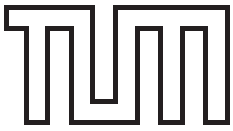


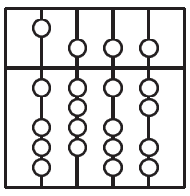
INSTITUT FÜR INFORMATIK

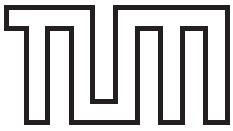
DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Bachelor Thesis

Modelling the Usage of IT Services

Author:	Marta Galochino Rodriguez
Proposed by:	Prof. Dr. Heinz-Gerd Hegering
Tutors:	David Schmitz Andreas Hanemann





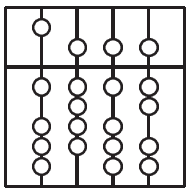
INSTITUT FÜR INFORMATIK

DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Bachelor Thesis

Modelling the Usage of IT Services

Author:	Marta Galochino Rodriguez
Proposed by:	Prof. Dr. Heinz-Gerd Hegering
Tutors:	David Schmitz Andreas Hanemann
Deadline:	15th March 2006



Hiermit versichere ich, dass ich die vorliegende Diplomarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 15th. March 2006

. . . . . .

(Unterschrift des Kandidaten)

Abstract

In today's IT service market, customers urge providers to grant guarantees for quality of service (QoS) which are laid down in Service Level Agreements (SLAs). To satisfy customers and to avoid penalties, service providers have to ensure that the agreed SLAs are met. Therefore, it is necessary to be able to effectively obtain the information about how different services are being used.

The aim of this thesis is the modelling of the usage of IT services. By monitoring the received service quality and the current and expected future service usage, this information is gained that will later on be used accordingly to different purposes. It could, for example, be used to obtain some statistics that reveal whether a service is increasing in popularity and therefore should be provided with more resources. In the case of failure in a service, the information gained through monitoring the usage can be used to assess the expected costs and to select an appropriate recovery alternative. Besides this short term perspective, a malfunctioning service has an impact in the whole service provisioning and by analysing this impact a lot of information will become available. The results of the impact analysis can be employed to identify critical resources and to improve the service provisioning.

For a view of the inner structure of service provisioning one needs to have in depth knowledge of the dependencies of the offered services on subservices and resources as well as the customers' SLAs, their QoS parameters, and the current service usage.

Today, service providers often achieve this by relying on the experience of their employees, a practice which has several drawbacks. Important influence factors are likely to be left out in the decision making process or employees can leave the company which will lead to an information loss.

It is for these reasons that a new methodology is needed to avoid the above mentioned problems. The creation of a model of IT usage of services can help address these problems in a more efficient way.

Zusammenfassung

In heutigem IT Service-Markt drängen die Kunden die Diensterbringer, Dienstgütegarantien (Quality of Service oder QoS) in den Dienstvereinbarungen (Service Level Agreements oder SLAs) anzubieten. Um Kunden zufrieden zu stellen und Vertragsstrafen zu vermeiden, müssen Diensterbringer sicherstellen, dass die vereinbarten SLAs eingehalten werden. Folglich ist es notwendig, die Information darüber, wie unterschiedlich Dienste verwendet werden, effizient zu bekommen.

Das Ziel dieser Arbeit ist das Modellieren der Nutzung von IT-Diensten. Durch das Überwachen der erhaltenen Dienstgüte und der gegenwärtigen und erwarteten zukünftigen Dienstnutzung wird diese Information gewonnen, die später dementsprechend zu unterschiedliche Zwecken verwendet werden kann. Sie kann zum Beispiel verwendet werden, um einige Statistiken zu erhalten, die aufdecken, ob ein Dienst sich in seiner Popularität erhöht, und folglich mit mehr Ressourcen versehen werden sollte. Bei einem Dienstausfall in einem Dienst kann die Information, die durch die Überwachung der Nutzung gewonnen wurde, verwendet werden, um die zu erwartenden Kosten festzusetzen und eine passende Wiederherstellungsalternative auszuwählen. Außer dieser Kurzzeitperspektive wirkt sich ein Dienstausfall in der ganzen Dienstbereitstellung aus und indem man die Auswirkung analysiert, wird viel Information daraus zur Verfügung gestellt. Diese Auswirkungsanalyse kann eingesetzt werden, um kritische Resourcen zu identifizieren und die Dienstbereitstellung zu verbessern.

Für eine Einsicht in die innere Struktur der Dienstbereitstellung muß man eingehendes Wissen, sowohl von den Abhängigkeiten der angebotenen Dienste von den Subdiensten und von den Ressourcen, als auch von den SLAs der Kunden, ihre QoS-Parameter und der gegenwärtige Dienstnutzung haben.

Heute erzielen Diensterbringer häufig dieses, indem sie sich auf die Erfahrung ihrer Angestellten verlassen, was einige Nachteile mit sich bringt. Wichtige Einflußfaktoren werden so möglicherweise in dem Entscheidungsprozeß ausgelassen oder Angestellte können die Firma verlassen, was zu einem Informationsverlust führt.

Aus diesen Gründen ist eine neue Methodik erforderlich, die die oben erwähnten Probleme vermeidet. Die Erstellung eines Modells der IT-Dienstnutzung kann helfen, diese Probleme in einer leistungsfähigeren Weise zu regeln.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The issue addressed in this thesis is the development of a concept for service usage modelling. Nowadays monitoring the usage of IT services often limits itself to the retrieval of information to produce a bill. That implies that, as far as a service is working, there will be no changes within the system that the dynamic of the usage might point at. If the services' usage modelling was more thoroughly done and with the intention of gaining information about the state of the system, it would be possible to attend those deficiencies that the daily usage of services is showing. Moreover, if a service breaks down because a subservice or a resource has broken down, a service provider needs to find the cause of it and fix the problem. We see then that some monitoring is necessary. An error in, for example, a subservice has an impact on the rest of services that depend on that subservice. To analyse the impact of these failures A. Hanemann, D. Schmitz and M. Sailer [HSS05] proposed the 'Impact analysis framework' that is shown in Figure 1.1.

Here subtasks which need to be performed during the impact analysis are defined and appropriate components are introduced. For some components there are already existing approaches, while others will have to be addressed in detail in future work. It is the aim of this thesis to present a first approach to the component 'service usage monitoring and prediction'.

The framework depicted in Figure 1.1 shows the main components communicating with each other and carrying out their task, which is performing the impact analysis and fault recovery. The beginning of this framework is determined when one or more resources fail. Services depending on these resources will also fail and some customers/ users will be encountering problems with the services offered. For some of them, the malfunctioning of the services will not represent too much trouble, but for others it may have serious repercussions. The service provider is supposed to react as quickly and efficiently as possible to restore the full functionality of the services which failed in the first place. Providers need to know first of all what are the affected resources and services and also what customers are affected by their malfunctioning. A provider must react quickly to restore the services or face the penalties specified in the Service Level Agreements (SLA). That is why, when different customers are having problems with different services and they use these services in a completely different way, providers face the task of having to decide which problem should get priority. In order to take the right decision, they need to consider two things:

- On the one hand they need to know the current usage of the service that is failing and the predicted future usage until the service is back to normal.
- On the other hand, they have to consider the SLA and calculate with the knowledge of the current and future usage of the service both the cost of repairing the service meeting SLA parameters and so avoiding penalties and also the cost of repairing violating SLA and therefore taking into account the penalties.

Then they have to compare both costs and decide what to repair first. If for example only a few small customers are affected and the penalties are relatively low, it will be more economic and subsequently

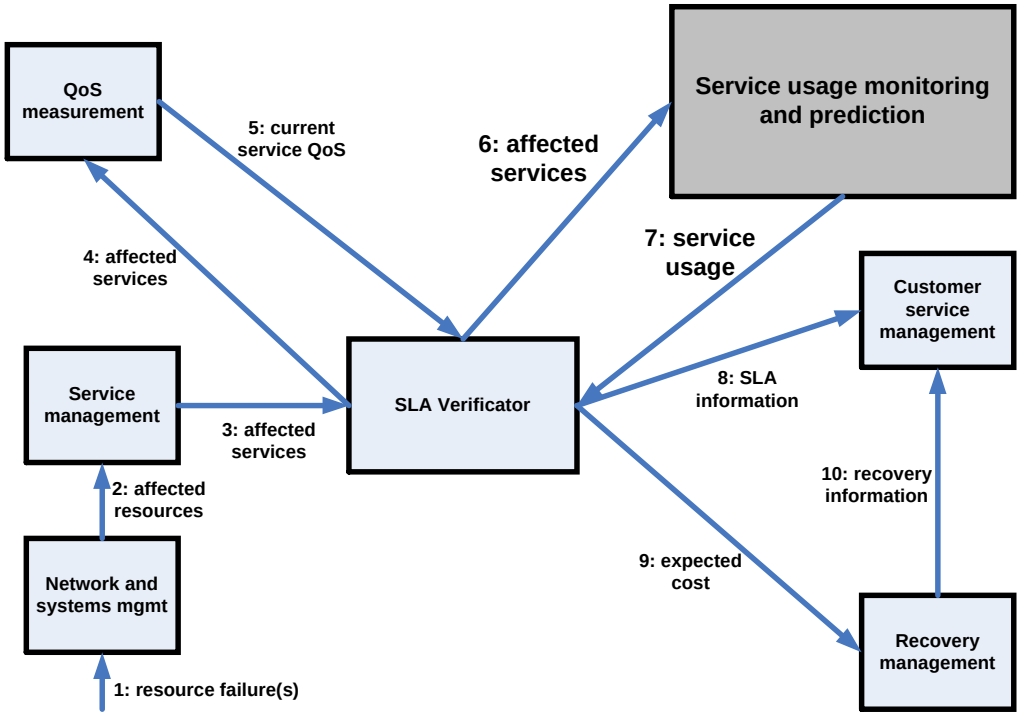


Figure 1.1: Simplified Impact Analysis Framework

better, to assign to the sorting of this problem a lower priority so that another problem that is affecting a lot more users can be sorted out first.

Today's impact analysis and recovery is mostly done by hand, using undocumented expertise and using best practices. This means that when, for example, staff is unavailable (e.g. through illness, leave or having left the company altogether), the service provider has difficulties to react in the rapid and efficient way that is wished. In other words, the problem resolution mechanism is unreliable and inappropriate.

Therefore the idea is to support and partially automate these best practice workflows. It is necessary to identify which components are needed and analyse related work for the reusability of components. Some concepts and tools will be adopted, while others that are missing will have to be developed.

For some of the components in the figure, there is already a concept. For other components the concept is being developed and here in this thesis the concept for the component 'Service Usage Monitoring and Prediction' is being proposed.

The main focus will be placed on the current service usage and the forecasting of future usage based on current and past values. The existing usage measurement and some accounting concepts have been adopted.

1.1 Need for new features in the monitoring of IT service usage

Up until now the standard has been to monitor the service usage according to those features that are going to be reflected in the bill for the services taken. Depending on the kind of access a user has to a specific service, it will be monitored. Or, in other words, a user paying per volume will be monitored in such a way that the total volume used is a known number for the provider. A provider will monitor the time using a

service when a user is paying per unit time. Since the monitoring nowadays is done with the purpose of producing a bill, the monitoring is incomplete in a sense. This results in the difficult and time consuming task of performing a check for each individual user to make sure that the service usage is correctly reflected in the bill. On top of that, the SLAs make the job of monitoring service usage even more complicated as each customer has an SLA that needs to be adhered to. So a service provider not only has to monitor the service usage individually but also the adherence to the pertinent SLA. A model to monitor service usage that is applicable to any service and any user is the aim to achieve.

1.2 Current monitoring deficiencies

Consider a service provider and 3 of its customers. Customer 1 has an SLA 1, customer 2 has an SLA 2 and customer 3 has an SLA 3. The service provider offers 98% availability and reflects this in SLA 1 and SLA 2. In SLA 3 the customer has the additional requirement that the 2% period of downtime lasts no longer than 5 minutes as anything more than a 5 minute downtime would have serious repercussions for his business. Service provider and customer agree on that and sign. Now comes a period of time when the service is down. The provider has nowadays no way of determining on time who is affected. An extensive job has to be done to find out the impact of this downtime. At the present time, there is no support for the experts involved in returning the service back to normality, and so they are facing situations where, for example, a parameter has been forgotten and the service had to be down for even longer time than expected.

Different customers having problems with different services make providers face the task of having to decide which problem to fix first. In order to take the right decision, they need to consider the current usage of the service that is failing and the predicted future usage until the service is back to normal so that they can calculate what the loss / cost for not repairing it straightaway is. This information allows a provider to decide the best recovery alternative. When only a few customers are affected and the penalties associated with the service degradations are relatively low, a service provider might, for example, assign a low priority to the sorting of this problem due to economic reasons to be able to deal first with another problem that affects a lot more users and involves more or higher penalties.

1.3 Proposal for monitoring IT service usage

Due to the above mentioned problems, it is desirable to create a service usage model that could be used for different purposes (for example SLA conformance). The proposed model intends to overcome the problems that service providers find nowadays when they collect the information needed for them to attend the demands of today's customers. Customers and users demand more and more that the services used are reflected genuinely in their accounts (instead of receiving a superficial and undetailed report of their takings). Today customers and users want to know exactly how they are using services and verify that they do get a good deal. Through customers' satisfaction, it has been shown [BL99] that the usage of services increases. In order to gain the information that customers are demanding, a model is here of advantage. This model should help service providers meter and collect the usage of their users.

The first step for the development of the model was to analyse the relevant features for monitoring IT services' usage and here two scenarios were presented and the requirements were derived. The second step was the proposal of a model that meets the requirements. Third and final step was to consider the current and the future usage of services for the creation of the forecast model.

This thesis is divided into 5 chapters:

Chapter 1 introduces the reader to the thesis. Chapter 2 introduces the MNM service model and makes an analysis of the requirements where the relevant features for the modelling of two example services namely an e-mail and a web hosting scenario are shown. Two different scenarios are shown here and a catalogue of requirements is derived. Chapter 3 presents the state of the art, which is analysed and evaluated with

respect to the requirements. Chapter 4 is concerned with the development of a general model for IT service usage as well as a comparison of today's methodology with the use of the model. Chapter 5 closes with the application of the general model of chapter 4 to the two example services: the e-mail and web hosting services. Finally chapter 6 summarises and concludes this thesis.

Chapter 2

MNM service model and requirements' analysis

This chapter begins with an introduction to the Munich Network Management (MNM) service model in Section 2.1. This model contains the basis that will lead to the creation of a model for the dynamics of IT service usage. According to the MNM service model, a service view consists of two main functionalities: usage and management. This division will be adopted for the analysis of the requirements .

Section 2.2 is concerned with an analysis of the requirements. Here two scenarios will be analysed, their requirements will be derived and the relevant features for the modelling of these scenarios will be reviewed. These two scenarios are e-mail and webhosting. After having seen these examples, a general catalogue of requirements will be developed in Section 2.3.

Finally the chapter closes in Section 2.4 with a summary.

2.1 The MNM Service Model

The MNM Service Model [GHH+01], which was developed by the Munich Network Management Team, is a generic model for IT service management. A distinction is made between customer and provider side. The customer side contains the basic roles customer and user, while the provider side contains the role provider. The provider makes the service available to the customer side. The service as a whole is divided into usage which is accessed by the role user and management which is accessed by the role customer.

The service view in Figure 2.1 shows a common perspective of the service for customer and provider. The functionality of a service is in the foreground, abstracting from details on its implementation. Everything that is only important for the service realization is not contained in this view. It should be pointed out here, that the explicit modelling of the service management on the right hand side of the model in contrast to service usage on the left hand side is an important feature of the model.

Relevant to the purpose of this thesis are the following components:

- **Service:** A service is defined as a functionality that is provided with a certain quality and cost at a Service Access Point (SAP). The customer side has two participating roles: *users*, who actually use the service and a *customer*, who is interested in maintaining a subscribed service and therefore performs all the management activities on the customer side. On the provider side all necessary activities for enabling service usage as well as service management are performed by the role *provider*.
- **Functionality:** The functionality of a service consists of two parts: the usage functionality covers the interactions needed by the user. These interactions represent the actual purpose of the service.

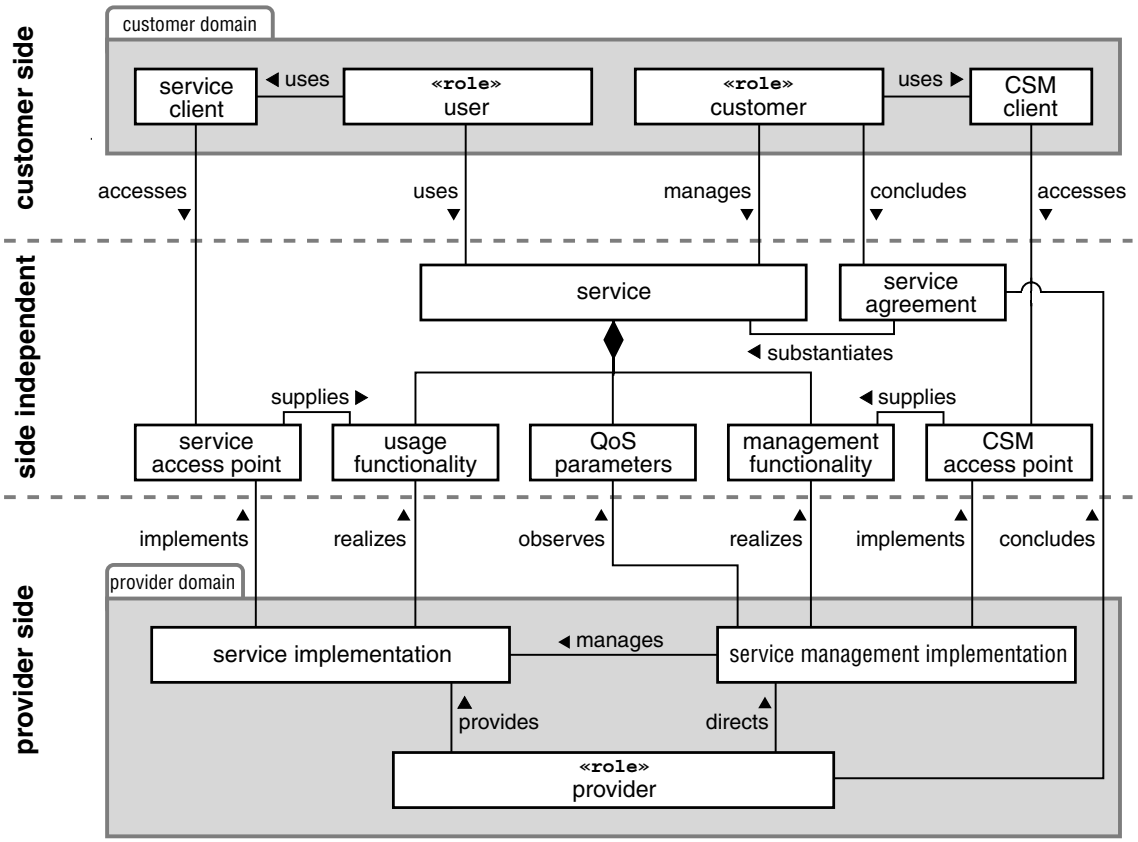


Figure 2.1: View of the MNM service model

Additionally, interactions beyond the service's purpose are needed to fulfill the customer's needs, to monitor and control the provider's service provisioning as well as for payment. The management functionality comprises these interactions.

- **QoS Parameter:** The above mentioned functionalities must satisfy a set of QoS parameters. These parameters define the minimum required service quality in order to be useful for the customer side. The QoS parameters are qualitative values.
- **Service Level Agreement (SLA):** Contract between service provider and customer where all details referring the conditions of the service provisioning are specified.

The introduction of the MNM service model here was motivated by the fact that it can be used as an starting point where the analysis of the requirements will base on. Let us now then begin with a couple of scenarios from which the requirements will be analysed.

2.2 Analysis of two scenarios

In this section we show two different scenarios: e-mail service in Section 2.2.1 and web hosting service in Section 2.2.2. For each of these scenarios, requirements are derived and the relevant features for modelling IT service usage are presented. Having seen these scenarios and their derived specific requirements, a generally applicable catalogue is developed in Section 2.3. As it will be seen, the MNM service model view was adopted for the division of the catalogue of requirements in requirements related to the service view and those related exclusively to general management.

2.2.1 E-mail scenario

An IT service that is widely used is the e-mail service so the E-mail service offered by the Leibniz Supercomputing Center (LRZ) [LRZa] will serve as an example IT service. The LRZ, which is the computing center for the Munich universities and runs the scientific network in Munich, offers e-mail access for students and staff of the universities and the LRZ itself.

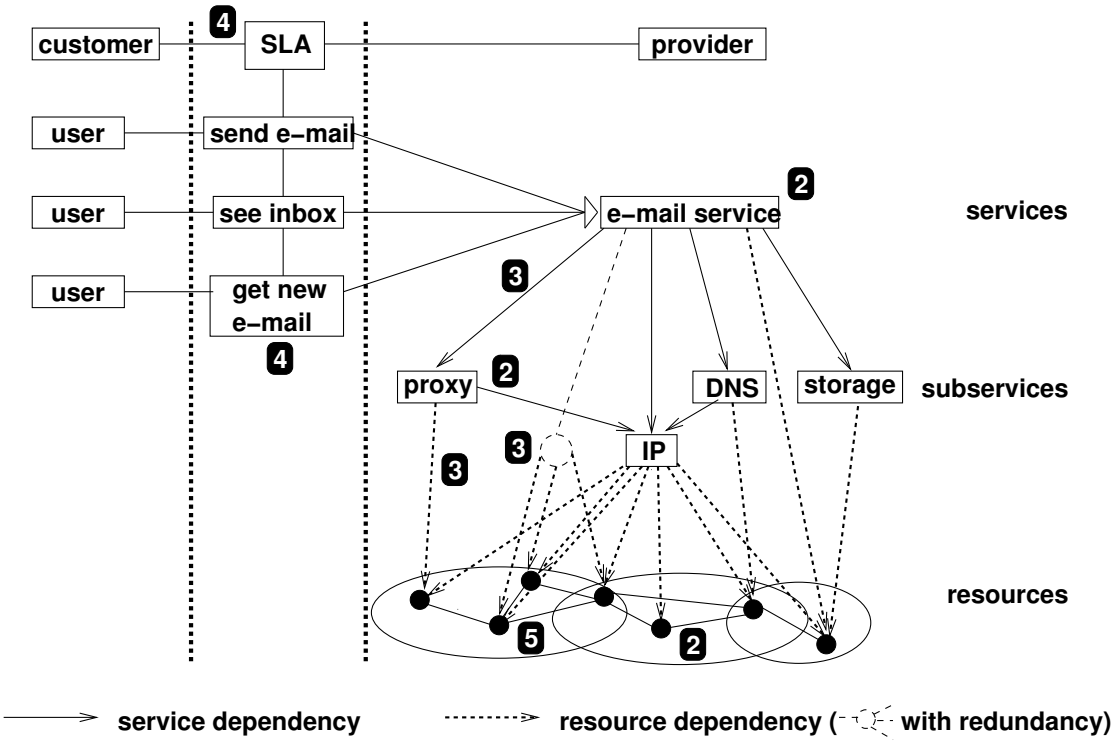


Figure 2.2: E-Mail Scenario

Figure 2.2 shows the main components of this E-mail scenario. The roles provider, customer and user are shown. The provider offers the E-mail service to his customers. The customers allow several users to use the service functionalities by granting them to create their personal e-mail accounts. The users have access to the service functionalities; here in this scenario these are ‘send e-mail’, ‘see inbox’ and ‘get new e-mail’. These functionalities and others that are not shown in the above figure of this scenario comprise the E-mail service. Between provider and customer there is an SLA where all details concerning the quality of the service (QoS) are specified. Examples of QoS parameters are availability and delivery delay.

SLA for this scenario

Availability: 99.9% during business hours, weekly basis.

Delay: Sending of mail (to next mail domain) takes less than 25 minutes in 99% of the cases.

Maximum size per Mail: 10 MB

Maximum amount of mails per user a day: 500

Penalty: 10.000 EU per month, immediate possibility to change the provider in case of violation.

On the provider side, the e-mail service is provided using subservices. In the scenario these subservices are DNS, proxy service, connectivity service (IP), and storage service. Both services and subservices depend on resources which they are provisioned upon. These are e.g., network components, network links, an end

system's main memory, or processes running on a server. As depicted in Figure 2.3 a service can depend on more than one resource and a resource can be used by one or more services.

Requirements derived by this scenario

The following issues are motivated by the scenario and need to be addressed in detail although not entirely in this thesis.

- **Modelling of services:** A service model is needed covering the service features relevant for the impact analysis, especially the QoS parameters. The QoS modelling should be independent from the provider's service implementation. This is a requirement made by customers in order to be able to compare the offers of different providers. This thesis is concerned with the modelling of the dynamics of IT service usage and although the services and resources that comprise a service need to be known in order to monitor its usage it must be remarked that the modelling of the services and resources is the concern of other theses.
- **Dependency modelling:** In the scenario, there are three kinds of dependencies, i.e., dependencies between different services, dependencies between services and resources, and dependencies on the resource level. It is important to identify the characteristics of these dependencies and their necessary attributes. An example for this is an appropriate modelling of resource redundancies. Here again it must be pointed out that the dependency modelling that is meant here is not the concern of this thesis.

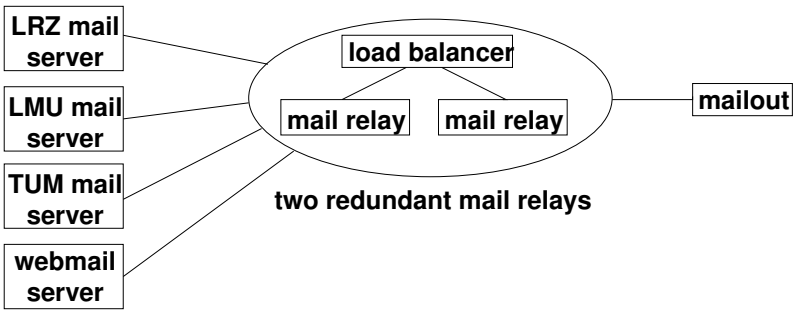


Figure 2.3: Subservices'dependencies of this e-mail scenario

Figure 2.3 shows the resources which are used for the provisioning of the E-mail service. For load sharing and redundancy reasons, the dispatching of mails is performed by two different mail relays controlled by a load balancer. Mails received can be accessed from different incoming mail servers depending on the user group (mail for LRZ employees itself or mail for staff/students of the Munich universities LMU and TU which are the customers of the LRZ). The E-mail service can also be accessed by using a dedicated web mail server. Examples of resource failures causing QoS degradations for the E-mail service might be:

  - Failure of a hard disk storing the E-mail inboxes on one of the incoming mail servers
  - Mail dispatching at the mail relays is very slow because of too many mails waiting in the mail queue (possibly caused by a lot of spam mails)
  - The load balancer is not working properly (causing high delay and packet loss) because of wrong routing tables

- **SLA:** As the impact analysis is performed with respect to SLAs, an SLA modelling is needed based on the QoS modelling mentioned above in page 6.
- **A monitoring component:** Besides an SLA repository, a monitoring component is required to check the provided QoS and to determine the effect onto the SLAs . To determine the actual consequences

more precisely, service usage monitoring is needed. The data retrieved from the monitoring should be kept in a database so that information is available for different purposes like predicting what the usage of a service will be like in a period of time, or assessing resources for optimization.

Relevant features for modelling the usage of the E-mail service

Most online services and Internet Service Providers (ISPs) offer e-mail, and most also support gateways so that you can exchange mail with users of other systems. Usually, it takes only a few seconds or minutes for mail to arrive at its destination. This is a particularly effective way to communicate with a group because you can broadcast a message or document to everyone in the group at once.

1. **Transfer distance:** An e-mail that stays within a domain is less likely to become lost and the delivery time will be less than if the e-mail has to go from domain to domain. In the last case, that is when the e-mail crosses domains, it also has to be considered how many domains it actually transverses, and the distance to be transversed. Obviously an e-mail sent to the antipodes takes longer to be received than an e-mail that remains within a domain and goes across a physical space of 1 Km.

2. **Size of the message:** Short messages are in general easier to handle. The bigger the message gets, the more resources that will need to be used, and the more time that will be needed to process it. A service provider usually supports the sending of e-mails up to a certain size. Once that size has been surpassed, it is for the service provider in terms of efficiency no longer desirable to further process an e-mail of excessive size because dealing with that large file would mean neglecting the efficiency of the rest of files and the general well being of the service offered would be in danger. Should it come to a failure in the transmission of data by TCP it would be necessary to retransmit the whole mail again, what is obviously undesired.

3. **Relevant content; little spam:** In addition to wasting people's time with unwanted e-mail, spam also eats up a lot of domain bandwidth. Consequently, there are many organizations, as well as individuals, who have taken it upon themselves to fight spam with a variety of techniques. Some online service providers have instituted policies to prevent spammers from spamming their subscribers.

4. **Amount of sent e-mails from one user:** An e-mail provider might restrict a user to send only a limited quantity of e-mail messages per hour or per session. By this way ISPs prevent their SMTP servers from overloading because SMTP servers spend much more time delivering these messages than receiving them from an e-mail client. There are also a lot of bulk e-mail accounts on the Internet that can be bought, which are located on huge high volume SMTP servers and capable of delivering thousands of e-mail messages per minute.

5. **Containing virus:** E-mail is now the most common way that viruses are transmitted between computers. The most common mechanism for this is in the form of an 'attachment' to the message. The attachment facility is normally used for e-mailing documents, images, and so on. However, it is also possible for attachments to contain programs which execute when the attachment is opened. Viruses can infect simply by reading, or in some cases, by previewing e-mail.

6. **Percentage of lost e-mail / percentage of received e-mail:** when someone sends a message, he expects it to be delivered and if this does not happen, the consequences might vary from disappointment to serious economic loss. Improving the quality of the services offered is nowadays the rule for every ISP.

7. **Confidentiality breaches:** Most confidentiality breaches occur from within a company. These breaches can be accidental, for instance by selecting a wrong contact in the To: field. However, confidentiality breaches can also be intentional, as Borland International Inc. experienced first hand: A Borland employee used the company's email system to send out confidential information to competitor Symantec, his new employer. The trade secrets included product design specifications, sales data and information regarding a prospective contract for which both companies were competing.

The employee and recipient were both charged with trade secret theft. Whether it is by mistake or on purpose, the result of the loss of confidential data is the same.

2.2.2 Web hosting scenario

Web hosting refers to the process of publishing a web site so that it is available to the world on the Web. Because probably every reader has already gathered some experience with browsing websites and so can have an idea to what this service is about, it was chosen to serve as an example.

The LRZ web hosting service [LRZb] hosts web sites for the Munich universities as well as for other research institutions. The following scenario will be defined using the structure of the LRZ web hosting service.

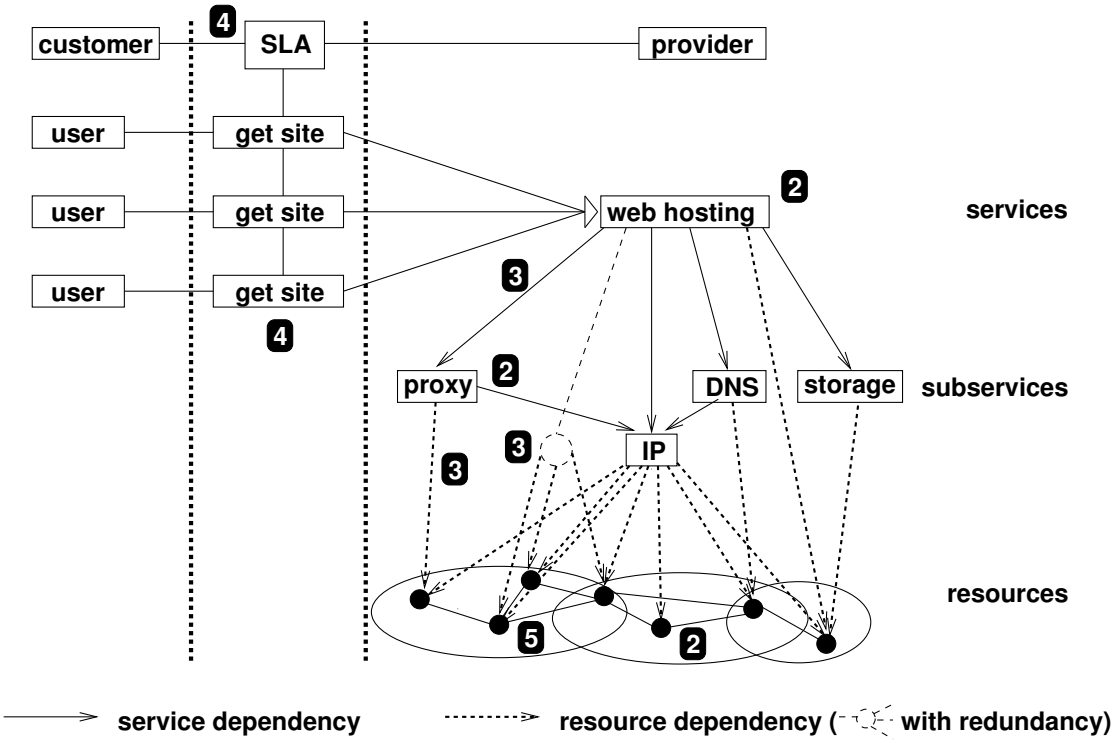


Figure 2.4: Web Hosting Scenario

Figure 2.4 shows the main components of the web hosting service provided by the LRZ. An SLA is signed between provider and customer and represents all those agreed interactions between provider and customer. Users access the functionalities of this service within the boundaries agreed on the SLA. In this figure the functionality 'get site' is being accessed by three different users. As it can be seen in the figure the web hosting service depends on subservices like the proxy, IP, DNS and storage subservices. There are also dependencies between subservices like the proxy and DNS subservices depending on the IP subservice. Subservices depend on resources and some resources depend on other resources. All these dependencies presented in this scenario are important for the modelling of IT service usage because monitoring the usage of a service implies monitoring the usage of each and all of its subservices and resources. Marking the boundaries of the service to be monitored makes the modelling a lot simpler.

Apart form the 'get site' functionality there are also other functionalities that belong to the web hosting service offered by the LRZ. A model of IT service usage should contain a list with all the different functionalities that are offered and are accessed by users and that need to be monitored from the provider.

The quality of the web hosting service offered is laid down in the SLA. An SLA contains usage and management parameters and in what follows the reader can see what a possible SLA for this scenario could look like.

Usage parameters:

- Availability: 99% during business hours, weekly basis
- Reliability: 3/100; this indicates the number of errors per total number of requests and responses generated or received by the service.
- Delay: 0.09 ms; the delay that occurs when accessing a website of average size 100KB, usually specified in milliseconds.
- Up-to-dateness: 30 sec; maximum duration until a content change comes into effect.
- Document accesses: 5; this specifies the total number of simultaneous requests that can be handled by the web hosting service.

Management parameters

- Data transfer per month: 1 GB
- Bandwidth for updating pages: 10 MB
- Maximum number of content changes per month: 10
- Time when the service provider is available for updating pages : Monday to Friday from 8AM to 5PM.
- Hours per month of downtime: 2 hours per month is the permitted duration of service unavailability due to failure.

Requirements derived by this scenario

The following issues which were derived from the scenario need to be addressed in detail although not entirely in this thesis.

- **Modelling of services and resources:** A service model is needed covering the service features relevant for the impact analysis. It is especially necessary to model the QoS parameters. A key requirement is that parameters are specified independent from a provider's service implementation. The provider may have subscribed services from other providers which drives the need for an implementation independent definition in order to reliably provide own services. In addition, the customers demand provider-independent QoS definitions to easily compare offers of different providers.

  This thesis is concerned with the modelling of the service usage and although the services and resources that comprise a service need to be known in order to monitor its usage it must be remarked that the modelling of the services and resources is the concern of other thesis.

- **Dependency modelling:** In the scenario, different kinds of dependencies can be distinguished. There are dependencies between different services, dependencies between services and resources, and dependencies on the resource level. It is important to identify the characteristics of these dependencies and their necessary attributes. For the modelling of dependencies between services and resources, redundancies in the service provisioning have to be covered by the modelling.

  Here again it must be pointed out that the dependency modelling that is meant here is not the concern of this thesis.

  Figure 2.5 shows the resources which are used for the provisioning of the Web Hosting Service. For load sharing and redundancy reasons the service is provided on five different web servers, therefore failures of certain servers do not result in an unavailable service. The load sharing is performed by

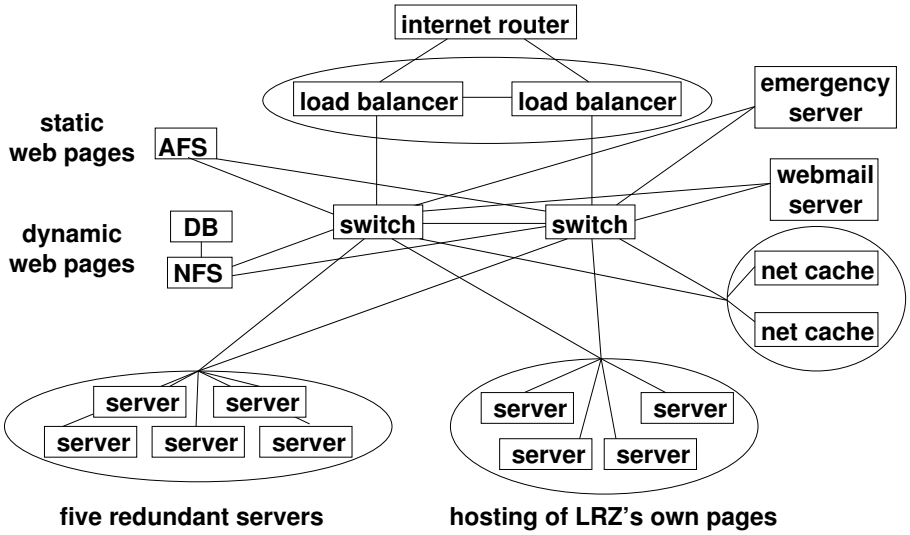


Figure 2.5: Subservices' dependencies of this web hosting scenario

a couple of load balancer switches. Static web sites are located in a distributed Andrew File System (AFS), while dynamic web pages are located in a Network File System (NFS). The LRZ's own web sites are located on four redundant servers. The e-mail service of the LRZ can also be accessed by using a dedicated web mail server. In case of severe network problems inside the LRZ an emergency server is available which only contains some web sites containing basic information.

When, for example, a web site is unreachable the service usage modelling will help identify the possible root causes. These could be a DNS problem, connectivity problem, wrong configuration of the load balancer, etc.

- **SLA:** As the impact analysis is performed with respect to SLAs, an SLA modelling is needed based on the QoS modelling mentioned above. In case a subservice has been outsourced to another provider, the consequences of a failure in this subservice also have to be considered. From a business point of view, it is necessary to ensure that a provider's SLA with a subprovider contains appropriate penalties. For instance, if the provider cannot meet the SLAs with its customers due to a failing subservice, these penalties have to cover the resulting costs. Accordingly, the SLA definition should allow for a derivation of such a mapping. A monitoring infrastructure is needed to measure the QoS as defined in the SLAs. Together with a history of past QoS violations, the current status of the SLAs can then be determined. The current service usage should be taken into account for performing the impact analysis. If e.g., a failure in a resource leads to a malfunction of a service, but the service is currently not used, there is no impact on the SLAs at the moment.
- **A monitoring component:** Besides an SLA repository, a monitoring component is required to check the provided QoS and to determine the effect onto the SLAs . To determine the actual consequences more precisely a service usage monitoring is needed. The data retrieved from the monitoring should be kept in a database so that information is available for different purposes like predicting what the usage of a service will be like in a period of time, or assessing resources for optimization.

Relevant features for the monitoring of the usage of the web hosting service

The idea here is to review a few of the features that a service provider has to monitor to get the knowledge he/she needs to analyse the current situation and be able to predict what it will be like in the future.

1. **Size of a website:** On average, most web sites take up very little space, especially for the site itself. It's quite feasible to have a web site containing 100 pages that uses less than 5MB of web space.

Websites are usually at least double the initial web space requirement in order to allow for future growth.

2. **Data transfer (bandwidth):** When someone downloads anything from a site whether it is a single web page, several pictures or music files, this is counted as data transfer. Data transfer is also used when checking and downloading e-mail or uploading (publishing) a site.

   It can be very difficult to predict how much data transfer a site is likely to use. If a site is new, then it will generally not have a great many visitors using up the data transfer allowance. However, if a site contains many pictures or music, the data transfer is a lot bigger.

   Imagine that a site used 1 MB of disk space including all graphics and html files. If a visitor were to load every page on this site, they would have used 1 MB of data transfer. If the negotiated data transfer is 1 GB (1024 MB), this will be enough for 1024 visitors to view this site in one month. Obviously most visitors will not load every page on a site (unless there is only one page) so this is a rather extreme worst case scenario.

3. **Size of music files:** The problem with music files is that they use a lot of space and represent a high data transfer. The use of mpeg files is recommended since they take up roughly 10% of the original size for CD quality and sound the same. A typical song recorded at high quality lasting for 3 minutes and 30 seconds will use 35.3 MB as a .wav file. As an .mpg file, this will be reduced to roughly 3.5 MB which is still fairly large. A website with very long music files, will be accessed by only a certain type of user e.g. a user having a slow internet connection (e.g. via a 56 k modem) will have to wait considerably longer to download such a file than another user who has a broadband connection.

4. **Size of graphic files:** Pictures can take up large amounts of space and data transfer on a web site. The amount of space used by a picture depends on the format of the picture. For example, .bmp pictures are not compressed and take up far more space than is required. Compressed formats like .jpeg or .gif use up significantly less space. For pictures with a lot of colours is .jpeg the best format and if a picture only contains several colours (such as a logo or diagram) converting these to the .gif file format will reduce the required space.

   Ways to cut down on data transfer used on a site: For a gallery of pictures, provide a page showing thumbnails (small versions of your pictures) so that the visitor to the site can choose which pictures to view. This not only cuts down on data transfer but also speeds up viewing for a visitor. Make sure that not too many pictures (or large pictures) are placed on the home page as this will generate a larger amount of data transfer for every visitor.

5. **Size of text files:** For very long text files also applies what was just mentioned for graphic files.

6. **Amount of music, graphic or/and text files:** A large quantity of music, graphic or/and text files amounts to a potential massive data transfer that could cause a lot of potentially harmful traffic.

7. **Amount of people visiting a website:** A website that has lots of visits per day must be kept as full functioning as possible. If some of the services offered are down or the website itself is down the number of users affected could be very big.

8. **Amount of people altering a website:** People altering the content of a website entails data transfer and that needs to be reflected in the usage of services.

9. **Amount of links in a website:** When the main webpage of a website includes lots of links to the other webpages of that website a global vision of the contents of the website is provided and so the data transfer is likely to be higher than if the main webpage had contained few references to the rest of the webpages that compose the website.

10. **Loading time of a website:** Depending on the size of a website the loading of a website will be greater or shorter. A large website means a long loading time for that website and a lot of data transfer.

11. **The prices for additional transfer and space** so a customer does not pay a fortune for going over his/her plan's limits.

12. **Guaranteed uptime** Uptime is the percentage of time that a web site is working. For example, if some host has an uptime average of 99.86%, this means that a site will be down for a total about 1 hour each month. A customer requires a refund for times when his/her host is down (no host stays up 100% of the time).

2.3 Development of a catalogue of requirements

The idea here is to develop a general catalogue of requirements by abstracting what we have seen in the previous section so that the resulting catalogue of requirements can be applied to any service. First of all we propose a general description of the requirements followed by a more detailed one.

Before starting with the requirements, it is important to review the impact analysis framework that we saw in Figure 1.1 in more detail, because this thesis is about one of the components of this framework and to fully understand it, one needs to see where it fits in. Now the framework will be fully shown as proposed by Hanemann, Schmitz and Sailer in [HSS05].

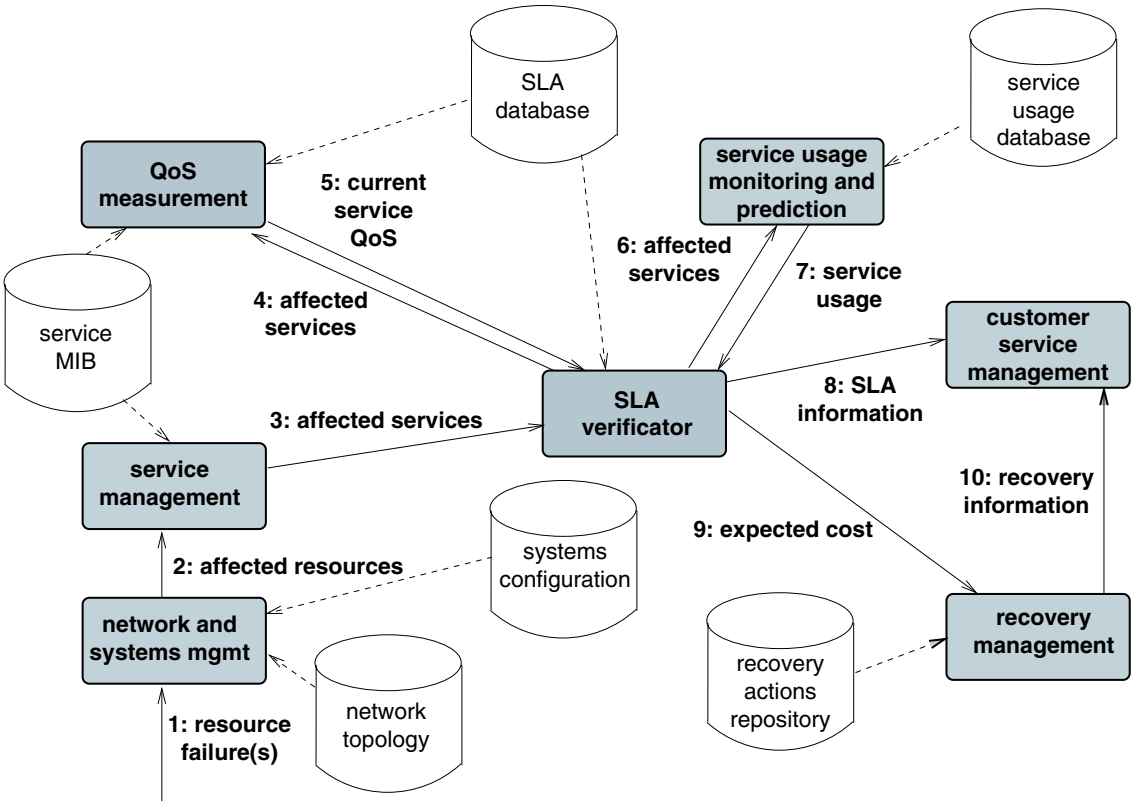


Figure 2.6: Impact Analysis Framework

Figure 2.6 shows the main components in grey boxes and the cones represent databases that are accessed to get information. The framework initial state is shown by the arrival of a/some resource degradation (for example, failure) at the network and systems management (step 1).

Quickly after this event, other resources which are affected by the failure get identified (step 2) by means of retrieving information in the databases. In the service management, the services that use the malfunctioning

resources are identified. The severity of the impact is also derived. The list of all affected services including the expected QoS degradation is transferred to the SLA verificator (step 3). The list of affected services is sent to the QoS measurement (step 4) and information about the severity of the service quality degradation is transferred back to the SLA verificator (step 5).

This thesis is concerned with steps 6 and 7, i.e. service usage measurement and prediction. Measuring the service usage and keeping a database with the results of this measured values serves as a base to analyse and predict what the service usage will be like in the period of time between the occurrence of a degradation and the time when the service functionality has been fully restored. To determine the expected costs for not correctly providing the service, the current service usage by customers (and their users) is taken into account. Prediction models can be used to get an expected service usage for future time intervals. To get such usage information, the affected services are sent to the service usage measurement and prediction component. That information is crucial for a service provider for the simple reason that this knowledge will allow the provider take the right decision about what to repair first, and when exactly he needs to act to prevent something worse from happening. If, for example, a service is not working properly, but it is only used by few customers whose SLAs do not contain severe penalties, then the impact can be classified as low.

Up until now there has not been an efficient way to see when a service needs to be paid attention to and so a service provider has been relying on the experience of their qualified employees to react to service degradations to the best of their knowledge. With service usage modelling, the job of these employees can be partially automated and supported. This means for a service provider that qualified employees can make a better use of their working time.

The result of monitoring and predicting service usage is received by the SLA verificator (step 7).

To keep the customers informed about the status of the services with respect to the SLAs, the information gathered so far is transferred to the Customer Service Management (CSM) (step 8).

From the collected information the SLA verificator can now determine an expected cost function over time for not repairing the resource failure(s). This information together with the resource failure(s) and corresponding repair possibilities are reported to the recovery management (step 9).

The recovery management decides which recovery steps should be performed and tracks the recovery progress. The customers are kept informed via the CSM (step 10).

Having explained why the modelling of IT service usage is important for service providers, it will be proceeded to the making of a catalogue of requirements that will help create the model.

2.3.1 General requirements

G1 - Framework Support

As the usage of IT services is a part of the ‘Impact Analysis Framework’, the desired model must support the component ‘Service Usage Monitoring and Prediction’ of the above explained framework.

G2 - Transparency

The metering of the usage of the service must occur transparently for both the user and the customer. Neither user nor customer need be aware that their usage of the service is being monitored. They are only aware of the different functionalities of a service.

G3 - Genuineness

Customers require that a bill reflects genuinely the usage that they have made of a service, not by showing whether the agreed QoS has been met, but by showing what the actual QoS has been like together with an accurate description of the actual usage of the service. The knowledge of the actual usage of services allows a provider to improve the provisioning.

2.3.2 Requirements related to the service view

These requirements are customer oriented and are derived from the service view of the MNM service model that was presented in Section 2.1. Therefore both usage and management functionalities are here referred from the customer point of view. Usage functionalities refer to the functionalities that are accessed by the users and that are directly related to the usage of the service. Management functionalities refer to the customization of the service according to user's needs. For example when a user needs a bigger mailbox, a management functionality will perform this task.

The International Telecommunications Union (ITU) [ITU] is a consortium of telecommunications companies worldwide who have, among other things, defined a series of recommendations that describe how a telecommunications management network (TMN) should be operated. The ITU members have adopted a model of management functions, often referred to as the FCAPS model after the initials of each of the major functions it describes.

TMN Function	**Description**
Fault Management:	Fixing what is broken.
Configuration Management:	Controlling the usage functionality so it works the way you want.
Accounting Management:	Knowing who is using how much of what, and maybe billing them for it.
Performance Management:	Making sure it all works acceptably quickly.
Security Management:	Controlling who can do what.

Table 2.1: FCAPS

The management functionality of a service involves many parts and it is important to consider them all in the requirements.

U1 - Specification of the functional granularity

A service provider must specify the granularity of the functionality (usage and management) that will be monitored. Some examples related to usage functionalities of, for example, the e-mail service are specifying that the incoming and the outgoing e-mail will be monitored. Within the incoming e-mail, the size of these e-mails and whether they come from the same domain as the receiver or from a different one will also be monitored. Spam could also be monitored.

As for the management related functionalities, they are usually subsumed under the interface 'Customer Service Management' (CSM) and enable customers to individually monitor and control their subscribed service. There is also a subdivision that must be specified. For example, a service provider could specify that should there be any problems with the 'incoming mail' functionality there is a 'contact help desk' function covering the management of this functionality. Another management functionality could be the opening of a new e-mail account or the alteration of the mailbox size. In other words, these management functionalities are visible to a customer and have been made available to support and manage the usage functionalities.

U2 - Specification of the time granularity

A service provider must specify how often it will be monitored. For example every session or every transaction. Depending on the functionality to be monitored an adequate time granularity will be chosen. A management functionality is the function that allows a customer to change granularity from, for example, monitoring how many mails were delivered to a mailbox every hour to monitoring the amount of delivered mails every 10 minutes.

U3 - Specification of what to monitor

A service provider must specify what will be monitored per session. The extent of that monitoring varies considerably depending on the functionality to be monitored. An example of usage functionality related to

the web hosting service is ‘load website’ and its equivalent management functionality is the ‘alter website’ function that allows the user customize the usage functionality according to his needs.

U4 - Distinction between actual and negotiated usage

A service provider requires that a clear distinction is made between the actual usage of a service and the negotiated usage as in the SLA. A customer might have negotiated that a message should not take any longer than 1 minute to be received but, if an error occurs and the delivery of the message takes 5 minutes, it is important to know the real delay to react appropriately. Actual usage must be metered and compared with the negotiated usage.

2.3.3 Requirements related to general management

These management requirements are referred to general management aspects that a provider needs to consider to offer a service and therefore, not necessarily supporting a concrete usage functionality as it was the case in last section. They are related to the internal management of a service provider.

Management related aspects form an integral part of a service, and need thus to be considered.

M1 - Identification of the resources used

A requirement for the making of the model is that resources that are being used when a service is being used are identified.

M2 - Dependencies on subservices

Dependencies of a service on different subservices need to be identified. The relationship between different subservices and their dependencies should be clear for the model. The consideration of all resources redundancies can help, for example, in case of a resource failure to decide what to do next.

2.3.4 Requirements related to prediction

P1 - Selection of relevant data with a certain granularity

In order to make a prediction it is necessary that some data is available to be selected according the forecasting method chosen and the particular situation to be forecast. The data selection must follow a certain functional or/and time granularity.

P2 - Selection and application of a forecasting method

Among all forecasting methods and under considerations of a particular forecast a matching forecasting method must be chosen and applied.

2.3.5 Catalogue of requirements

The requirements that we have identified and explained in the previous sections are here in form of a table summarised to a catalogue of requirements. This catalogue serves on the one hand as an instrument to assess the state of the art (see Chapter 3) and on the other hand to develop a new solution (see Chapter 4).

2.4 Summary

In this chapter we first introduced the MNM service model with the intention of setting a foundation for our analysis of the requirements.

Catalogue of requirements
General Requirements
G1 - Framework support
G2 - Transparency
G3 - Genuineness
Requirements related to the service view
U1 - Specification of functional granularity
U2 - Specification of time granularity
U3 - Specification of what to monitor
U4 - Distinction between actual and negotiated usage
Requirements related to general management
M1 - Identification of resources used
M2 - Dependencies on subservices
Requirements related to prediction
P1 - Selection of relevant data with a certain granularity
P2 - Selection and application of a forecasting method

Table 2.2: Catalogue of requirements

Secondly two different scenarios were analysed, their requirements were derived and the relevant features to model the usage of these services were presented.

Lastly a general catalogue of requirements generally applicable was developed. The requirements were classified under three categories: requirements related to the service view, requirements related to general management and requirements related to prediction.

This catalogue represents the starting point that will help assess the present status of the subject we are concerned with and also it represents a platform to develop a proposal.

Chapter 3

State of the art

In the previous chapter a catalogue of requirements for the modelling of IT service usage was created. In this chapter the representative specifications of standard groups will be reviewed in respect to the modelling of IT service usage and will be evaluated.

Section 3.1 reviews the specifications of standard groups and analyses them with respect to modelling IT service usage. Each specification is subsequently evaluated.

Section 3.2 illustrates how SLAs are nowadays monitored. An example of a commercial product widely used for the monitoring of IT services is given. The usability of today's methodology to monitor SLAs is evaluated.

Section 3.3 covers general prediction concepts that are relevant for the purpose of this thesis and explains how these concepts can be adopted within the concept of this thesis.

The chapter closes with a summary in Section 3.4 where the conclusions of the analysis of the state of the art are reviewed.

3.1 Related Work in the industry and standard groups

The following sections are concerned with the specifications of the standard groups that are relevant to the modelling of IT service usage and their evaluation. These specifications are TINA, ITIL, eTOM, CIM and AGIMO.

3.1.1 Telecommunication Information Networking Architecture (TINA)

The TINA service architecture [Con97] introduces a set of concepts, principles, rules and guidelines for constructing, deploying, operating and withdrawing TINA services. TINA defines 'service session' when referring to service usage and associates this concept with the beginning and end of a service usage. The concept 'service transaction' is also introduced by TINA offering in that way an alternative granular unit to measure service usage.

TINA describes the separation of access and usage, and within usage the separation of service session and communication session. As the reader will have a chance to see later on in this thesis, these relevant concepts will be partially adopted for the making of the model of IT service usage.

TINA also describes the environment in which services operate together with the way the different components are combined, and the way they interact. Although its focus is set on specifying a software architec-

ture in order to implement (telecommunication) services rather than on modelling services, it still offers a good reference that can be used for the modelling of services.

As the elements of the service architecture are specified in various models and a modelling method is missing, in order to use this work for the purpose of this thesis lots of care must be taken to be consistent throughout the modelling of specific scenarios. Nevertheless, the business model introduced in [MNM+00] can serve as a source for checking completeness regarding the service model.

The evaluation conclusion obtained after reading TINA is positive since it represents a reference to the modelling of IT service usage.

3.1.2 IT Infrastructure Library (ITIL)

An approach to how IT services are used can be found in ITIL, an infrastructure library developed in the UK. ITIL is a widely accepted approach to IT service management in the world. ITIL provides a cohesive set of best practice, drawn from the public and private sectors internationally. It is supported by a comprehensive qualifications scheme, accredited training organisations, and implementation and assessment tools. The best practice processes promoted in ITIL support and are supported by the British Standards Institution's standard for IT service Management (BS15000).

Currently the ITIL documents are facing a restructuring process resulting in six volumes of which two are published: Service Support [Iil00]and Service Delivery[Iil01]

Service Delivery is concerned with the management of the IT services. A number of management practices to ensure that IT services are provided as agreed between the service provider and the customer are examined. Reviewing existing services, producing and monitoring the Service Level Agreement (SLA), establishing priorities, planning for service growth, performance monitoring, workload monitoring, resource forecasting, demand forecasting and modelling are all relevant aspects from where this thesis can profit. The document consists of 5 disciplines and can be accessed in http://www.itil-itsm-world.com/delivery.htm

Although the focus of ITIL is on IT management, it does offer valuable hints towards IT service usage and therefore the evaluation of ITIL with respect to the purpose of this thesis is a very positive one.

3.1.3 Enhanced Telecom Operations Map (eTOM)

The Enhanced Telecom Operations Map (eTOM) [For05] is the ongoing TeleManagement Forum (TMF) initiative to deliver a business process model or framework for use by service providers and others within the telecommunications industry. The TMF eTOM describes all the enterprise processes required by a service provider and analyzes them to different levels of detail according to their significance and priority for the business. eTOM is more formal than ITIL by specifying a process framework that postulates a set of business processes that are typically necessary for service providers to plan, deploy and operate their services.

The specification concentrates on giving service providers valuable hints and recommendations concerning what processes are necessary to provide an overall integrated service management covering several services. Although the 'Service Planning and Development', 'Service Configuration' and 'Service Quality Management' of the 'eTOM Business Process Framework' represents a valuable reference for this thesis, the majority of the work is concerned with points that are out of the range of this thesis and so the contribution of eTOM for the purpose of the modelling IT service usage is not very great.

3.1.4 Common Information Model (CIM)

The Common Information Model (CIM)[For02] is a conceptual information model for describing management that is not bound to a particular implementation. CIM is composed of a Specification and a Schema.

The Schema provides the actual model descriptions, while the Specification defines the details for integration with other management models.

The CIM Schema itself is structured into three distinct layers of which two are relevant for this thesis: the Core Schema is an information model that captures notions that are applicable to all areas of management and the Common Schema are information models that capture notions that are common to particular management areas, but independent of a particular technology or implementation.

The most important aspect of CIM for this thesis is that CIM aims to address both FCAPS management (fault, configuration, accounting, performance and security management) and to support the abstraction and decomposition of services and functionality.

Finally it must be mentioned that the great extension of CIM makes this source a valuable one, but the formality and abstraction of CIM endangers the application of concepts for the purpose of this thesis.

3.1.5 AGIMO's 'better practice' in online service delivery

The Australian Government Information Management Office (AGIMO) has created 'The Better Practice in Online Service Delivery Program' [(AG04] to help ensure that government continues to be an effective and exemplary user of IT. They have also created 'Better Practice Checklists' to help web managers, business unit owners, and others quickly enhance their understanding of a range of issues associated with technology enabled government. In particular one of these checklists: 'AGIMO Checklist in website usage monitoring and evaluation' [AGI] offers significant contributions to the monitoring of service usage. This checklist has been created to help agencies to evaluate the usage of their websites and although it only refers to website usage and is not intended to be comprehensive, it represents a valuable source of information to this thesis.

3.1.6 Assessment of the related work

Now that the specifications from standard groups has been reviewed and evaluated, a general assessment with respect to fulfillment of the requirements mentioned in Section 2.3 will be illustrated through the means of a chart.

Figure 3.1 presents an overview of the results of the evaluation of the works reviewed in previous sections. From this table the reader can quickly see whether these pieces of work comply with the requirements or not. The symbol '+' was used to indicate compliance. The symbol '-' indicates no compliance and the symbol 'O' was used to indicate that the requirement is fulfilled but with some restrictions.

	TINA	**ITIL**	**eTOM**	**CIM**	**AGIMO**
G1	-	-	-	-	-
G2	+	+	+	+	+
G3	-	-	-	-	O
U1	+	-	+	-	-
U2	+	-	-	-	-
U3	-	-	-	-	O
U4	-	-	-	-	O
M1	O	O	O	O	-
M2	O	O	O	O	-
P1	-	+	-	-	O
P1	-	+	-	-	O

Table 3.1: Assessment of the related work

TINA complies well with three requirements and although M1 and M2 are touched by TINA it is often in a context outside the purpose of this thesis. ITIL does not comply with any of the requirements related to

the service view but it is the only piece of work of this thesis that actually comments on forecasting and predicting. With respect to the requirements related to general management it is also the case that although the main points are addressed they are reviewed with another perspective and so these requirements are marked with a O to indicate that there exist some restrictions. The contribution of eTOM to this thesis is very reduced since only one of the general requirements is fulfilled and it must be remarked that all other reviewed standard specification also fulfilled this requirement. That is why eTOM is not well suited to model a particular service provisioning scenario. CIM offers a foundation for modelling dependencies between services and resources, but is not concerned with any of the requirements related to the service view. M1 and M2 were mentioned in CIM but with a different perspective from this thesis. Finally AGIMO offers a great reference for monitoring IT service usage but the problem is that these checklists focus on specific matters and so they are not always applicable in this thesis. For five of the requirements a 'O' was the result of the evaluation. The fact that these requirements are mentioned in a checklist that focuses on only a specific IT service, i.e. the checklist is not meant to be applicable for any IT service, is the reason for this marking.

To conclude this chapter it is worth mentioning that all these approaches focus on particular issues related to IT service. Although the monitoring aspect is mentioned in many cases, a general service usage model that can be used in different scenarios and environments is missing. The better practice checklists offered by ITIL and AGIMO that were reviewed in Sections 3.1.2 and 3.1.5 provide a lot of valuable information about monitoring services but they do not offer a service usage model. This thesis explores this issue and makes a proposal in Chapter 4.

3.2 Monitoring Service Level Agreements (SLAs)

'Service level agreement (SLA)', is a contract between a service provider and the customer that stipulates and commits the service provider to a required level of service. An SLA specifies then the services offered and the quality with which they will be offered.

Rather than the actual usage of services it is often today the rule to monitor SLAs. SLA conformance can be done in different ways and here it is worth mentioning that the violation of an SLA is related to penalties. In this thesis it will be shown that if service provider knew the current and future usage of services, they would be supported in many of his/her tasks. When, for example, an error occurs and affects the quality of several service provisionings, providers face the task of fixing the problems. Nowadays they only have the information retrieved by the monitoring of SLAs and this information is not complete because it only covers the aspects on which customer and provider agreed. If service providers monitored the current service usage, they will gain a lot more information and therefore they would be supported in the process of decision making. The knowledge gained from the monitoring of the current service usage will reveal that it will sometimes be better not to do anything at all about a service not functioning well, than to repair it quickly. Because today the rule is to monitor the parameters that define a certain quality of service, it is worthwhile at this point to have a look at that a customer will be demanding from a service provider.

QoS parameters a user demands

A customer needs to define the kind of services he wants to receive from a service provider. He might need to know things like how long he should wait before he could transmit his data, how long the receiver will take to receive a message, whether the receiver will get an error-free message or not, how high the error probability is, whether he can be sure he will always be able to send data at any time, how much he should pay for the service with the level of quality he is getting, how flexible services are, whether he can directly control QoS parameters or if a customer is protected against any unauthorized monitoring or modification of data. All those facts can be summarised in the following QoS parameters:

- Time related parameters: latency, jitter

- Bit rate: constant or variable
- Cost
- Reliability
- Flexibility
- Availability
- Security level

Time related parameters (latency and jitter) and bit rate (constant or variable) parameters can specify data transfer delay. Flexibility of a service means that customer and service provider agree on a range of quality of service that is acceptable. Availability means that services are available when a user requests it. The larger data transmission capacity and the lower failure rate a network has, make a service less likely to be faulty and therefore a higher availability will be provided. Security level is a parameter through which a customer can specify if the network service provides any security protection for transferred data.

Although these quality of services are offered and a provider commits himself to them, in the real world, the following facts can seriously affect QoS:

- Random hardware failure
- Flow congestion

Although random hardware failure happens rarely, when it happens it can seriously damage a provider's quality of service image. Flow congestion also happens very rarely, but when it does happen, the service provider cannot guarantee any QoS parameters.

Since nowadays the main part of the monitoring process for IT services is mainly done using different commercial products, an example is given to the reader of one of the many software products that are available.

The 'Service Monitor' from ElectraSoft Software is a program that monitors a website and informs the owner of the website if and when the site goes down and for how long. If a service or an event fails, it will alert the owner of the website with a variety of methods and the down time will be logged. If a customer pays a service provider to host his/her site or service, he should be able to get at least a partial refund from the site hosting provider if any serious down time occurs. The network connection monitor feature of this network monitoring software allows you to write a simple script for testing any service performance.

ElectraSoft software is used by many well known establishments including military, government, education, medical, business, and individuals. Users of ElectraSoft software include the United States Armed Forces, the United States Senate, the University of Utah, the Central Iowa Health Care System, the San Diego Health Department, IBM, and millions of small businesses and individuals.

After having presented the main aspects considered when monitoring SLAs and having seen a example commercial product, it must be explained now that even though monitoring SLAs is essential, the problem nowadays is that the real usage of services is not usually being monitored. By limiting the monitoring to SLA conformance lots of information are not being taken into account to, for example, optimize performance. Optimization of services should be a direct consequence of the results of the daily monitoring of service usage. Only in this way a service provider gains first hand the information about what needs attention.

For the purpose of this thesis the monitoring of SLAs is necessary but also other important aspects that are being neglected at the moment. It is for this reason that it can be affirmed that although SLAs monitoring is necessary for the creation of a model for IT service usage it is not sufficient. In other words, the information gained through SLA monitoring brings very useful information but also neglects other very important information.

3.3 Prediction concepts

Foresight is probably one of the most important parts of management. Correctly recognizing emerging changes in the usage of services and accurately predicting future ones are prerequisites for the future success of a service provider. Strategy, or at least a good part of it, must be based on some form of forecasting and a realistic assessment of the uncertainty involved in all types of future predictions.

At the short-term, the role and contribution of forecasting towards planning is clear. Forecasts are the 'most likely value' or a best estimates about the future. These values are found by identifying and extrapolating established patterns and/or existing relationships. They are accurate as long as the future is a continuation of the past. Should it be thought it will not be so, judgment must be used to adjust the extrapolative forecasts. At the same time the uncertainty surrounding the forecasts is accepted and measured.

In the medium term, the role of forecasting is also well defined, although the uncertainty surrounding the forecast is higher because of unusual or unexpected events. Predictions about the medium term are based on the 'average' of say, past recoveries as well as an interpretation of the special circumstances surrounding each event.

In the long term the role of forecasting is less obvious. Forecasting is needed to develop foresight about what is to come and evaluate the extent and directions of forthcoming changes as well as their impact. Forecasting is indispensable for identifying potential opportunities as well as dangers in the business environment and appreciating the extent and impact of future uncertainty.

The critical question is how correct forecasts can be made and how these forecasts can be effectively incorporated in order to improve a service provider provisionings. In order to make a prediction about the future usage of IT services, it is necessary not only to access the data contained in the IT service usage database but also to apply some of the concepts and methods of prediction like, for example, the ones explained in the book of Makridakis [MSH98].

Makridakis calls the 'full range of major forecasting methods'. These comprise the traditional time series methods of decomposition, exponential smoothing, simple and multiple linear regression and Box-Jenkins' ARIMA models. Further to those, the 3rd edition very wisely includes some more advanced forecasting methods such as dynamic regression, neural networks, state space modelling as well as some 'new ideas for combining statistical and judgmental forecasting' amongst others. The authors have aimed at providing a complete description of the methods' essential characteristics. They have also presented the advantages and drawbacks of the methods with the intention of helping in model selection. In what follows relevant forecasting techniques will be explained.

3.3.1 Smoothing techniques

Smoothing techniques are suitable for preparing short-term forecasts for a number of different items. A typical example would be the forecast of the amount of e-mails that a user will be sending or receiving within a period of time. The nature of these situation can be assumed to change only slightly during each subsequent time period. Obviously there can be occasions on which it might change a considerable amount in a single period, but generally speaking many of these items exhibit a fairly stable series of values over a short time horizon. These are the techniques used to predict unemployment figures on a short-term basis. The most common and those that will be discussed first are moving averages and exponential smoothing. These approaches to forecasting are nonstatistical in nature and are based largely on simple intuitive principles. The historical data is used to obtain a 'smoothed' value for the series which becomes the forecast for some future period. Thus in applying a smoothing technique there are two steps to the process. In the first some kind of smoothed value is computed based on historical data, and in the second that value is used as a forecast for some future time. The basic notion inherent in moving averages, exponential smoothing, and other forms of smoothing techniques is that there is some underlying pattern in the values of the variables to be forecast and that the historical observations of each variable represent the underlying patterns well as random fluctuations.

- **Simple moving average (SMA)**

  The term 'moving average' is used because as each new observation becomes available a new average can be computed and used as a forecast. As an example, the share price of a company on the stock exchange will be monitored over a period of time and a forecast of the closing price for a month will be made. A simple moving average is calculated by computing the average (mean) price of a share over a specified number of periods. For example: a 5-day simple moving average is calculated by adding the closing prices for the last 5 days and dividing the total by 5.

  10 + 11 + 12 + 13 + 14 = 60

  $\frac{60}{5} = 12$

  The calculation is repeated for each price bar on the chart. The averages are then joined to form a smooth curving line - the moving average line. Continuing the example, if the next closing price to be taken into account of in the simple moving average calculation is 15, then this new value would be added and the value of the oldest day, which is 10 in this case, would be dropped. The new 5-day simple moving average would be calculated as follows:

  11 + 12 + 13 + 14 + 15 = 65

  $\frac{65}{5} = 13$

  Over the last 2 days, the SMA has increased from 12 to 13. As the values for new days are added, the values for the old days will be subtracted and the simple moving average will continue to vary over time.

Day	Daily Close	10-day SMA
1	67.5	
2	66.5	
3	66.44	
4	66.44	
5	66.25	
6	65.88	
7	66.63	
8	65.55	
9	65.63	
10	66.066	66.39
11	63.94	66.03
12	64.13	65.79
13	64.5	65.6
14	61.81	65.24
15	61.88	64.8
16	62.5	64.46
17	61.44	63.94
18	60.13	63.3
19	61.31	62.87
20	61.38	62.4

Figure 3.1: Moving averages

In Figure 3.1, a 10-day SMA is being calculated using the closing prices of the company's shares. Obviously day 10 is the first day on which it will be possible to calculate the 10-day simple moving average. As the calculation continues, the value for the newest day is added and the value for the oldest day is subtracted. The 10-day SMA for day 11 is calculated by adding the prices of day 2 through day 11 and dividing by 10. The averaging process then moves on to the next day where the 10-day SMA for day 12 is calculated by adding the prices of day 3 through day 12 and dividing by 10.

The chart in Figure 3.2 is a plot that contains the data sequence in the table. The simple moving average begins on day 10 and continues.

This simple illustration highlights the fact that all moving averages are lagging indicators and will always be 'behind' the actual price. The share price of the company is trending down, but the simple moving average, which is based on the previous 10 days of data, remains above the price. If the price were rising, the SMA curve would most likely be below the curve representing the actual share price development. Because moving averages are lagging indicators, they fit in the category of trend

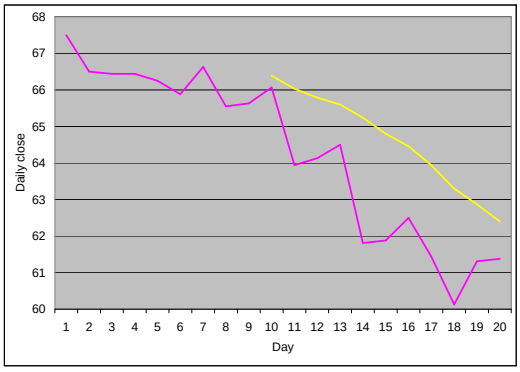


Figure 3.2: A 10-Day Simple Moving average

following indicators. When prices are trending, moving averages work well. However, when prices are not trending, moving averages can give misleading signals.

- **Exponential Moving Average (EMA)**

  In order to reduce the lag in simple moving averages, technicians often use exponential moving averages (also called exponentially weighted moving averages). EMA's reduce the lag by applying more weight to recent prices relative to older prices. The weighting applied to the most recent price depends on the specified period of the moving average. The shorter the EMA's period, the more weight that will be applied to the most recent price. For example: a 10-period exponential moving average weighs the most recent price 18.18% while a 20-period EMA weighs the most recent price 9.52%. The calculating and EMA is much harder than calculating an SMA. The important thing to remember is that the exponential moving average puts more weight on recent prices. As such, it will react quicker to recent price changes than a simple moving average.

  Exponential Moving Averages can be specified in two ways - as a percent-based EMA or as a period-based EMA. A percent-based EMA has a percentage as it's single parameter while a period-based EMA has a parameter that represents the duration of the EMA.

  The formula for an exponential moving average is:

  EMA(current) = ((Price(current) - EMA(prev)) x Multiplier) + EMA(prev)

  For a percentage-based EMA, 'Multiplier' is equal to the EMA's specified percentage. For a period-based EMA, 'Multiplier' is equal to $\frac{2}{1+N}$ where N is the specified number of periods.

  For example, a 10-period EMA's Multiplier is calculated like this:

  $\frac{2}{TimePeriods+1} = \frac{2}{(10+1)} = 0.1818(18.18\%)$

  This means that a 10-period EMA is equivalent to an 18.18% EMA.

  The results of an exponential moving average calculation are:

  For the first period's exponential moving average, the simple moving average was used as the previous period's exponential moving average. From period 11 onward, the previous period's EMA was used. The calculation in period 11 breaks down as follows:

  $(C - P) = (61.33 - 63.682) = -2.352$

  $(C - P) * K = -2.352 * .181818 = -0.4276$

  $((C - P) * K) + P = -0.4276 + 63.682 = 63.254$

The 10-period simple moving average is used for the first calculation only. After that the previous period's EMA is used.

Note that, in theory, every previous closing price in the data set is used in the calculation of each EMA that makes up the EMA line. While the impact of older data points diminishes over time, it never fully disappears. This is true regardless of the EMA's specified period. The effects of older data diminish rapidly for shorter EMA's. But, they never completely disappear.

These forecasting techniques attempt to identify, or at least approximate, that basic underlying pattern. Now another technique will be briefly presented that goes beyond trying to approximate the underlying pattern in a time series. This is the simple regression technique. It is here assumed not only that such a basic pattern exists but also that the form of that basic pattern is linear. This means that when the data is plotted it falls approximately along a straight line.

3.3.2 The simple regression technique

It is usually the case that forecasting involves a period of time. Thus, when this situation is plotted, the time variable is on the horizontal axis and the variable that is to be forecast on the vertical axis. Regression analysis is a technique that is not limited to this type of relationship only. A relationship between any two variables and then base a forecast of one of the values on the other is perfectly possible. As an example, consider the situation faced by a large mail-order house. Each day a tremendous amount of mail is received, much of which contains orders that have to be filled. The mailing department has noted over several months that the number of orders to be filled seems to be related to the weight to predict the number of orders that will have to be filled that day, and thus help to schedule the time of people who will fill those orders. As a first step in determining whether such a relationship exists, they have kept a record over several days of the weight of the mail each day and the corresponding number of orders. These pairs of values can then be plotted on a graph, in which a trend relationship between the weight of the mail and the number of orders is made clear. In this situation the department could approximate this relationship with a straight line. Then, when they receive a certain number of kilos of mail, they could use that straight line to forecast the number of orders. This procedure assumes a causal model or relationship between weight and the number of orders.

In the use of simple regression the starting points is the assumption that a basic relationship exists between two variables and can be represented by some functional form. Mathematically, it can be written as

$Y = f(X)$,

which simply says that the value of Y is a function of (or depends on) the value of X. In simple regression this is a straight-line relationship, and therefore the mathematical function can be written as

$Y = a + bX$

Since this is the general form of any linear relationship, it is important that the reader understand just what this means. Suppose that the value of X is zero. In such a case Y would have the value a. Thus a is the point at which the straight line intersects the Y axis. Referring again to the example above this would mean that when the weight is zero kilos the number of orders would have the value of a, which could be reasoned to be zero, since if no mail is received no orders are received. The value of b in this equation is called the regression coefficient and indicates how much the value of Y changes when the value of X changes one unit. Thus, if the number of orders are compared when the kilos of mail increase from 40 to 41, it would be expected an increase of b orders for the 41 kilos.

Following with the above linear equation, the values of a and b need to be estimated. These values are referred to as the parameters in the equation for a straight line. Several methods can be used to approximate these parameters. Perhaps the most straightforward technique is to plot the historical observations. Once this is done, the values of the parameters a and $b$could be read off the graph.

The regression equation is a statistical model, and thus it is possible to make statistical statements about the accuracy and significance of regressions. The use of these statistical properties will also allow to make

statements about the likelihood that future values will vary from the forecast by certain amounts and the accuracy of the coefficients *a* and *b*.

To finish this section it is necessary to mention the usability of these concepts for this thesis. A manager needing to make a forecast about any aspect related to IT service usage has the possibility to do so by applying the prediction concepts introduced in this section. It could, for example, forecast what the usage of a specific IT service would be like for a particular customer in the following month. In order to make this prediction, the service provider would need to define precisely what is to be forecast. That information together with his knowledge about forecasting techniques would allow him to select one forecasting method and according to this method the service provider would proceed to the selection of relevant data at a certain level of granularity. Finally the service provider would only need to apply the method and interpret the results.

3.4 Summary

In this chapter representative related work from standard groups and related to monitoring SLAs and prediction concepts were reviewed and assessed.

Regarding the monitoring of SLAs, it was mentioned that nowadays monitoring IT service usage limits itself to monitoring SLAs. The need of a service provider to collect and manipulate information about the current service usage was addressed. The conclusion is that not only SLAs should be monitored but also the current service usage.

With respect to prediction, relevant techniques were reviewed and their usability for this thesis to obtain information about future service usage according to a service provider's needs was addressed.

The result of this research is, that although valuable work has been done, there is no place to find a piece of work that complies with all the requirements that were proposed in the previous chapter.

Summing up, the prediction concepts that were found are directly applicable in this thesis but the rest of points concerning this thesis like 'monitoring', 'service usage' and 'modelling' are not exposed in the literature in a way that satisfies the requirements that were proposed, and so in what follows of this thesis, a proposal that unites these concepts will be made.

Chapter 4

Development of a model for the IT service usage

Models are the things we build to help us understand things better. When dealing with concepts of reality, it is often the case that abstractions are made and a model is developed. In situations where it is simply too costly to build the real thing, models are built to help understand things better. In short, models are simplifications or abstractions of reality intended to promote understanding.

Most of the models that are build are static in nature. That is, they are models that represent a snapshot of something at a particular point in time. Yet, reality is not static. Reality is constantly changing through our interactions with it, and the interactions between all of its parts; reality is dynamic in its nature. The object of this thesis is to produce a model of the dynamics of IT services so the question arises as to how to build a static model of a dynamic reality.

This is the way it was proceeded to get to such a model. To start with, a reflexion on what the characteristics of the dynamic usage of IT services are was needed with an eye on providers monitoring their customers service usage and then these characteristics were analysed. Then the requirements that we exposed in our catalogue in Section 2.3 to make the model were considered. The developed model is presented in Figures 4.1 and 4.2 and is fully explained.

4.1 Developing an IT service usage model

When talking about monitoring IT service usage, a lot of characteristics come to mind. What features a service provider should monitor is not an easy question to answer. Each service has its own functionality and what a service provider should monitor depends not only on the service itself but also on the purpose of monitoring. The development of a model should help a service provider optimize the services offered. In order to optimize services, a service provider must have the knowledge of what he is currently offering and how customers find these services taking into account their personal needs. The way to gain this information is through monitoring IT service usage. So now that the reader knows what a provider should do i.e. monitoring service usage and the purpose of it i.e. optimize services, the time has come to create a model as general as possible that can be applied to any IT service and that helps a service provider achieve the goal of improving the services offered. At this point it must be remarked that the model presented in this thesis does not intend to be complete, but an overview of some of the features that should be in a complete model. The reason why this is so is obviously the extent of the work that a complete model would entail and the fact that the time allowed to create this thesis is well shorter than that.

Figure 4.1 shows us five aspects that a service provider needs to consider for every service when monitoring is carried out.

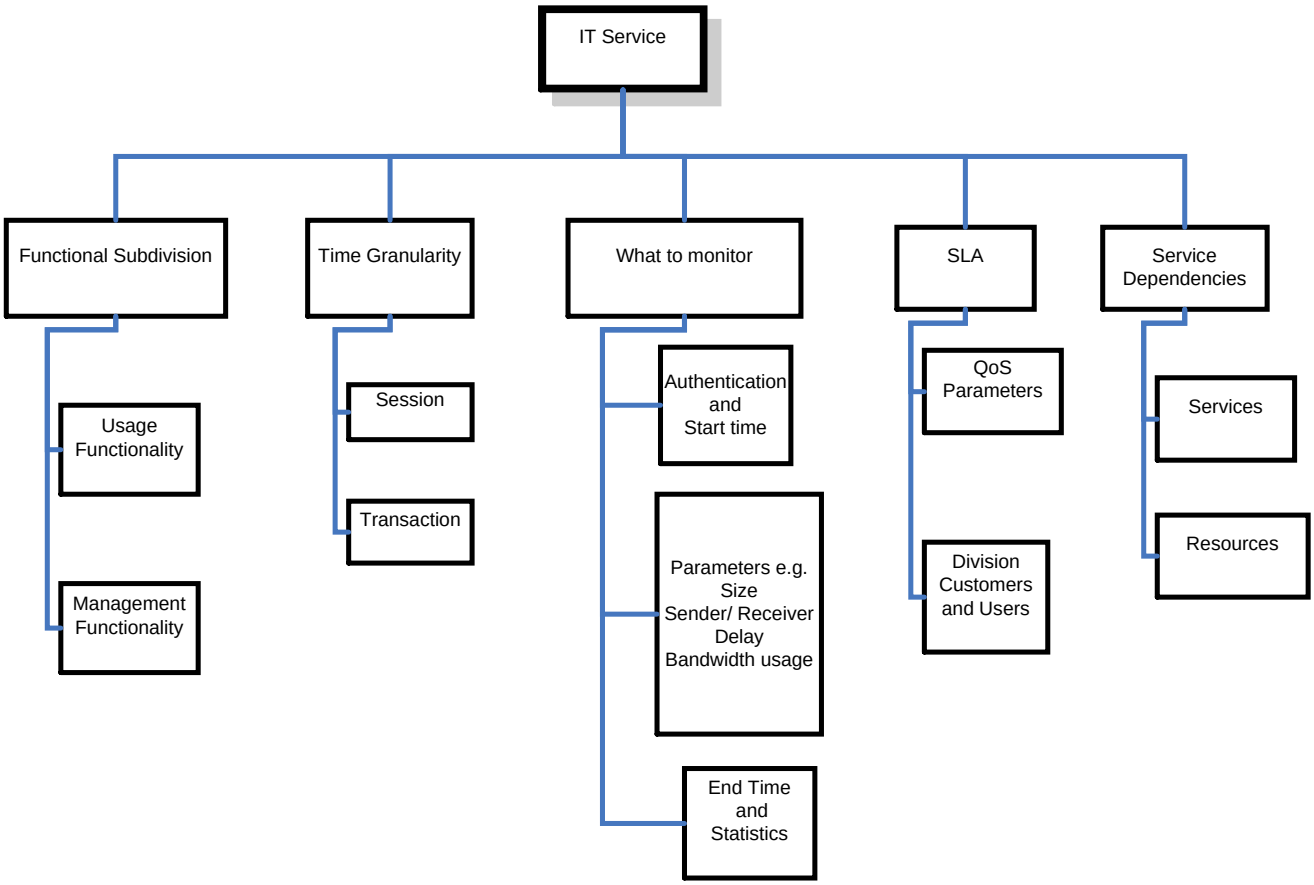


Figure 4.1: Modelling an IT service

These aspects are:

1. **Functional Subdivision:** When dealing with complexity, the rule is that it is divided into smaller and more manageable units and here it is exactly that what we want to do. When a service has a complex functionality dividing it into subfunctions according to the type of interaction is the way to get a closer look into the functionality of this service.

   Figure 4.2 shows the next stage in the making of the model. Here both sides of the functionality of a service according to the MNM service model are shown and how these two can be further divided into different areas is also easy to see.

   A first subdivision is as it was seen in Section 2.1:

   (a) usage functionality: interactions that represent the actual purpose of the service.

   (b) management functionality: interactions beyond the service's purpose that are needed to fulfil the customer's needs.

   The modelling of these functionalities to the benefit of the service provider helps find on time the information that is needed to react quickly and efficiently to service usage degradations.

2. **Time Granularity:** Any time granularity can be viewed as the partitioning of a time domain in groups of elements, where each group is perceived as an indivisible unit (a granule). The description

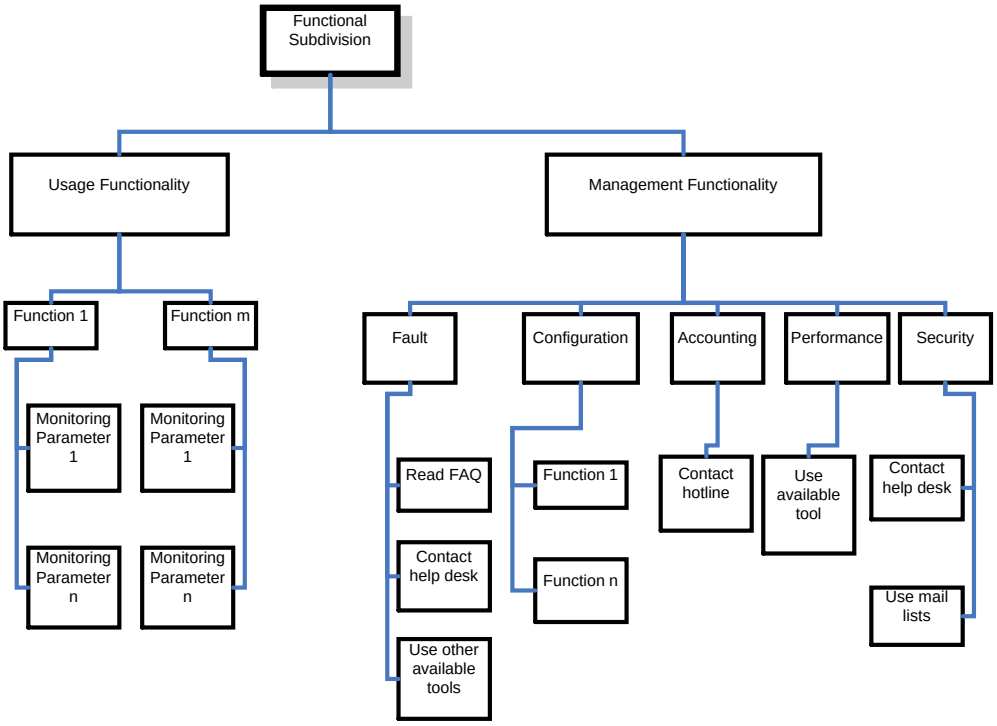


Figure 4.2: Modelling the functional subdivision

of a fact, action, or event can use these granules to provide them with a time qualification, at the appropriate abstraction level.

Applications may involve processes and actors spread over different sites, even in different parts of the world. It is therefore essential to have a common representation for time granularities between provider and customer so that they can correctly understand each other when referring to service usage.

Granularity of time is, in other words, the way in which a unit of time can be considered, i.e. a second, a minute, a day, a year, etc. and needs to be defined. Since we cannot expect a provider to monitor and record every second of their customer's usage of services, it is necessary to find a unit that leaves the time when a user is not using any service unmonitored and that focuses on specific aspects when the user is indeed using a service. That efficient unit could be:

- A Session:

  First of all it must be specified here what a usage session is from the provider's point of view.

  Figure 4.3 represents how a session can be divided into seven phases.

  (a) A user makes a request and this request is filtered by the service provider.

  (b) The service provider filter checks authentication of the service request using the supplied signature and gives a fault response in case of no authentication or a 'Authentication succeeded' message back to the service requester.

  (c) The user's request is sent to the meter service provider, requesting it to count the service and the meter service provider validates the service provider's request against the contract

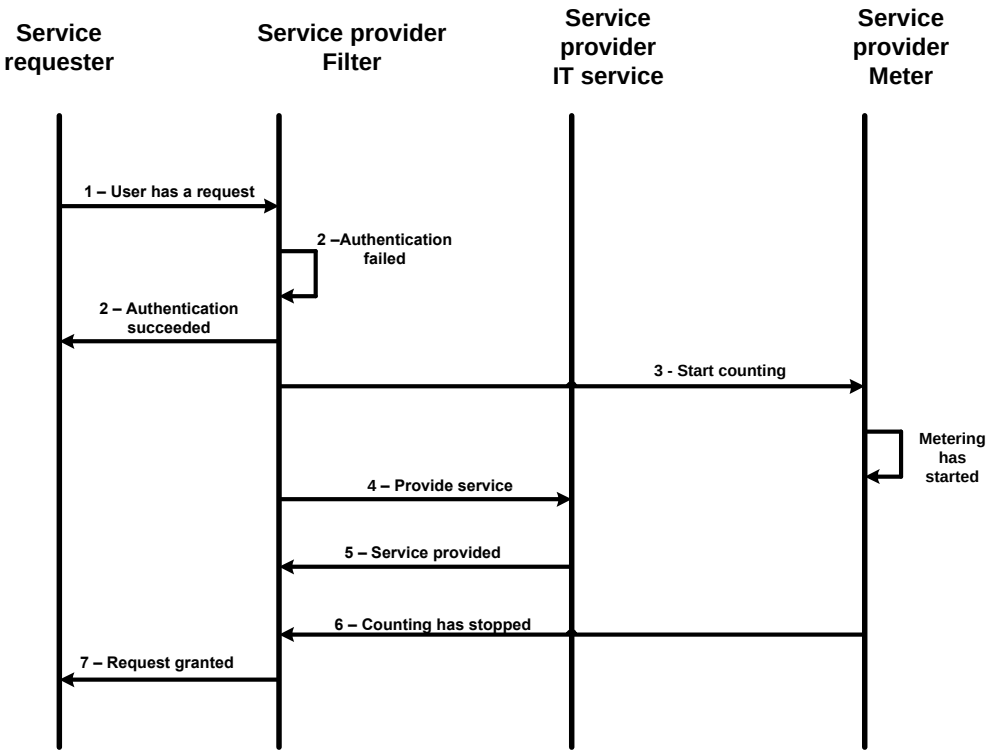


Figure 4.3: Monitoring of a usage session

details and starts counting. The usage session has begun and the information retrieved so far is user name and start time of session.

(d) The service provider executes the service. Here most of the monitoring takes place. For an e-mail service session, it is here when a service provider monitors the functionalities that a customer is accesing. If the user wanted to get his mail, it will be recorded when he intended to get his e-mail, when he actually got his mail, how long it took for the e-mail to be accessible to him, who the sender was, how big the e-mail(s) was(were), whether it contained spam, whether the mail was urgent, and so on.

(e) The service request is completed. At this point there is nothing else to meter but the time at which the execution of the service is finished.

(f) A request message is sent to the resource counter to stop counting. The service usage session is now finished.

(g) A response message is sent to the service requester indicating that the service usage session is completed.

Prior to the session's start the monitoring is concerned with the authentication process.

Once the user has been identified, the monitoring begins and the usage database starts being filled in with the raw data like session start and end time, bandwidth used for transfer data, etc.

At the end of the session the monitoring functionality is still not completely finished, for some data that needs to go into the usage database has to be processed to obtain some statistics. Examples that can be mentioned here are total number of sessions within the day, week, month, year, total bandwidth used, and basically all total figures.

When these totals are updated the monitoring for that user/customer and that session is completed. The service usage database is up to date and information is then available.

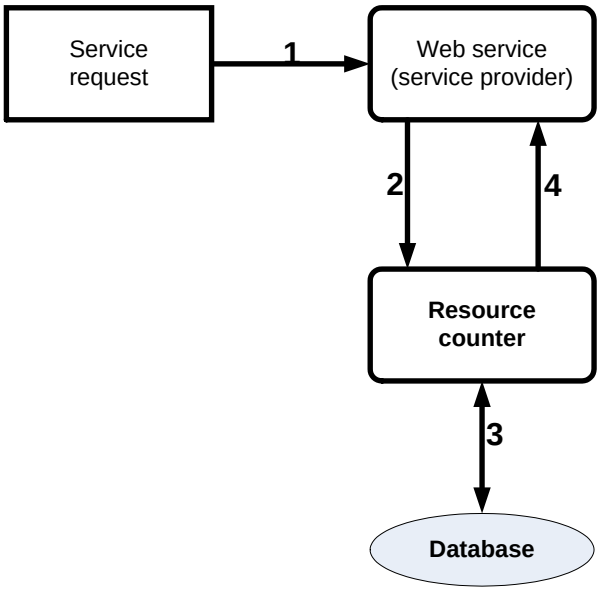


Figure 4.4: A simplified view of the monitoring of a usage session

Figure 4.4 summarises the above seven phases in four:

(a) request for a service.

(b) start counter as soon as the service is available to the customer.

(c) monitor the usage of the service and record it in a database.

(d) stop counter.

For simplicity's sake, minor interactions (that is, error situations and the certification authority needed to validate signatures) were omitted.

- A Transaction:

A transaction is a group of logical operations that must all succeed or fail as a group. As an example, consider a typical banking operation, moving 500 EU from a savings account to a current account. This seems like a single operation to the user but in fact consists of two: debiting the savings account by 500 EU and crediting the current account by 500 EU. If the debit operation succeeded and the credit did not, the 500 EU would disappear. Each of these two main operations may consist of several sub-operations, each of which may or may not fail; therefore we'll use these two 'main' operations (debiting and crediting) as a generic alias for any number of elementary operations in a generic transaction.

Transaction processing systems allow these two operations to be 'grouped' into a single transaction so these sorts of consistency problems cannot occur. They do this by making copies of the data in question and then running the operations on the copied data. When both commands have successfully completed, the changed data are written back to the system in a single operation. If either operation failed, the copied data are simply discarded, and an error is reported.

The requirements are that:

  - The database must not be left inconsistent because of a hardware or software failure in mid-transaction,
  - Other processes running while a transaction is in progress should have a consistent view of the database,
  - Completed transactions should be logged and re-run if necessary after a system failure.

3. **What to monitor:** Considering a service provider monitors sessions or transactions, this is what to monitor at what moment:
   - **Before the session/transaction has begun:** A service provider monitors here the identity of a user, it authenticates to give access to a service and, supposing the user gets access, the service provider keeps track of the start time of the session.
   - **During the session/transaction:** Here is when most of the monitoring takes place, so here we could have a long list of things we have to monitor (meter and record). A service provider monitors here different parameters depending on the service that is being monitoring.
   - **After the session/transaction:** As soon as the session is finished, a service provider needs to keep track of the end time; in this way it finds out how long the session lasted. When the session is over, the monitoring is concerned with restructuring the raw data to get some statistics that help provider and customer better understand what the real usage of the service was like.

   For examples we refer the reader to Sections 5.1 and 5.2.
4. **SLAs:** An important aspect to consider when monitoring the usage of services is what was negotiated in the SLA. Here it is specified what the quality of service is supposed to be like and what the penalties are for violation of this contract. Examples that could be applicable to every service are: delay and availability. A service provider offering for example 98% availability in an SLA means that they commit to the service being available 98% of time, but of course things don't go always the way one wishes, so a service provider must differentiate between negotiated usage and actual usage of services. That differentiation is only possible through monitoring.

   A service provider that has information about actual and negotiated usage is ready to inform its customers of its performance as a provider of services. Since they know that satisfied customers will maintain or even increase the level of usage of these services [BL99], it is a goal for a service provider to show customers good performance figures.
5. **Service dependencies:** Figure 4.5 shows the modelling of a service dependencies. Here dependencies between services and between services and resources are reflected.

   Resources can be considered strong if there is no alternative resource for them or weak for the case when for a resource there is another equivalent resource that in case of failure of the first one this alternative resource can be used; weak resources are in other words redundant resources.

4.1.1 Modelling the usage functionality

The usage functionality depicted in Figure 4.2 consists of a group of functionalities that are inherent to the service that is being monitored. The monitoring of each and all of these functions makes the monitoring of the service itself. The functions themselves could not be specified in this general model for they depend on the service that is being monitored.

A further subdivision for all functions into parameters to be monitored is necessary here. The kind and number of the parameters to be monitored depends on the functionality that is being monitored, so a list of parameters to be monitored here should suffice the purpose of the model.

For an example the reader is referred to Sections 5.1.1 and 5.2.1.

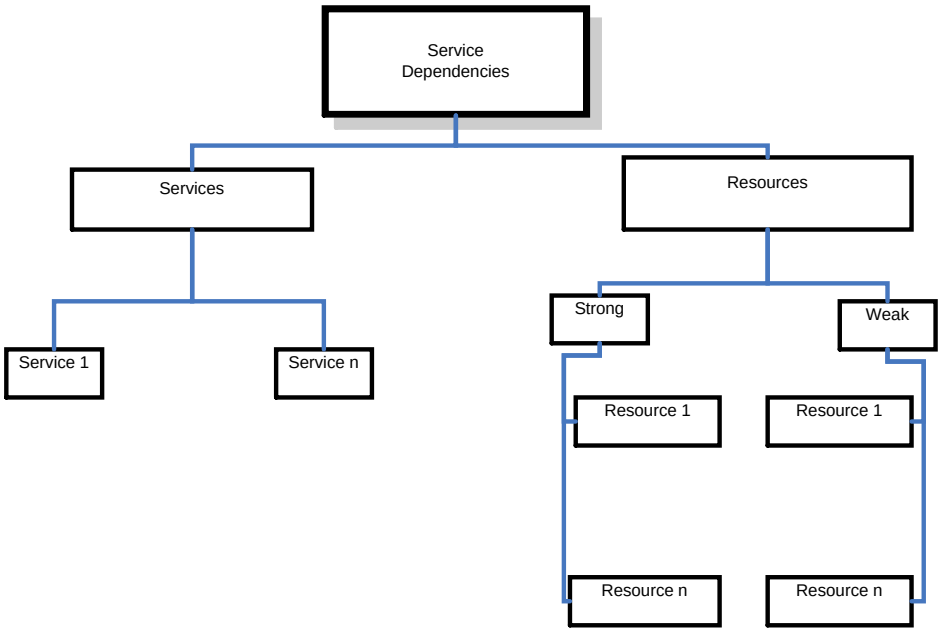


Figure 4.5: Modelling service dependencies

4.1.2 Modelling the management functionality

All those other interactions beyond the service's purpose to fulfil the customer's duties, to customize the service according to user's needs, to monitor and control the provider's service provisioning as well as for payment are here reflected.

In Figure 4.2 the management functionality is divided into five levels; each of those levels are here explained in more detail.

- **Fault**

  At the fault level, domain problems are found and corrected. Potential future problems are identified, and steps are taken to prevent them from occurring or recurring. In this way, the domain is kept operational, and downtime is minimized.

  Examples of management functionalities by fault are:

  1. **Read FAQs**

     A first approach to sort out a problem for a user is to read the FAQs. Service providers need a way to know how many people are actually reading FAQs. They need to know that in order to decide if it is worth offering FAQs to start with. It is also important to know whether people who are reading FAQs are getting the solution to the problems they have. To offer FAQs that do not help anyone is obviously undesired for both customers and providers and considered to be only a waste of time.

  2. **Contact help desk**

     A telephone line is usually offered to users that need some kind of help with the usage of services. At the help desk, the majority of problems get solved straightaway, but for the few that don't, a trouble ticket is created with the details of the request for help and this is sent to the appropriate department for revision and solution of the problem. When the line at the

help desk is not available or simply one prefers the usage of another resolution mechanism, an e-mail address is also offered in most cases. Here again monitoring how people are using a service is essential to find out whether everything is going well or whether some changes are needed. Providers need to know if people are using this service and if they are happy with the way the service is offered to them.

3. **Use other available tools**

   Apart from the help desk and e-mail contact address, service providers usually offer users some other kinds of tool to diagnose the user's problem. For the provider it is essential to fix the problem before it gets worse or has a bigger impact. For users it is evident they want to get out of trouble as soon as possible. Both users and customers want efficient tools that sort out the problems a user is encountering. A service provider needs to monitor how successful these tools are in order to optimize the quality of the service.

   A tool like the above mentioned should provide a service provider with the information it needs to act if necessary. A common way of getting this information from users is to ask if the problem was fixed with the information provided by the tool in question.

- **Configuration**

  At the configuration level, for example, network operation is monitored and controlled. Hardware and programming changes, including the addition of new equipment and programs, modification of existing systems, and removal of obsolete systems and programs, are coordinated. An inventory of equipment and programs is kept and updated regularly.

  Examples of management functionalities related to configuration are here difficult to mention since they are also inherent to each service. For some examples the reader is referred to Sections 5.1.2 and 5.2.2

- **Accounting**

  The accounting level involves tracking service usage and informing customers about the usage of resources and the cost associated with their usage. When computing resources are scarce, it may be necessary to set limits on the usage of resources. This level is also responsible for ensuring that customers are billed appropriately.

  A typical example of management functionality related to accounting is the view of bills.

- **Performance**

  The performance level is involved with managing the overall performance of the service. Throughput is maximized, bottlenecks are avoided, and potential problems are identified. A major part of the effort is to identify which improvements will yield the greatest overall performance enhancement.

  For an example of management functionality related to performance the reader is referred to Sections 5.1.2 and 5.2.2.

- **Security**

  At the security level, the network is protected against hackers, unauthorized users, and physical or electronic sabotage. Confidentiality of user information is maintained where necessary or warranted. The security systems also allow network administrators to control what each individual authorized user can (and cannot) do with the system.

  An example of management functionality related to security is an e-mail address that providers usually offer customers in case they have security issue concerns. By monitoring the amount of e-mails received with requests for help or advice and by monitoring the satisfaction of the customers in the way their request was handled, a service provider is able to learn whether it is worth offering such an e-mail service or whether some other alternative way of helping people with security issues should be found.

4.1.3 Prediction

A key aspect of any decision-making situation is being able to predict the circumstances that surround that decision and that situation. Such predictions, generally handled under the title of forecasting, have been identified as a key subpart of the decision-making process. As a natural consequence of the increased emphasis placed on systematic management, the area of predicting and forecasting has been studied extensively, and methods of making predictions more objective and reliable have been developed. These techniques vary considerably in their sophistication and usefulness.

No single forecasting method can meet the needs of all decision-making situations. Forecasting is merely a means of improving decision making and is not an end in itself. Decisions should be based on reliable forecasts.

Although each situation in which decisions must be made and in which a forecast might be helpful are different in nature, some elements are common to all. They are what make it possible to develop and use each method of forecasting for a number of different situations. The *first* element that will be noticed is that all these situations deal with the future and time is directly involved. Thus a forecast must be made for some specific point in time, and changing that point generally affects what the forecast will be. A *second* element that is always present in forecasting situations is uncertainty. If management were certain about what circumstances would exist at a given time, the preparation of a forecast would be a trivial matter. Virtually all situations faced by management involve uncertainty, however, and judgments must be made and information must be gathered on which to base a forecast. The *third* element, present in varying degrees in all the situations, is the reliance of a forecast on information that is contained in historical data. The amount of information contained in such data is a measure of how relevant that data is to decision making. Thus we could have a tremendous amount of data without having much information on what will happen in the future and vice versa. Generally speaking, forecasts are based directly or indirectly on information that is obtained from historical data. Although other elements may be present in a number of forecasting situations, these three are the most important.

This thesis is concerned with IT service usage and prediction and in this section the focus is placed on the prediction of IT service usage. The first question that has to be answered is what is to be predicted. This question has no standard answer that can be modelled and applied to each and all IT services. All that can be said about what to predict is that in order to predict we need historical data. This historical data originates from the service usage monitoring that was recorded in a database with the purpose of using this data to predict. The same usage functionalities that were monitored should be the subject of the forecasting. A service provider that has been monitoring the usage of services needs to keep record with all the details that were monitored. This record is stored in the service usage database and in order to predict what the usage of services will be like in the future this database has to be analysed.

There is a number of methods or techniques, that have been developed during the last two decades. These can be separated into two broad classes: quantitative techniques and qualitative techniques. This classification generally reflects the extent to which a forecast can be based directly on historical data in a technical fashion. Those techniques that start with past data values and then, following a certain set of rules, develop a prediction of future values fall into the category of quantitative methods. Situations in which such data is not readily available or applicable and in which much more management judgment must be inserted are generally best suited to the application of qualitative forecasting methods.

The area of quantitative forecasting methods is the most important for us and comprises a number of techniques whose common element is that the forecasts are based almost exclusively on historical data. Some of the more widely used techniques in this class include exponential smoothing, decomposition methods, and regression analysis. In these techniques data is used to help predict what will happen in some future time.

Quantitative forecasting techniques have gained wide acceptance over the last few decades for at least three reasons. One has been that they have developed a record of accuracy as a means of preparing forecast. A second important factor has been the development and adoption of computers. The computer can be used

not only to make the many computations that quantitative forecasting methods require but also to store historical data and then retrieve that data rapidly and efficiently when it is needed for the preparation of a new forecast. The last reason is that quantitative forecasts are, generally, much cheaper to obtain than any of the available alternatives.

Because of the difficulty (and cost) of working with qualitative methods of forecasting, they are generally applied only to long-term situations and to those of major importance. Qualitative forecasting methods are not yet well developed and are still largely intuitive and therefore they have only recently begun to gain wider acceptance.

The next question to answer is what forecasting technique should be used. This is another question that does not have a unique answer. The characteristics of the decision-making situation for which a forecast is to be prepared deserves special mention. The period of time over which a decision will have an impact and for which the manager must plan clearly affects the selection of the most appropriate forecasting method. One must be aware of the level of detail that will be required for the forecast to be useful in making decisions. The number of items to be forecast is another factor of importance. Apart from these considerations the decision maker must also consider the characteristics of the various forecasting methods in making his selection. It is often most effective to start with a simple forecasting method that does not require much data until the manager can build up a set of records that can then be used as the basis for applying a more sophisticated method. One decisive aspect is the consideration of the time allowed for preparing the forecast. The urgency in particular situations influences the selection of the method.

Central to any application of a forecasting technique is the role of data. Dealing with quantitative methods of forecasting it must be pointed out that each of these techniques requires that considerable amount of data be selected. This satisfies the requirement P1 of the catalogue of requirements. In order to comply with P2 a forecasting technique must be selected so that it can be applied.

Depending on how appropriate and accurate the available data is, the accuracy of the forecast will be determined. It is therefore essential that the necessary data be collected from a database that is regularly updated.

It has already been mentioned that the nature of some forecasting methods is associated with data requirements. The nature of data acquisition problems and the handling of data in a manner appropriate for forecasting has to be examined.

Most works on forecasting generally assume that the variable to be forecast is known and well defined. Although that is obviously true in situations in which a forecasting method is already being applied, in new situation it is not necessarily the case. The initial step in most new applications of forecasting is to determine the variable to be forecast that will be most useful to the manager and for which it is feasible to obtain historical information. Five aspects define the variable to be forecast:

- **the time span that should be covered by each value of the variable**

  Forecasts that generally contribute to longer range decision making can generally be based on observed data values for fairly long periods of time, such as quarterly or annually. Forecast aimed at controlling day-to-day operations would need to be based on data values that cover a time period of one day or perhaps even an hour.

- **the required level of detail**

  For one situation it may be satisfactory to forecast as a whole for a given period of time. In another situation the forecasting might be precised in more detail. It is always much more efficent to collect data at the most detailed level possible and then to aggregate it rather than collect aggregates and later dicover that they must be broken down into finer detail.

- **the frequency with which historical data is required**

  If data is used only on an annual basis, there is no need to have it collected within one or two days of its occurrence. If it its to be used on almost a daily basis, it must be collected much more rapidly.

- **the most appropriate units of measurement**

  Converting units of measurement represent a loss of information in most forecasting situations. The unit that is naturally associated with the variable should be used and if necessary a conversion of unit should be done after the raw data has been stored. This allows to go back to the raw data in its original unit.

- **the required level of accuracy**

  The factors that determine the most desirable level are the importance of the management situation and the role of the forecast in effecting that situation. When the forecast needed for an important management situation is peripherical to that situation, the level of accuracy required is not great. When the management situation is only of medium importancd but the forecasting represents the basis of decision making, a high degree of accuracy is desired.

The source from which data can be selected is the database. Data is collected on a number of different variables and then stored on some easily accessed system so that it will be available for selection when needed. Generally three types of data can be included in a data base for forecasting:

- data that is required and available
- data that is currently available but not currently required and
- data that may be required in the future but not currently available

In order for the requirement P1 to be fulfilled, it es necessary that the data is available. A database that focuses only on required and available data is generally the most straightforward and least expensive system to develop. With a small incremental cost such a base can be expanded to include the collection of available data that is not currently required but may be in the future. There are a number of sources of error in the data collection process, and thus it is necessary to run periodic checks on the data to make sure that such errors are not creeping in systematically and that the data sill represents what it is supposed to.

Matching the forecasting method with a particular situation is the most important step of forecasting. There are six criteria that can be used in selecting a forecasting method:

- The lead time for which the method is most appropriate (often referred to as the ‘time horizon’. This lead times can be divided into short (referring to a one- to three-month time horizon), medium (referring to three months to two years), and long term (referring to two years or more).
- The pattern of data that can be recognized and handled. This can be divided into horizontal (when the data are about evenly distributed over time, that is, when there is no apparent growth or decline over time) and trend (when there is a pattern of growth or decline in the data over the time span referred to).
- The type of model inherent in the method. It can be time-series (when time is used as an independent variable), casual (when other independent variables can be used in preparing a forecast), statistical (provide managers with a single point forecast and also supply the information needed to develop a confidence interval or range of values around that point forecast) and nonstatistical (output only a single value and do not provide the information necessary to test its significance).
- The cost associated with using that method.
- The accuracy of the method.
- The applicability of the method can be defined as the time required to obtain the forecast and the easiness to understand the results.

Once a method has been chosen it must be adapted to the situation. The specific forecasting techniques to which these criteria will be applied are those that have already been covered in this thesis in some detail in Section 3.3.

		Smoothing		Regression
Factors		**SMA**	**EMA**	**Regression**
Time horizon	Short term	✓	✓	
	Medium term	✓	✓	
	Long term			✓
Pattern	Horizontal	✓		
	Trend		✓	✓
Type of model	Time-series	✓	✓	✓
	Causal			✓
	Statistical			✓
	Nonstatistical	✓	✓	
Cost	0 smallest 10 highest	1	1	4
Accuracy	0 smallest 10 highest	2	3	5
Applicability	Time required to obtain forecast	1	1	3
	Easiness to understand the results	10	7	9

Table 4.1: A comparison of forecasting techniques on six basic criteria

Table 4.1 summarizes this comparison. The purpose of this table is to serve as a guide to help the manager in his selection process when he is faced with a particular situation that requires a forecast. A few brief comments should help the reader to understand this table. For the first three criteria the symbol ' ✓' has been used to indicate those techniques that are suitable for that particular criterion. For the last three criteria used in evaluating forecasting methods a point scale of 0 to 10 has been used to evaluate the various techniques, where 0 means smallest and 10 highest.

To finish this section the stages involved in the process of forecasting are summarised:

- **Phase 1:** Understanding the alternative forecasting techniques
- **Phase 2:** Selecting the forecasting method
- **Phase 3:** Selecting the historical data with a certain granularity
- **Phase 4:** Tuning the selected forecasting technique

4.2 Comparison of today's methodology with the use of the model

In the introduction of this thesis in Chapter 1 some deficiencies of today's methodology to monitor the usage of IT services were mentioned and, now that a new methodology has been proposed, a comparison of these two methods can be made and conclusions can be drawn.

The main problem with the way monitoring is done today is the lack of a structure that gives the subtasks involved in the process of monitoring the role they have in relation to the global structure. This lack of inner structure makes the monitoring superficial and as global as it can possibly be. Unfortunately this means that when unexpected events occur a patch has to be quickly put in place. It may well be that the patch works, but this should not disguise the fact that it is hardly ever an efficient way of proceeding. The result of this way of dealing with problems is that some subtasks are neglected in favor of others that apparently are more important or urgent. The reality is that service providers often find that, had they had more time to fix the problem, it would have been done in a more efficient way. Sometimes the fixing of problems has led to the occurrence of other problems. Other times the problems have reemerged and another solution was tried. In the majority of cases the economic loss is the most significant of all drawbacks. If service providers had more time to react appropriately to the different events they are exposed to, it would be a solution, but unfortunately time cannot be controlled and events do take place when they do, so the idea of

this thesis was to present a model that gives the whole monitoring of IT service usage an inner structure. A service provider has to deal with problems more efficiently; if he were to discover that a model that covers the full task of monitoring had become available, the service provider would profit considerably from it. With the use of a model, subtasks are automated and are therefore less likely to be forgotten or dealt with too late. The model presented here not only ensures that every part of the problem is treated but also at the right time which makes the process of restoring service functionality a lot faster.

Taking into account the proposed requirements, it was seen in Chapter 3 that these were not fulfilled in the present. After having developed the proposed model it is time to check whether the requirements are fulfilled. Section 4.1.1 and Section 4.1.2 cover the fulfillment of the requirements related to the service view; i.e. functional granularity (U1), time granularity (U2) and what to monitor (U3) are reflected on the proposed general model for IT service usage. Requirement U4 makes reference to the distinction between actual and negotiated usage and this is also fulfilled by the model. The IT service usage model is divided into 5 areas and while the actual service usage is modelled under 'What to monitor', the negotiated usage is modelled under 'SLA'. As it can be seen, the actual and the negotiated service usage are well detached in the model. The requirements related to general management are also covered by the model, since service dependencies and resources are also shown as another area of the model. Finally the requirements related to prediction were covered in Section 4.1.3. The selection of relevant data (P1) and the selection of a forecasting method (P2) are absolutely necessary to make a prediction. The conclusion is that the proposed model for IT service usage and prediction complies with all the requirements.

4.3 Summary

In this chapter a IT service usage model was proposed and explained in detail. This model complies with all the requirements that were proposed in Section 2.3. A comparison of today's methodology with the use of the model was made where it was shown how the deficiencies of today's monitoring of service usage vanish with the deployment of the proposed model.

Chapter 5

Applying the model to IT services

In this chapter it will be shown how the model that was developed in Chapter 4 can be applied to different services. This will hopefully help the reader better understand the general model and see its usability. The reader is invited here to jump back and forth to contrast some points in the general and in the concrete case.

The chapter begins with the application of the model to the e-mail service in Section 5.1 and then another example of application of the model to the web hosting service is presented in Section 5.2. The chapter closes with a summary in Section 5.3.

5.1 Applying the model to the e-mail service

Following the structure of the general model, those five aspects that were mentioned in the general model will now be here illustrated. Figure 5.1 represents an instantiation of the general model.

1. **Functional Subdivision:**

   The divide and conquer procedure for functions that make up a service is also applied here. The e-mail service has a complex functionality and, by dividing it into subfunctions, the reader gets a closer look into the functionality of this service.

   Figure 5.2 shows the next stage in the application of the model. Both sides of the functionality of the e-mail service are shown and how these two can be further divided into different areas.

   A first subdivision is as before:

   - usage functionality
   - management functionality

   More details about these functionalities are given in Sections 5.1.1 and 5.1.2.

2. **Time Granularity:**

   When considering the e-mail service, time granularity refers to the time unit that will be used to monitor. Every time a user tries to use a function of the e-mail service, a session begins. Once the session has started, the provider starts monitoring the usage and, when the session has finished, the provider produces some statistics with the raw data obtained from the measuring. To take the transaction as a granule or unit of time is not applicable in this example.

3. **What to monitor:**

   What exactly will be monitored is something that needs to be specified. Taking a session as a unit, this is what will be monitored at every moment:

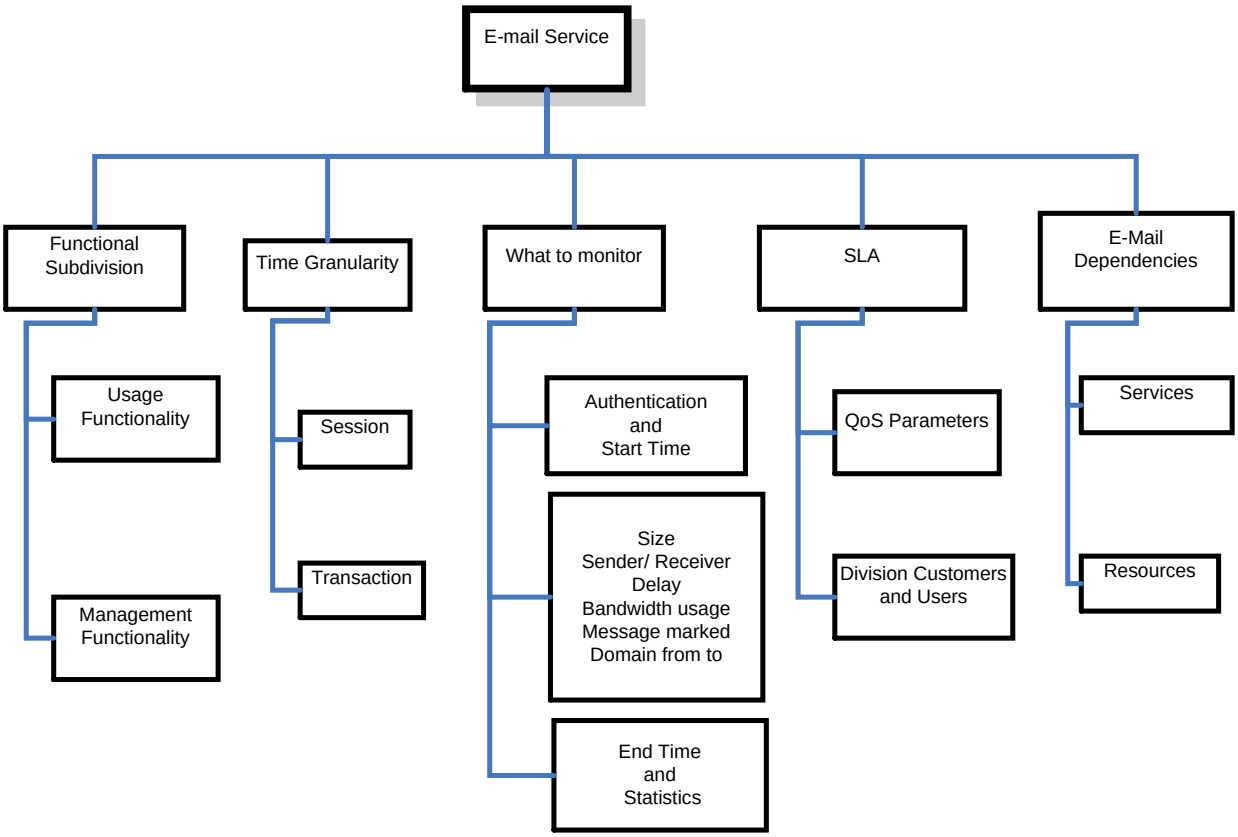


Figure 5.1: Modelling the E-mail service

- **Before the session has begun:** Here the user is identified, authenticated and given access to a service (supposing the authentication process succeeds). The starting time of the session is recorded.
- **During the session:** Here is when most of the monitoring takes place. Depending on what was specified the monitoring will take place. Suitable examples for the e-mail service are: Size of message in Kb, who the sender and the receiver are, what the delay in receiving the message is, the bandwidth that is used, whether the message is in someway marked (for example, as urgent), from what domain it comes from into what domain, whether the sender asks for a receipt.
- **After the session:** As soon as the session is finished, the end time is recorded so that the length of the session can be calculated. After that, the monitoring is concerned with restructuring the raw data to get some statistics that help provider and customer to better understand what the current service usage is like.

4. **SLAs:**

   An important aspect to consider when monitoring the service usage is what is negotiated in the SLA. A few examples are: delay, availability, maximum size per mail, penalties. A service provider must always have in mind what the negotiated service usage is and ensure that it can be realized.

5. **E-Mail dependencies:**

   Figure 5.3 illustrates the E-Mail service dependencies. Here dependencies between services and

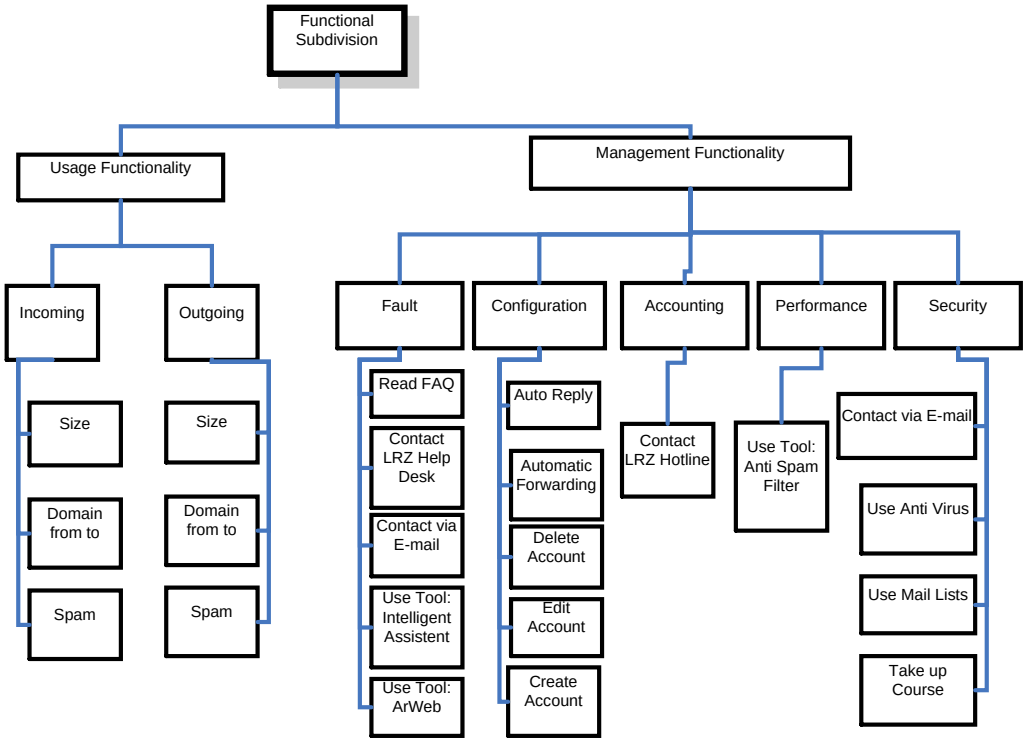


Figure 5.2: Modelling the functional subdivision of the E-mail service

between services and resources are reflected.

The E-Mail service depends on the proxy, IP, DNS and storage services. A failure in any of those services would imply that the E-Mail service is failing.

As the figure shows, there are two redundant E-Mail Relay servers at the LRZ, so if there is something wrong with one of them, the E-Mail service can still be functioning using the other server.

5.1.1 Modelling the E-mail usage functionality

All the functions that the service comprises should be considered, but for simplicity's sake the focus will be placed on two usage functions:

- **Receive messages or update incoming mailbox:** A service provider will monitor the incoming mailbox to obtain information about the current state of the service usage. A service provider can use this information later on for different purposes. If he is to predict something related to the service, he will certainly need this information at a certain level of granularity, so he should ensure that the information gained through the monitoring is kept properly.
- **Send messages or update outgoing mailbox:** The same as above applies here to this other functionality.

And here a few examples on what to monitor in these functionalities will be given:

- **Size:** Controlling the size of messages can help a service provider detect an error that might have been caused by too large a message. Some other uses of monitoring are checking SLA's conformance and optimization of the quality of a service to fit a user's needs.

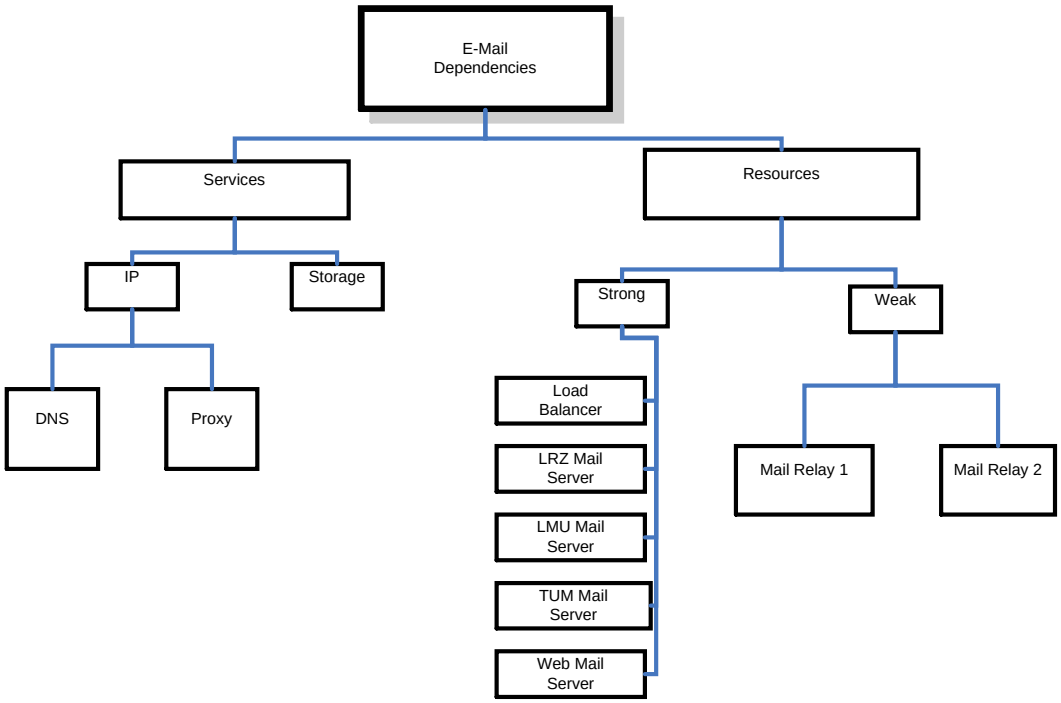


Figure 5.3: Modelling E-Mail dependencies

- **Spam:** This is also important because users who are flooded with spam are not getting the kind of standard that they want and that could cause a user to drop his/her provider for a better one.
- **Traversed distance:** This is also a feature that will surely help a provider detect and deal with failures more quickly. The longer the distance traversed by the e-mail for its delivery, the higher the chances are that something might go wrong.

5.1.2 Modelling of the management functionality

As it was mentioned earlier, the management functionality refers to all those other interactions beyond the service's purpose to fulfill the customer's needs. It is through the management functionality that a customer has the possibility to establish contact with the service provider to report issues concerning an IT service.

In Figure 5.2 was illustrated how the management functionality is divided into five levels and here examples of each of those levels are given.

- **Fault**

  Examples of management functionalities by fault at the LRZ are:

  1. **Read FAQs**

     As it was mentioned in Chapter 4 when the general model for IT service usage was developed, FAQs represent a first approach to sort out a problem. The reader is reminded here that the purpose of monitoring this functionality for a service provider is to find out whether people who are reading FAQs are getting the solution to the problems they have.

2. **Contact LRZ help desk**

   A telephone line is offered to LRZ's users that need some kind of help with the usage of services. The LRZ help desk solves most problems straightaway. Otherwise a trouble ticket is created and sent to the appropriate department for revision and solution of the problem.

3. **Contact the hotline per e-mail**

4. **Use tool: Intelligent assistant**

   This is a tool offered to users to diagnose the problem of a user. The user identifies himself and answers a few questions by clicking one of the alternative answers. These answers help the expert diagnose the problem and fix it. The user gets a trouble ticket number that he can use at the help desk to find out more details about the state of his query. When the problem is fixed, the user gets a message that informs him that the service is back to its full functionality.

5. **Use tool: ArWeb**

   When the intelligent assistant does not help to describe the problem or when someone has a question to ask, ArWeb is another tool that can be used. Here one fills in a form describing the problem or/and asking a question. A trouble ticket is generated and the problem is usually sorted out within 24 hours. After the generation of the trouble ticket, the user gets a trouble ticket number so that he can contact the help desk to inform himself about the state of the resolution of his problem.

For all these examples that have just been given, it is important to point out that monitoring how people are accessing these functionalities is essential for a provider to find out whether everything is going well or some changes should come into play.

- **Configuration**

  Examples of management functionalities related to configuration at the LRZ are:

  1. **Create a new account:** When the LRZ monitors the creation of new accounts it is being ensured that customers have access to their service usage on the one hand and on the other hand they can observe how many customers are accessing the e-mail service. This can be used, for example, to assess if their resources are coping well with demand or more resources are needed.

  2. **Edit an account:** Customers editing an account are concerned with something related to the e-mail service and that means that either they want to see what they are getting, maybe because they are reviewing price vs. quality, or they are altering their account because their needs have changed. Anything that concerns a customer is or should be important to the service provider.

     By editing an account it must be considered:

     – View parameters: it was mentioned that a customer viewing parameters of his or her account could mean that the customer is checking price vs. quality. This implies that he might be thinking about changing his service provider. This means for the latter that he might be losing a customer.

     – Modify parameters: a customer with new needs has to let his/her service provider know about it and the wish of a service provider is to satisfy a customer with his service, so monitoring the usage of this function is important to make sure that a service provider is still active in the market.

     Furthermore when a customer is modifying some parameters like the user's name, the user's e-mail address or the size of his mailbox, a customer is showing that the current usage of this service does not longer satisfies his/her needs and finding out the reason why this is so is very important for the service provider. It might not be due to anything important like a user's name

no longer being desired, but it might also be due to the fact that for example, the user is not happy with his mailbox size.

3. **Delete an account:** Keeping track of the amount and kind of accounts that are active is a must for the service provider to assess its resources.
4. **Automatic forward of incoming mails:** This functionality needs to be monitored to check it is working properly.
5. **Auto reply:** Not all customers will be using this function, but a service provider should know exactly who is and who is not, so that for example, in case of error with this functionality, the affected customers can be easily identified.

- **Accounting**

  An example of management functionality related to accounting at the LRZ is an accounting hotline available for users. View a bill is another example.

- **Performance**

  An example of management functionality related to performance at the LRZ is the anti-spam filter 'SpamAssassin'. This mail filter scans and attempts to block spam. Each incoming e-mail is scanned for signs that it may be 'spam', and if it is determined that it's likely to be spam, it is altered to clearly reflect this, so that you can decide whether to delete or keep it. Supervising the performance of a service is essential for a service provider that wants to optimize it.

- **Security**

  Examples of management functionalities related to security at the LRZ are:

  1. **Courses and Information**

     The LRZ offers several courses and information about security issues for every concerned user. People taking part in courses or asking for information about security issues need to be monitored. If the LRZ knows whether people are being helped by this means, how many people are using this services and how satisfied they are with them, they can then decide whether a course is updated or scrapped.

  2. **E-mail address**

     An e-mail can be sent to security@lrz.de to address issues related to security. By monitoring the amount of e-mails with requests for help and following their satisfaction with the way their request was handled with, a service provider can find out whether it is worth going on with way of helping with security matters or something else should be done to improve the quality of this service.

  3. **Anti-virus software**

     The anti-virus software from Sophos is available to users. The LRZ needs to know how many of their users are actually using this software. It is also important for a service provider to know whether the software is user friendly, for this will give the users the satisfaction of being in control of their systems.

  4. **Mailing list**

     In order to keep tabs on security issues, the LRZ offers its users the use of mailing lists which a user can subscribe to find out information related to security.

     Majordomo is a perl script which automates the management of Internet mailing lists. It is executed via electronic mail; users send e-mail to Majordomo with instructions in the body of the message, and the perl script performs the requested actions and responds with the results. Any text in the 'Subject:' line is ignored.

Here we see where to find information about some important security issues:

Information about warnings: win-sec-ssc@cert.dfn.de

Discussions around ‘Security’: win-sec@cert.dfn.de

Discussions around ‘PKI’ etc.: win-pca@pca.dfn.de

News about ‘Security’: security-news@lists.lrz.de

Information and discussions about ‘Security in MWN’: mwn-security@lists.lrz.de

The LRZ as a service provider needs to have control over all these mailing lists by monitoring what is happening with them. Once more, the satisfaction of the user is a decisive aspect for a service provider to react to daily demands of users and customers.

5.1.3 E-mail prediction

This section is concerned with predicting future e-mail usage functionalities that were monitored and appropriately stored in a database.

A service provider using the proposed model can determine the current and future usage of this service and by doing so he is gaining the information he needs to react adequately to different events. For example, by forecasting how much e-mail a customer will be receiving in a month’s time according to his/her usage history database, he can find out if the quality of this service provisioning is endangered in any way. According to the manager’s knowledge about forecasting methods and given the low importance of the concrete forecasting situation, the manager checks up his table (illustrated in Figure 4.1) for selecting the most appropriate forecasting technique. Having considered all the six criteria he decides to apply the simple moving average technique. He defines exactly what he wants to forecast as the number of e-mails that customer X will be receiving in the month of april, which happens to be the following month. He proceeds to the collection of historical data. For the selection of historical data he specifies that the amount of e-mails received in the last 3 months will be needed on a monthly basis. He accesses the database and retrieves these values. He then applies the simple moving average technique and interpret this result.

Another example of application of prediction concepts on the e-mail service could present a customer that sells books and orders are to be made by e-mail. Service provider and customer sign the pertinent SLA with the details about the service provisioning. The customer specifies that he/she needs the service from Monday to Friday and from 7AM to 3PM and this is, of course, reflected in the SLA. Violations of the SLA will incur into penalties. By knowing the current and future usage of the E-mail service a service provider has the information he needs to make a decision when, for example, a service degradation has occurred. In order to restore the full functionality of the E-mail service the service provider must project the consequences of applying different recovery alternatives. It is by predicting what it will be like in the future that a service provider is in the position of making ‘the right decision’ about what needs fixing and when.

5.2 Applying the model to the web hosting service

Here the general model will be applied to the web hosting service. The five aspects shown in the general model of service usage will be explained in the context of the web hosting service. Figure 5.4 represents an instantiation of the general model.

1. **Functional Subdivision:**

   The web hosting service is a very complex one as it has lots of functionalities. Applying the general divide and conquer method, the full functionality of the web hosting service can be addressed. For simplicity’s sake, just a few of those functionalities will be considered.

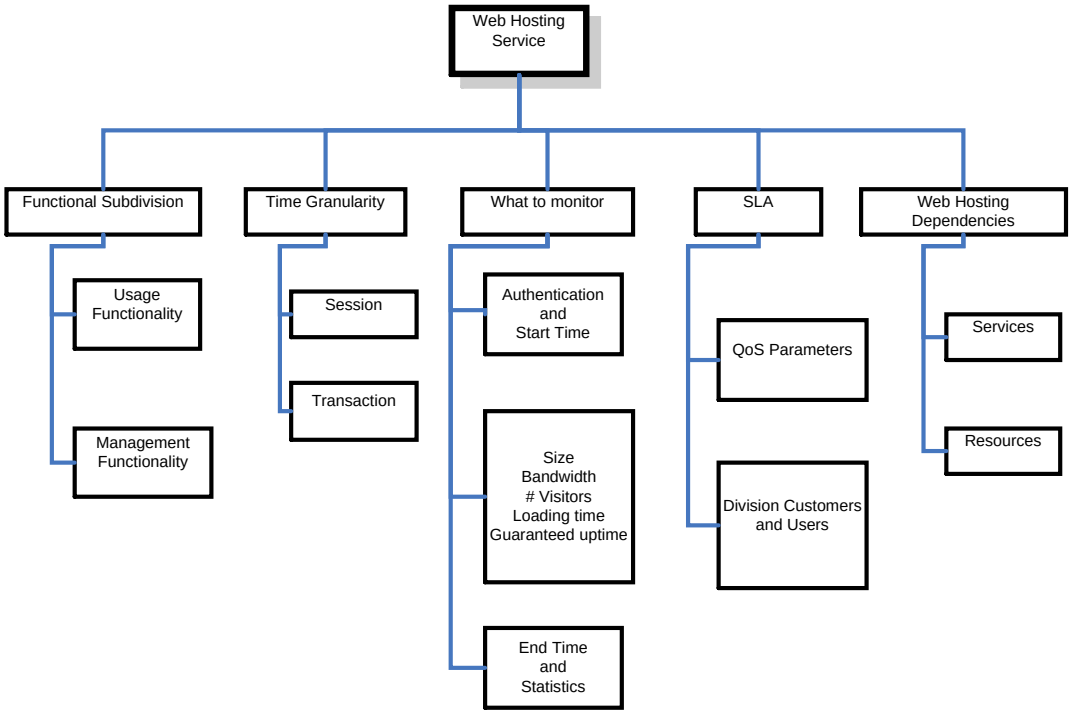


Figure 5.4: Modelling web hosting usage

Figure 5.5 illustrates the next stage into the application of the general model. Here the functionalities and subfunctionalities of the web hosting service are shown. A first hierarchy level refers to usage and management functionalities. For an explanation of a further hierarchy level the reader is referred to Sections 5.2.1 and 5.2.2.

2. **Time Granularity:**

   Every time a user tries to access a functionality of the web hosting service a session begins. Once the session has started the provider starts monitoring the usage and when the session has finished the provider produces some statistics with the raw data obtained from the measuring. To take the transaction as a granule or unit of time is applicable in this example. The example on page 28 about bank transfers could be used here to explain why it is important to specify a time granularity. Once a transfer has started, a service provider must go on monitoring until the transaction has finished. Imagine what would happen if, in the middle of a transaction, the service provider stops monitoring for whatever reason it might have. By loosing track of the development of the transaction, the service provider has no accurate information about the current web hosting usage. Lack of accuracy in data could be disastrous for, for example, a forecast that uses the inaccurate data.

3. **What to monitor:**

   Here what to monitor at what time will be examined:

   - **Before the session has begun:** The user will be identified and authenticated. This information will be stored together with the starting time of the session.
   - **During the session:** Examples of things that could be monitored for the web hosting service are: Size of a website, loading time of a website, the bandwidth used, whether the website is altered, how many times it has been altered, percentage of uptime for a website, etc.

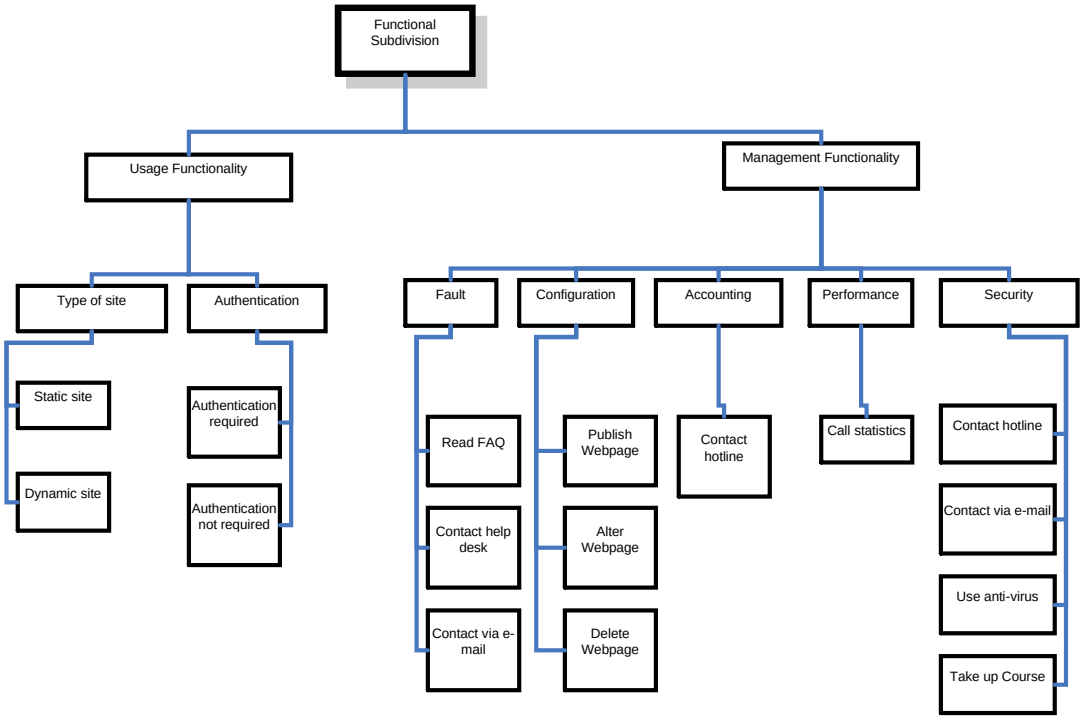


Figure 5.5: Modelling the functional subdivision

- **After the session:** As soon as the session has finished, the end time will be stored so that the length of the session can be calculated. When the session is over, the monitoring is concerned with restructuring the raw data to get some statistics that help provider and user to better understand what the current usage of the service is like.

4. **SLAs:**

   A few examples about what should be monitored related to SLAs are: maximum permited loading time, availability, maximum size per website and penalties.

5. **Web hosting dependencies:**

   Figure 5.6 illustrates the web hosting dependencies. Here dependencies between services and between services and resources are shown. The web hosting service depends on the well functioning of the proxy, IP, DNS and storage services. The LRZ has five redundant servers for the web hosting service and also four redundant servers hosting their own LRZ pages.

In the next two sections, the modelling of the web hosting service functionality will be examined.

5.2.1 Modelling the usage functionality

Figure 5.4 illustrates the modelling of the functional subdivision and here in this section the usage functionalities are shown.

- **Type of website:** Depending on the type of contents of a website, it can be divided into:
  - Static websites' contents remain so until they are altered or deleted. The content was written directly by an author, and when the user goes to the site, that code is downloaded into a browser

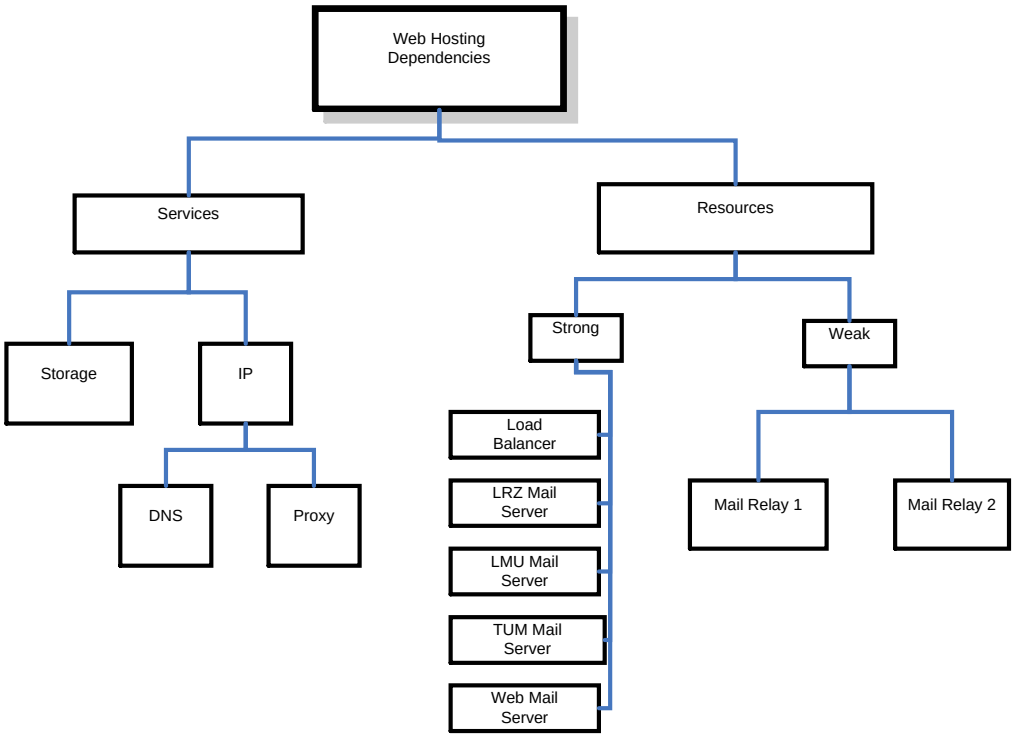


Figure 5.6: Modelling web hosting dependencies

and interpreted. At the LRZ static websites are located in an Andrew File System which is a location-independent file system that uses a local cache to reduce the workload and increase the performance of a distributed computing environment. A first request for data to a server from a workstation is satisfied by the server and placed in a local cache. A second request for the same data is satisfied from the local cache

– A dynamic website on the other hand uses programming in addition to the layout to not only allow the flow of data in and out of the site but to make meaningful liaisons with the data. For example an online business directory may allow local businesses to sign up and create profiles for their businesses. Visitors to the site may then search the directory based on their needs. For example, one may conduct a search for a lawyer in their postcode who speaks Spanish or for a dance school in the area that offers tap dancing classes. This type of websites are called dynamic and they are located in a Network File System at the LRZ. The NFS is a distributed file system which allows a computer to access files over a network as easily as if they were on its local disks.

Modelling the type of websites helps a service provider have a more accurate idea of what resources are affected in case of service degradations.

- **Authentication:** Whether a website needs authentication or not typically means whether the user must log in (enter user ID and password) first in order to access the desired website. Many websites use authentication to restrict the access of a page to a specific set of users. Modelling this usage functionality helps a service provider to detect more easily where the problems might be.

Examples on what to monitor follow here:

- **Size:** Controlling the size of a website is important for a service provider because it helps detect

other degradations that might occur that are related to the fact that a website is too large like, for example, the loading time of a website being too long due to the fact that it contains lots of sound files. As access to the server becomes more difficult, the number of packets lost increases. For small files, such as web pages and images, this is not such a problem; packet loss is (statistically) less likely to occur on small files. On the other hand, larger files are more likely to loose data during transfer since there are more opportunities for packets to become lost. These lost packets can adversely affect the quality of the download. Files can become corrupted; streaming media gets more jittery the more data is lost. A provider takes all this into consideration and might want to suggest to the website owner the use of links to each different media file so that the page containing the links can be loaded faster.

- **Publicity content:** This is also important because too much publicity on a website is in general annoying. People looking for something in a website do not want to be distracted with publicity. Publicity content could cause a user to drop his/her provider for a better one.
- **Number of people altering a website:** This is a feature that will surely help a provider detect and deal with failures more quickly.
- **Loading time of a website:** Both user and provider want a website to load quickly. If 500 web surfers all ask for the same page at the same time, the server can get bogged down, and loading speed will slow down. The home page at Yahoo gets 4 million hits a day. A personal web page could get only 4 hits per month. Web hosting providers will tune their servers to apply more computing power to the sites that get the most hits.
- **Number of visits to a website:** A website which monitoring has shown to be very popular needs to be kept clear from trouble, because the chances are that it will be visited a lot again in the future. A service provider that has the knowledge about what websites are more active than others can decide quickly which websites encountering problems should have preference when a fault needs to be remedied.
- **Number of times a website gets altered:** The fact that a website gets altered quite often, makes it more likely to be susceptible to errors. Keeping track of websites that get altered frequently can help a service provider find affected customers more quickly.

5.2.2 Modelling the management functionality

In Figure 5.5 we saw that the management functionality could be divided into five levels and as we did for the previous example, we now give examples for each level.

- **Fault**

  The same examples of management functionalities by fault at the LRZ that were reviewed in the previous example are here applicable:

  - Read FAQs
  - Contact LRZ help desk
  - Contact the hotline per e-mail

  For more details about what should be here monitored and why see page 40.

- **Configuration**

  Examples of management functionalities related to configuration at the LRZ are:

  - **Publish a website:** Monitoring this function allows the service provider assess its resources. If its resources are sufficient everything shall run smoothly, but if they run a bit short it is time to act before a service degradation occurs.

  - **Delete a website:** Following the previous point, we have now the reverse case: resources available to realize this function exceed the real needs.
  - **Alter the content of a website:** Do the changes in the website that a customer intends entail resource shortages? This is an example question that could be answered after having monitored this functionality.

- **Accounting**

  The LRZ hotline serves as an example of management functionality related to accounting. The LRZ provides a variety of mechanisms like online surveys for users to give feedback; such mechanisms are essential for a service provider to know how the service is running.

- **Performance**

  An example of management functionality related to performance at the LRZ is the invitation to use of a pop up blocker. With a pop up blocker, a visitor to a website can get to the content of the website faster than with annoying advertising. A service provider concerned with performance needs to address the issue of unwanted pop ups.

- **Security**

  Examples of management functionalities related to security at the LRZ are like in the previous example:

  - Courses and information
  - E-mail contact address
  - Anti-virus software
  - Mailing lists

  For more details the reader is referred back to page 41.

5.2.3 Web hosting prediction

In this section an example related to prediction for the web hosting service will be illustrated. It must be remarked that this example was specially chosen to show how all the concepts developed in this thesis fit in the impact analysis framework that was shown in page 11.

The role of the customer is taken by someone who runs his business through the web hosting service offered by a provider. The customer sells his products through the internet so he/she has signed a contract with the web hosting service provider in which it was determined what exactly he/she was going to have access to. The customer was worried that the service might not be available 100% of the time. The provider promised that the service would be available 98% of the time everyday of the week and 24 hours a day during a calendar month. The customer had the added requirement that a degradation in quality should not last longer than 1 hour. Given the fact that the customer needs something more than a guarantee better than a promise about the availability of the web hosting service he is signing for, he and the service provider come to the agreement that the service provider would pay compensation should the service fail either partially or totally. In the event of a degradation in the quality of the service, they proposed three periods of time as crucial:

1. the service is degraded for a period of time between 1 and 5 hours
2. between 5 hours and 24 hours
3. more than 24 hours

A degradation of service quality lasting less than 1 hour would be tolerated by the customer so long as the 98% availability of the overall service was upheld.

The customer knows exactly what his or her sales are likely to be within these periods of time and, in the case where the web hosting service is not fully working, the customer will lose these sales. He and the provider agree that any loss resulting from degradation of service will be made good by the provider in the form of a penalty, the penalty increasing as the period of service degradation lengthens which will have a knock-on effect on the customer's sales.

Up until this point, the focus was on the customer's needs and the SLA. Let us now have a look at the service provider's perspective and see what actually happens after this contract has been signed. As always happens in reality, not everything goes the way it has been planned. Imagine the following scenario: one day a problem with the DNS server causes the web hosting service to be unavailable for 3 hours. The web hosting service has been running well up until this date and has been meeting the 98% availability criterium offered by the provider, but now the part of the contract concerning the length of time of the degradation has been violated. If the provider has been able to fix the problem within an hour, there would have been no consequences for him. But, since the problem with the DNS server has been ongoing for 3 hours, he will have to pay the penalty due. But with the problem with the DNS server still not being fixed, the service provider is faced with the possibility of having to pay yet a higher penalty in case the service has not been fully restored to full functionality within the first 5 hours.

It should not be forgotten that this customer will not be the provider's sole customer and so other customers will also be feeling the effects of the degradation of service quality. As a result the service provider will also need to keep their SLAs and their respective penalties in mind. Some of those customers might not even be using the service at all so they will not even notice the service malfunctioning and will consequently not be affected. The service provider would need to know what the service usage was like just before the degradation of service quality occurred and also what the usage is likely to be in the period of time during which the repairs are taking place so that he can assess who exactly has been affected and in what precise way. The service provider will have to analyse the situation with the conclusion being that the consequences of the application of different recovery alternatives must be forecast. Drawing on his knowledge about forecasting techniques and considering the characteristics of the values he wants to get, he decides to go for a particular forecasting technique. He specifies the granularity with which the historical data must be selected and, once he is in possession of this data, he applies the chosen method. The results of the forecasting reflecting the most convenient recovery alternative are passed over to the recovery management component shown in Figure 2.6 where this alternative will be selected for application.

5.3 Summary

The general model was applied to the e-mail and web hosting service and it was seen that there is a lot of features that need to be monitored. It was also mentioned why a service provider profits from the knowledge that monitoring these services' usage gives him. To summarize all the reasons a provider has to monitor these services it could be said, is to optimize the quality of services. Sometimes that means a service degradation is sorted out more efficiently and other times that a customer gets a more customized service to his purposes. Modelling the usage of IT services has proven to be of considerable help in finding deficiencies and improving the quality of the services offered.

Chapter 6

Summary and Conclusion

Chapter 1 was an important chapter. An introduction was presented as to what the reader could expect from each chapter and the thesis was put into a broader context so that the reader would be able to understand not only what is being done in this thesis but also why it is needed.

Chapter 2 was concerned with offering the reader a starting point for the development of the thesis. The MNM service model was chosen for this purpose and subsequently a catalogue of requirements that would need to be fulfilled by the desired model was proposed.

It was necessary to do some research for related work and analyse everything that was found to test it against the proposed requirements. That is Chapter 3. The assessment proved that not one of those pieces of work fulfilled all the requirements.

So Chapter 4 was needed to create the model that would fulfill all the requirements. When the model was ready, a comparison with the methodology used today seemed to be the logical way of proceeding to show the contributions of this thesis to the service usage monitoring.

The balance was positive so Chapter 5 was written to show how well the newly developed model could be applied to the e-mail and web hosting services.

Before the drafting of this thesis service providers often felt unsupported in tasks such as restoring services after occurrence of failures. They had to rely on the experience of their employees and that had several drawbacks. Now that a model of IT service usage is available through this thesis, service providers have at their disposal a tool that was badly needed to automate processes and that enables them to react efficiently and on time to different events.

It was mentioned that service usage was being monitored but it was often the case that the information the service provider gained from it came too late. It was clear that a better organization of time would lead to the right solution and nothing could be more appropriate to organise information than in a model. So that is what was done!

The only drawback would appear to be in the time needed to prepare such a model. As it was said, the model presented here is not a complete model even though it took a few months to be created. The aim of this project was to offer a concise model that gave a service provider an idea of how to model service usage and how to profit from it. Now it can confidently be said that a manager who has understood the importance of structuring and automating processes can look into the future and easily conclude that, although the modelling of IT service usage involves time and money, it will definitely be worthwhile. Thus, the creation of a full model that fits the characteristics of a particular service provider is a challenge, but looking to the future, the advantages that this modelling brings about are too good not to be taken into consideration.

Last words to this thesis must be given to what still needs to be addressed in future work in the field of IT service usage modelling. A deeper insight into the modelling of the functionalities together with a more

detailed modelling of services dependencies and resources used are areas to be covered in future work.

The final conclusion drawn from this thesis is that it represents a contribution to the optimization of IT service provisioning.

Bibliography

[(AG04] The Australian Government Information Management Office (AGIMO). Better practice checklists in online service delivery, 2004. http://www.agimo.gov.au/.

[AGI] Agimo checklist in website usage monitoring and evaluation. Website. http://www.agimo.gov.au/practice/delivery/checklists/evaluation.

[BL99] Ruth N. Bolton and Katherine N. Lemon. A dynamic model of customers' usage of services: Usage as an antecedent and consequence of satisfaction. *Journal of marketing research*, may 1999.

[Con97] TINA Consortium. Service architecture version 5.0. tina baseline. *TINA Consortium*, june 1997. http://www.tinac.com/specifications/specifications.htm.

[For02] Distributed Management Task Force. Common information model (cim) core policy model white paper. *Distributed Management Task Force*, march 2002. http://www.dmtf.org/standards/documents/CIM/DSP0108.pdf.

[For05] Telemanagement Forum. Enhanced telecom operations map (etom) the business process framework for the information and communications services industry - gb921 v6.0 r6.0. *Telemanagement Forum*, december 2005. http://www.tmforum.org/browse.asp?catID=1647.

[GHH+01] M. Garschhammer, R. Hauck, H.G. Hegering, B. Kempter, M. Langer, M. Nerb, I. Radisic, H. Roelle, and H. Schmidt. Towards generic service management concepts - a service model based approach. In *Proceedings of the 7th International IFIP/IEEE Symposium on Integrated Management (IM 2001), Seattle, Washington, USA*, may 2001. http://www.nm.informatik.uni-muenchen.de/Literatur/MNMPub/Publikationen/smtf01/smtf01.shtml.

[GHH+02] M. Garschhammer, R. Hauck, H.G. Hegering, B. Kempter, I. Radisic, H. Roelle, and H. Schmidt. A case-driven methodology for applying the mnm service model. In *Proceedings of the 8th International IFIP/IEEE Network Operations and Management Symposium (NOMS 2002), Florence, Italy*, april 2002. http://www.nm.informatik.uni-muenchen.de/Literatur/MNMPub/Publikationen/ghhk02/ghhk02.shtml.

[GHK+01] M. Garschhammer, R. Hauck, B. Kempter, I. Radisic, H. Roelle, and H. Schmidt. The mnm service model - refined views on generic service management. *Journal of Communications and Network*, 3(4), December 2001. http://www.nm.informatik.uni-muenchen.de/Literatur/MNMPub/Publikationen/ghkr01/ghkr01.shtml.

[HAN99] H.G: Hegering, S. Abeck, and B. Neumair. *Integrated Management of Networked Systems - Concepts, Architectures and their Operational Application*. Morgan Kaufmann Publishers, ISBN 1-55860-571-1, 1999.

[HR00] R. Hauck and I. Radisic. Monitoring application service performance - classification and analysis of existing approaches. In *Proceedings of the 7th International Workshop of the HP OpenView University Association (HPOVUA 2000), Santorini, Greece*, June 2000.

[HSS04a] A. Hanemann, M. Sailer, and D. Schmitz. Assured service quality by improved fault management. In *Proceedings of the 2nd International Conference on Service Oriented Computing (ICSOC04), 183 to 192, ACM Press, ACM SIGSOFT and SIGWEB, New York City, NY, USA*, nov 2004.

[HSS04b] A. Hanemann, M. Sailer, and D. Schmitz. Variety of qos - the mnm service model applied to web hosting services. In *In 11th International Workshop of the HP OpenView University Association (HPOVUA 2004), 2004, Paris, France*, june 2004.

[HSS05] Andreas Hanemann, David Schmitz, and Martin Sailer. A framework for failure impact analysis and recovery with respect to service level agreements. In *Proceedings of the IEEE International Conference on Services (SCC 2005), IEEE, Orlando, Florida, USA*, juli 2005. `http://www.nm.ifi.lmu.de/pub/Publikationen/hss05c/PDF-Version/hss05c.pdf`.

[Iil00] ITIL IT infrastructure library. *Service Support*. Stationery Office Books, Norwich, UK, june 2000. `http://www.itil-itsm-world.com/support.htm`.

[Iil01] ITIL IT infrastructure library. *Service Delivery*. Stationery Office Books, Norwich, UK, april 2001. `http://www.itil-itsm-world.com/delivery.htm`.

[ISO95] Information Technology ISO. Information technology - open systems interconnection - systems management - part 10: Usage metering function for accounting purposes is 10164-10. *ISO*, dec 1995.

[ITU] The international telecommunications union (itu). Website. `http://www.itu.int/ITU-T/`.

[LLN98] Michael Langer, Stefan Loidl, and Michael Nerb. Customer service management: A more transparent view to your subscribed services. *Technical Report*, 1998.

[LRZa] The e-mail service at the leibniz supercomputer center. Website. `http://www.lrz-muenchen.de/services/netzdienste/email/`.

[LRZb] The web hosting service at the leibniz supercomputer center. Website. `http://www.lrz-muenchen.de/services/netzdienste/www/v-server/`.

[MNM+00] Christian Mayerl, Z. Nochta, M. Müller, Martin Schauer, A. Uremovic, and Sebastian Abeck. Specification of a service management architecture to run distributed and networked systems. In *Proceedings of the Third International IFIP/GI Working Conference on Trends in Distributed Systems: Towards a Universal Service Market*, Sept 2000.

[MSH98] S. Makridakis, S.Wheelwright, and R. Hyndman. *Forecasting: Methods and Applications*. Wiley, 1998.

[Rod03] Gabi Dreo Rodosek. A generic model for it services and service management. In *Integrated Network Management*, pages 171–184, 2003. `http://dblp.uni-trier.de`.

[SB03] Mathias Salle and Claudio Bartolini. Management by contract. *Technical Report*, 2003.