Will a no-regulation community result in chaos, or will the users protest, demanding some kind of control or regulation?

In an SLA-regime we may see the control as distributed regulation, because there may be different control elements in each SLA for each link in the chain of providers. A provider distributing services to a million users may for instance have a million different SLAs.

In this chapter the extreme extents are not considered. Instead the gray area illustrated below is the area being discussed. The free market relies somewhat on the laws, and the laws considered are defined in a way that allows actors to interpret them and define own versions of service agreements.

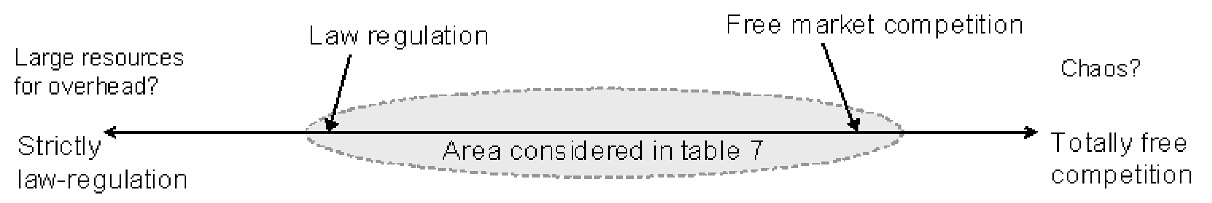


Figure 33: Area of consideration

11.1: Discussion

Some of the points discussed in previous chapters are considered in relation to each of the regulation-regimes mentioned above. These are described below, and summed up in table 9.

(1) **Competition**

Competition between actors in a multi-provider environment will always be present one way or the other. Service providers offering similar or the same services, may make their own versions of the service descriptions so that their services do not appear exactly the same to the customer and therefore appeal to different customers.

Note that if the strictly law regulation-regime applies; there will not be much competition for other than prices. The law regulation will have pre-defined all the qualities and characteristics of which a service should include, which the providers would have to fulfill, meaning similar services would be exactly the same, only diverging each other by price. Since this is only theoretical, and the analysis is made for a more realistic point of view, we say competition is present in all areas of regulation.

(2) **Legal (government) regulation**

Market regulation may be executed to make sure the market competition is maintained on a “healthy” basis. Actors in the same market/environment should have good relationships, and keep to the laws existing in that environment. The “sale of goods” law regulates this very well, and as mentioned in chapter 10; SLAs should add up to such a “perfect” scenario as well, especially if standard SLA templates as [12] (chapter 8) are acknowledged.

In the free market competition providers may, as previously mentioned; turn a blind eye to the laws and/or make their own modified agreement versions to accomplish their needs and wishes. In the worst case scenario, there is no regulation in a totally free market competition environment; no existing agreements between a user and its provider, thus leaving the users with no rights.

(3) **Guaranteed end-to-end QoS**

Service providers should be able to provide end-to-end QoS if they have defined well enough SLAs with their sub-providers and other involved actors. Using one-stop responsibility (see chapter 7.2) guaranteed end-to-end QoS should be possible, at least in theory, but it may be harder to realize in real-life.

However, law regulation or free market competition can not guarantee this; although the laws include somewhat information about that a service purchased should have similar conditions compared to similar services (see point 11).

In the extreme case though, strictly law regulation may support guarantied end-to-end QoS, if e.g. the laws define margins for the maximum range a provider may act, how much QoS-parameters may oscillate (see chapter 7.2 and 8.2.3) etc.

(4) **Change of provider**

If problems occur and the user of a service is not satisfied, the easiest way to handle it is to change service provider. (In a multi-provider environment we expect there are several options of choosing a provider offering similar services.) Price-regulation is also a reason why a user might want to change his/her provider.

Neither the law nor the free market competition prevents the user from changing provider, but a user bound to an SLA may have problems because the agreement may limit the user to hold on to the service provider for a certain amount of time. Especially in situations where the service in the SLA is formed to suit the user’s needs and demands, it may be hard for the user to get out of the agreement “light headed”.

Note that we only consider the users’ point of view. The service provider may have reasons why they do not want to keep a customer (e.g. bad behavior, distributing viruses etc – see chapter 9), and the user may be “forced” to change service provider.

(5) **Trouble shooting**

Trouble shooting can be performed in various ways and some of them are spoken of in this reflection analysis. Trouble shooting should be done either to prevent service degradation due to faults etc, or to define how actors can/may deal with faults, service degradation etc if they occur.

a. **Problem detection**

If there is a problem in delivering the service, it will be detected in either environment by the user and/or the provider. However, if the problem lies somewhere within a network in which the provider delivers its data communication service etc, the provider may have to apply some kind of monitoring equipment or similar to detect the problem(s).

An SLA should specify how problems/faults are detected, but may specify that the provider renounces this kind of responsibility. The laws specify how goods should be delivered, what the goods should consist of, and the actors’ rights if something diverges from service specification. Problems in free market competition may be detected by either part (provider or user), but does not necessarily provide any information or action.

b. **Fault handling**

As with problem detection, fault handling may be specified in the SLA, but the provider may renounce responsibility for various faults. The laws specify which rights the user has if something is wrong with the service, compared to similar services. Faults detected by providers and/or users in the free market may or may not be dealt with, depending on how good service the service provider offers.

(6) **Traffic monitoring**

The Norwegian “ekomloven” (Law for electronic-communication) [32] specifies that “the provider of a publicly available electronic communication service must measure and inform the regulating instance about the service quality offered to the end-user.” Traffic monitoring is therefore necessary.

Actors in the free market competition do not have to perform any traffic monitoring. However, service providers committing to SLAs should specify how the quality of service will be guaranteed by monitoring and measuring traffic, but may do it differently depending on the service content and the customer priority level. See chapters 8 and 9.

(7) **One-stop responsibility**

The one-stop responsibility (chapter 7.2) applies to the SLA-environment as we assume the service providers write SLAs with all immediate “neighboring” actors, meaning it is not responsible for any third party provider’s actions. The sale of goods law [20] specifies that "if a delay depends on a third party performing tasks given by the salesman, the salesman is free of responsibility…"

(8) **Stability/predictability**

How predictable or stable an actor-actor relationship may be depends on how well the agreement between them is written. This also applies to how predictable the service is; which service qualities and characteristics that can be expected at all time.

In the free market competition there may be no stability, no guarantees as to how the relationship or services work (see chapter 10.2 where the service providers renounce responsibility). An SLA though should state these points quite clear in terms of how a relationship is started, how it is terminated, and include service description (chapter 7-9). The law also includes regulations of terms of notice, what to expect from the service, and how matters can be solved if the service expectations are not fulfilled.

(9) **Mobility**

For some users mobility may be an issue of quality of service. To be able to access the service from different locations with the same resources available may be important. Some problems concerning mobility are mentioned in chapter 8.2.4.

a. **Location dependent service**

Whether the service is location dependent or not, depends on the service provider's relationship to the network operator(s) and the type of service. Some services may be easier to distribute across different networks than the one or those regularly being used, others harder. Services may be dependent of location or not, but should either way be mentioned in an SLA. Laws and free market competition does not apply to this.

b. **Remote access to service**

Some service providers offer remote access to parts of a service, e.g. remote logon to check e-mail, which is quite common. How, and if this is possible should be stated in an SLA, but may or may not be accessible in the law or free market competition environment.

(10) **Standard time-scheme**

It is normal to implement a time scheme when selling/buying a service or any kind of goods (see initial terms in chapter 9.1.5). The time scheme should include

- how and when the service can be purchased/ordered,
- how and when the service will be delivered,
- delivery time (time from order to delivery),
- when and in which order payments should be done (e.g. who is responsible for delivery costs) etc.

All of the above are covered in the sale of goods law which state how a salesman and the buyer should relate to a sale; what should be done prior to, during and after a sale. These issues should also be covered in an SLA, although it may be treated independently by different actors.

(11) **Standard service quality**

An aim is to be able to easily regulate market competition by implementing that similar services have similar quality and characteristics; making it easier to i.e. compare the services. The laws include this by stating that the purchased goods should "suit the purpose which similar goods are normally used for, have the characteristics shown by demonstration or similar, and correspond to those demands made for type, quanta, quality and other characteristics for the specific service" [20]. It can therefore be claimed that the law sets a standard to what a service should consist of according to similar services in the market, without specifying every type of service in detail.

The same may apply in an SLA-environment if the SLA-proposal made by the Norwegian Post and Telecommunication Authority [12] is approved as standard template, but it is not regulated in a free market competition.

(12) **Incentives**

(Economic) incentives may add up to the actors' urge not to break the agreements made. In case of the law; breaking the law may be punished by a fine to the government or other regulating instances. Ekomloven [32] for instance states that "the government can set an economical penalty/fine accumulating each day until the illegal activities have ceased…"

In the free market competition, bad services and/or customer-handling may lead to loosing customers. SLAs normally have parts where reaction and compensation patterns are outlined. See e.g. examples in chapter 9.3.

Incentives give the actors a motivation to stick to the bargain, rather than running their own game.

Table 9: Comparing different regulation-environments

	Conditions:		In an SLA-environment:	Using the law	Free market competition:
1	Competition		Y	Y	Y
2	Legal (government) regulation		To some extent	Y	N
3	Guaranteed end-to-end QoS		Perhaps	To some extent	N
4	Change of provider		Y/N	Y	Y
5a	Trouble shooting:	Problem detection	Y/N	Y	Perhaps by coincidence
5b		Fault handling	Y/N	Y	?
6	Traffic monitoring		Y/N	Y	N
7	One-stop-responsibility		Y	Y	N
8	Stability/predictability		Y	Y	N
9a	Mobility	Location dependent service	Y/N	?	?
9b		Remote access to service	Y/N	?	?
10	Standard time scheme (how to handle a purchase before, under and after – ref. initial terms in chapter 9.1.5)		To some extent	Y	N
11	Standard service definition		N	Y	N
12	Incentives		Y	Y	Y
	SUM: “Is quality of service guaranteed?”		**7.5**	**12.5**	**-2**

The sum is calculated by giving each Y 1 point, each N (-1) point and Y/N/perhaps-situations are given 0.5 points. E.g. SLA-environments have

Y: 4 x 1 point	= 4
+ N: 1 x (-1) point	= -1
+ maybes: 9 x 0.5 points	= 4.5
giving the total score of	= 7.5 points

11.1.1: Weighting the values

Summing up the table's results, the weight for each point has been set equally important. It may therefore seem as if the overall best solution is to enforce the law at every point, although, as described in chapter 10, the laws lack of details concerning each and every service.

Other ways to use the table may be to weigh the conditions differently in relation to different services and/or point of views. Considering the conditions above in relation to a service which has functioned perfectly for several years, e.g. a PSTN service, the most important point may for instance be competition in terms of pricing. Other conditions are less relevant since it has worked more or less "perfectly" "for ever". Mobility issues are for example not important for a fixed telephone service, and traffic monitoring and fault handling may not be so important while there are none or a very small number of problems detected.

On the other hand, considering a completely new service, e.g. VoIP, it may be important to monitor the traffic to detect and correct errors when it is tested in real life. The conditions about end-to-end quality of service and trouble shooting may therefore be the most important ones.

A point of view may be seeing the conditions from an end-users perspective, another from the provider's point of view. Some actors may use completely different values/priorities than other actors when weighing the conditions. It is therefore *hard to conclude which regulation-environment is the better one*. It depends on which type of service is being considered, how long it has been "on the market", and which actor's point of view is being taken. In general the best solution may seem to be to use a combination of SLAs and the law. Actors who model service level agreements should always take the law into consideration in every point so that these two elements do not diverse in any way. SLA negotiations will also try to make the best out of all viewpoints.

12. Conclusion

12.1: Concluding remarks

Because of high market competition between services in today's multi-provider environment, means to regulate quality of service is required to guarantee providers', users' and other actors' rights when purchasing or selling a service. There are already a multiple of existing services available to users, and with implementation of NGN or similar future network propositions, the number of services and different providers are expected to increase rapidly. The needs for regulation are therefore even more present in NGN.

Creating a model with effective traffic and network handling methods for packet based networks, using tools such as MPLS, DiffServ, routing mechanisms and SLAs is one of the future goals. This thesis has concentrated on methods of how such a model can be realized, using different tools.

One proposal of regulating quality of service (QoS) is to implement service level agreements (SLA) as standard, regular agreements being used whenever an exchange of service resources is made. An SLA should include service description, user's and provider's rights in terms of delivery, faults, service degradation, monitoring etc, pricing and other quality of service terms which are service specific. SLAs can also be used by customers to compare similar services, that way improving the competing environment.

Standardizing SLAs has been a main focus for many standardization organizations; how can they guarantee quality of service, as opposed to free market competition and legal regulation. A naïve perspective may be to say that the laws in Norway and free market competition are sufficient to regulate the multi-provider environment, but there are reasons why it is not so. Service Level Agreements are used to ensure that all relationships in an actor network environment, be it service providers, network operators, customers etc, operate in the "correct" way; ensuring quality of service, traffic engineering, and economic and legal issues. This is why this is one of the "hottest" topics among big actors in the telecommunication sector.

12.2: Check list

In the introduction a few questions or problems where outlined, which where proposed to be answered in this thesis.

- A main concern is that delivering services in a multi-provider environment is complex, and that the complexity grows with NGN implementation when the number of actors evolves. It has been stated that service provisioning needs regulation because users require good quality for their services. Competition between service providers in terms of pricing, quality and other service characteristics may not be the best way to ensure the users' needs.

- Actors in a multi-providing environment should find ways to define and operate mechanisms for QoS-management towards all other relevant actors. One way to do this is to include service level agreements between all relevant parties involved in delivering a service. Such agreements should include regulations in terms of quality, delivery time etc, and comprise service characteristics and priorities of every part, to make the actors involved aware of their responsibilities. Other mechanisms for QoS-management can be to rely on the legal system, or even to believe that service providers will see market competition as a competitive edge, and thus make an effort to satisfy their customers.

There are two main perspectives taken in this thesis, referring to end users; the situations seen from users in the private market's point of view and users in the enterprise market's point of view.

- In the private market one may conclude that the situation as it is today works at an acceptable level; it is quite easy for a user to buy his/her services from another provider if the current one is inadequate. On the other hand, it may be problematic and/or tiresome to never have a satisfactory service and having to switch service provider "all the time".
- In the enterprise market the situation is a bit differently. Big companies as the Norwegian National Bank (see chapters 9.1.1 and 10.3) etc depend their business on the services they purchase, meaning the services should function "perfectly" all the time. It may therefore be more critical for them to have specified SLAs to describe how the service will be experienced and how to regulate possible service degradation problems and/or divergences from the agreements.

A third perspective is the one seen from the providers and operators point of view, referring to the one-stop responsibility and their responsibilities if the end-users and/or sub-providers are dissatisfied. All of these actors see advantages of implementing standard methods for service regulation etc.

Chapter 11 sums up the different methods to regulate service provisioning in a multi-provider environment, and concludes that it is difficult to decide which kind of regulation is the better. However, it is made clear that regulation by free market competition is not sufficient.

12.3: Further studies

Figure 34 is similar to the thesis' scope outlined in figure 2. It shows the main themes described and/or discussed in the thesis. In general it has been considered what can be done with the tools we have today; the SLAs we have, existing protocols and systems etc. The question is whether these tools are sufficient to solve the problems occurring.

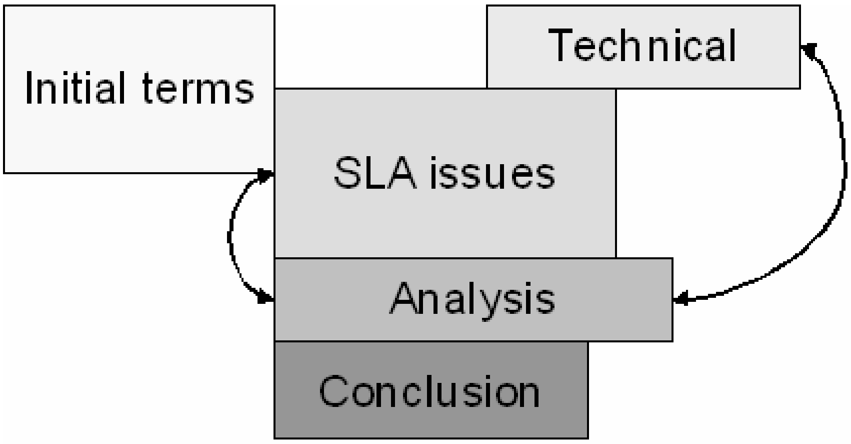


Figure 34: Future work

What could be done to continue this thesis' work is to look through all the "holes" discovered and look back to the technical protocol and system descriptions. Do they fill all the needs discovered in the analysis, or do they have to be expanded, changed or discarded? Maybe there even are needs for new protocols. Perhaps an effort done to set a standard for a common QoS-framework in networking would improve actor relationships?

A few SLA models are outlined in chapter 8 and some examples are proposed in chapter 9. Do they fulfill the SLA-demands made in the analysis? Do the models described in chapter 8 explain all the necessary matters in an actor-actor relationship? Can an SLA solve all the problems mentioned in chapter 8? Or do the proposed SLA models need to be improved? There may even be better ways to guarantee quality of service than using service level agreements. Perhaps better and more precisely defined laws are better than setting SLA-standards?

A.1: References

Books

“Data and Computer Communication” Sixth Edition, William Stallings, Prentice Hall International, Inc. 2000, ISBN 0-12-086388-2

“Internetworking with TCP/IP, Vol. 1: Principles, Protocols and Architecture” Forth Edition, Douglas E. Comer, Prentice Hall, Inc. 2000, ISBN 0-13-018380-6

Recommendations etc

ITU-T Rec. E.800 (08/1994): “Terms and definitions related to quality of service and network performance including dependability”

ITU-T Rec. E.801 (10/1996): “Framework for service quality agreement”

ITU-T Rec. E.860 (06/2002): “Framework of a service level agreement”

ITU-T Rec. Y.100 (06/1998): “Series Y: Global Information Infrastructure – General”

ETSI’s ETR 003, October 1994: “Network aspects (NA); General aspects of Quality of Service (QoS) and Network performance (NP)”

ETSI’s ETR 138, Second edition December 1997: “Network Aspects (NA); Quality of Service indicators for Open Network Provision (OPN) of voice telephony and Integrated Services Digital Network (ISDN)”

ETSI’s EG 202 009-3, 02/2002: “User Group; Quality of telecom services; Part 3: Template for Service Level Agreements (SLA)”

IETF’s RFC 1633, June 1994: “Integrated Services in the Internet Architecture: an Overview”

IETF’s RFC 2475, December 1998: “An Architecture for Differentiated Services”

IETF’s RFC 2805, April 2000: “Media Gateway Control Protocol Architecture and Requirements”

EURESCOM Project P1203, May 2003: “The Operators’ vision on systems beyond 3G, - Business modelling for systems B3G” Eskedal, Venturin, Grgic, Andreassen, Francis, Fischer, Danzeisen

EURESCOM Project 806-GI, June 2000: “EQoS – A common Framework for QoS/network provider in a multi-provider environment - Deliverable 4”

Technical Report DSL Forum TR-058, September 2003: “Multi-service architecture & framework requirements” Elias and Ooghe

Articles etc

[1] “Traffic Engineering with Traditional IP Routing Protocols” Fortz, Rexford and Thorup, IEEE Communications Magazine, October 2002, p118-124

[2] “SDH pocket guide”, Wavetek Wandel Golterman - Communication Test Solutions: SDH, page 1-39, http://www.lkn.ei.tum.de/studium/veranstaltungen/pra/sdh_Pocket_Guide_english.pdf [15.01.2004]

[3] “Satisfying the hunger for IP QoS” Hilde Hemmer, lecturer (første amenuensis) at Oslo University College, Electronics and Electrical Engineering department, spring 2001

[4] “Multiprotocol Label Switching (MPLS)” Web proforum tutorial, http://www.iec.org/online/tutorials/mpls/index.html [20.02.2004]

[5] “MPLS Advantages for Traffic Engineering” George Swallow, IEEE Communications Magazine, December 1999, p. 54-57

[6] “Evolution of Multiprotocol Label Switching” Viswanathan, Feldman, Wang and Callon, IEEE Communication Magazine, May 1998, p. 165-173

[7] “Key Exchange in IPSec: Analysis of IKE” Perlman and Kaufman, IEEE Internet Computing, November-December 2000, p. 50-56

[8] “H.323” Web proforum tutorial, http://www.iec.org/online/tutorials/h323/ [05.03.04]

[9] ”Final report for the European Commission: IP Voice and Associated Convergent Services”, 28.January 2004, performed by Analysys

[10] “ATM traffic shaping” http://cell-relay.indiana.edu/cell-relay/FAQ/ATM-FAQ/d/d3.html [01.03.04]

[11] “Modeling and Performance Comparison of Policing Mechanisms for ATM Networks” Erwin P. Rathgeb, IEEE journal on selected areas in communications, vol.9, no. 3, April 1991

[12] “Modell for Service Level Agreement (SLA) – Service Leverings Avtale”, Post og Teletilsynet/Norwegian Post and Telecommunication Authority, February 2004

[13] Canada’s Innovation Strategy FAQs, http://broadband.gc.ca/pub/faqs/faqscomplete.html [22.03.04]

[14] “Learning about DSL” http://www.dslforum.org/about_dsl.htm [29.03.04]

[15] “Local Multipoint Distribution System (LMDS)” Web proforum tutorial, http://www.iec.org/online/tutorials/lmds/ [29.03.04]

[16] “A framework for Managing QoS in Multi-Provider Configurations” Jensen, Grgic, Espvik, Telenor Research and Development, Norway

[17] “Generalized Multiprotocol Label Switching: An Overview of Routing and Management Enhancements” Banerjee, Drake, Lang, Turner, Kompella, Rekhter, IEEE Communications Magazine, January 2001

[18] “QoS Handbook” COM2-C54-E, ITU Study group 2 – Contribution 54, March 2004

[19] “Planning for Telecommunications Disaster Recovery” by Paul Kirvan, http://www.nedrix.com/presentation/1003/Breakout%20Sessions/Kirvan_Telecom.pdf [29.04.04]

[20] “LOV 1988-05-13 nr 27: Lov om Kjøp.” (Sale of goods law), last edited 01.07.2002 http://www.lovdata.no/all/nl-19880513-027.html [10.05.04]

[21] “LOV 2002-06-21 nr 34: Lov om forbrukerkjøp” (Law for Consumer Purchases), last edited 01.07.2002 http://www.lovdata.no/all/hl-20020621-034.html

[22] "LOV 2000-12-21 nr 105: Lov om opplysningsplikt og angrerett m.v. ved fjernsalg og salg utenfor fast utsalgssted", last edited 09.05.2003 http://www.lovdata.no/all/hl-20001221-105.html

[23] AOL Webmaster Info – Glossary http://webmaster.info.aol.com/glossary.html [14.04.04]

[24] Products & Services: e-Business glossary http://www.paint.org/resources/glossary.cfm [19.05.04]

[25] Symbian OS technology, Symbian Glossary: http://www.symbian.com/technology/glossary.html [20.05.04]

[26] Wireless Glossary, Thinkative http://www.thinkative.com/wireless/glossary.html [20.05.04]

[27] Henning Schulzrinne, http://www.cs.columbia.edu/~hgs/internet [05.03.04]

[28] Presentation made by Petter Møller, Statens Forvaltningstjeneste at a NORTIB seminar 01.04.2004, and made available at http://www.nortib.no/gjenseminarinfo.asp?SID=33

[29] Session Initiation Tutorial, Kuthan and Sisalem http://www.iptel.org/sip/ [05.03.04]

[30] "Differentiating Network services" http:/www.telenor.com/rd/publisering/dns00.shtml [12.05.04]

[31] "Krangler om SMS-krøll" Aftenpostens morgennummer, 03.01.2004

[32] "LOV 2003-07-04 nr 83: Lov om elektronisk kommunikasjon (ekomloven)", last edited 01.05.2004 http://www.lovdata.no/all/hl-20030704-083.html

Miscellaneous

Alcatel Telecommunication Review, 1st Quarter 2003

"Sammensmelting av Tale- og data-tjenester i Neste Generasjons Nettverk" – Hovedprosjekt ved Høyskolen i Oslo, avd. For ingeniørutdanning, spring 2002, written by Nina Sørsdal and Jørgen Helgheim

A.2: Abbreviations

AAL – ATM Adaptation Layer

ABR – Available Bit Rate

ADSL – Asynchronous DSL

AH – Authentication Header

AS – Autonomous System

ATM – Asynchronous Transfer Mode

BGP – Border Gateway Protocol

BI – Business Interface (in SLA relationship)

CAC – Connection Admission Control

CBR – Constant Bit Rate

CCITT - Consultative Committee on International Telegraphy and Telephony, former name of ITU before it was renamed in 1993

CDV – Cell Delay Variation

CLP – Cell Loss Priority

CP – Content Provider

CPE – Customer Premises Equipment

CRC – Cyclic Redundancy Check

CRM – Customer Relationship Manager

DiffServ – Differentiated Service

DCL – Digital Carrier Loop

DLCI – Data Link Connection Identifier

DSL – Digital Subscriber Line

DSLAM – Digital Subscriber Line Access Multiplex

DWDM – Dense Wavelength Division Multiplexing

ESP – Encapsulated Security Payload

ETSI – European Telecommunications Standards Institute

FDM – Frequency Division Multiplexing

FEC – Forward Equivalence Class

FSN – Full Service Network

FTP – File Transfer Protocol

GK – Gate Keeper

GII – Global Information Infrastructure

GMPLS – Generalized MPLS

GUI – Graphical User Interface

HDSL – High data rate DSL

ICMP – Internet Control Message protocol

IEC – International Electrotechnical Commission

IEEE – Institute of Electrical and Electronics Engineers

IETF – Internet Engineering Task Force

IGP – Interior Gateway Protocol

IKE – Internet Key Exchange

IN – Intelligent Network

IntServ – Integrated Services

IP – Internet Protocol, implemented in networks as version 4 and/or version 6

IPSec – IP Security

ISDN – Integrated Services Digital Network

IS-IS – Integrated System – Integrated System

ISAKMP – Internet Security Association and Key Management Protocol

ISM – Industrial-scientific-medical frequency band (902-928 MHz, 2400-2483 MHz, 5725-5780 MHz)

ISO – International Organization for Standardization

ISP – Internet Service Provider

ITU-T – International Telecommunication Union – Telecommunication sector

KAM – Key Account Manager

LAN – Local Area Network

LDP – Label Distribution Protocol

LEX – Local Exchange office

LMDS – Local Multipoint Distribution System

LMP – Link Management Protocol

MGCP – Media Gateway Control Protocol

MGW – Media GateWay

MPLS – Multi Protocol Label Switching

MSS – Maximum Segment Size

MTU – Maximum Transfer Unit

NAT – Network Address Translation

NGN – Next Generation Networks

NM – Network Management

NO – Network Operator

NOK – Norwegian Crones

NT – Network Termination

Oakley – Oakley Key Determination Protocol

OPT – OPTimal routing

OSI – Open Systems Interconnection

OSPF – Open Shortest Path First

OXC – Optical Cross Connection

PABX – Public Access Branch eXchange

PC – Personal Computer

PCM – Pulse Code Modulation

PCR – Peak Cell Rate

PLMN – Public Land Mobile Network

POTS – Plain Old Telephone System

PPP – Point-to-Point Protocol

PSTN – Public Switched Telephone Network

QoS – Quality of Service

RAS – Registration, Admission and Status

RFC – Request For Comment

RSVP – Resource reSerVation Protocol

RTCP – RTP Control Protocol

RTP – Real-time Transfer Protocol

SS#7 – Signaling System Number 7 – used i.e. for signaling in PSTN

SA – Service Availability (ref SLA)

SA – Service Association (ref. IPSec)

SCR – Sustainable Cell Rate

SDH – Synchronous Digital Hierarchy

SDP – Session Description Protocol

SDSL – Symmetrical DSL

SF – Statens Forvaltningstjeneste (office for public administration)

SHDSL – Symmetrical High bit rate DSL

SIP – Session Initiation Protocol

SLA – Service Level Agreement

SMS – Short Message Service

SOHO – Small Office/Home Office

SP – Service Provider

SSL – Secure Sockets Layer

STM – Synchronous Transfer Mode

SUA – Service UnAvailability

TCP – Transport Control Protocol

TDM – Time Division Multiplex

TE – Terminal Equipment

TI – Technical Interface (in SLA relationship)

ToS – Type of Service

TTL – Time To Live

UBR – Unspecified Bit Rate

UDP – User Datagram Protocol

UPC – Usage Parameter Control

URL – Uniform Resource Locator

VBR – Variable Bit Rate, is known in two forms; real-time (rt-) VBR and non-real-time (nrt-) VBR

VC – Virtual Circuit

VCC – Virtual Channel Connection

VCI – Virtual Circuit Identifier

VDSL – Very high-speed DSL

VoATM – Voice over ATM

VoD – Video on Demand

VoDSL – Voice over DSL

VoIP – Voice over IP

VoP – Voice over Packet

VP – Virtual Path

VPC – Virtual Path Connection

VPI – Virtual Path Identifier

VPN – Virtual Private Network

WAN – Wide Area Network

WDM – Wavelength Dimension Multiplexing

WLAN – Wireless Local Area Network

xbps – x bits per second, where x can be k – kilo, M – Mega, G – Giga etc

xHz – Hertz – unit for frequency, where x can be k – kilo, M – Mega, G – Giga etc