University of Oslo
Department of Informatics

A Conceptual Model for Service Availability

Judith E. Y. Rossebø
Mass Soldal Lund
Knut-Eilif Husa
Atle Refsdal

Research Report 337
ISBN 82-7368-292-7
ISSN 0806-3036

3rd October 2006

Abstract

Traditionally, availability has been seen as an atomic property asserting the average time a system is "up" or "down". In order to model and analyse the availability of computerised systems in a world where the dependency on and complexity of such systems are increasing, this notion of availability is no longer sufficient. This report presents a conceptual model for service availability designed to handle these challenges. The core of this model is a characterisation of service availability by means of accessibility properties and exclusivity properties, which is further specialised into measurable aspects of service availability. We outline how this conceptual model may be refined to a framework for specifying and analysing availability requirements.

Contents

1 Introduction **5**

2 Requirements to a Refined Notion of Service Availability **6**
2.1 Classifying Availability 6
2.2 Classification of Threats and Means 7
2.3 Viewpoints for Analysing Availability 9
2.4 Requirements of Different Services 10
2.5 Measuring Availability 11

3 The Requirements Summed Up **12**

4 Properties of Service Availability **12**
4.1 Exclusivity 13
4.2 Accessibility 13

5 Means to Ensure Service Availability **14**
5.1 Incident Prevention 15
5.2 Incident Detection 16
5.3 Recovery from Incident 17

6 Threats to Service Availability **17**
6.1 Active Threats 18

7 Conceptual Model for Service Availability **19**

8 Conclusions **21**

References **23**

A Definitions **25**

B Abbreviations **28**

1 Introduction

Availability is an important aspect of today's society. Vital functions as e.g. air traffic control and telecom systems, especially emergency telecommunications services, are totally dependent on available computer systems. The consequences are serious if even parts of such systems are unavailable when their services are needed.

Traditionally, the notion of availability has been defined as the probability that a system is working at time t, and the availability metric has been given by the "uptime" ratio, representing the percentage of time that a system is "up" during its lifetime [20]. This system metric has been applied successfully worldwide for years in the PSTN/ISDN[1] telephony networks along with failure reporting methodologies [8]. This metric does not sufficiently measure important aspects of service availability.

With this traditional understanding, a web-based application such as a concert ticket sales service may have $99,999\%$ availability, however if it is down for the 5 minutes when concert tickets to a popular artist are put out for online sale while at the same tickets can be purchase via competing distributors, this means a considerable loss of profit for the adversely affected ticket sales website even though the service is considered to be highly available along traditional lines [2]. Service availability needs a more enhanced metric in order to measure availability in a way that meets the demands of today's services which have been shown to have much more bursty patterns of use than traditional PSTN/ISDN services [6].

Such burstiness in usage patterns also affects the ability of the service to provide to all users requiring the use of a service at a given moment, as illustrated in the following example. The Norwegian tax authorities provide on-line services for delivery of tax returns. In recent years, the service has been broadened to allow individuals to make changes to the return on-line (prior to this a report return had to be completed). In 2005 there was an increase in web-based returns to 1.255.000 in 2005 from 675.000 in 2004. However, the service was not able to handle the increase in demand on the final day, resulting in a large number of users being refused by the server. As a result, the tax authorities had to extend the deadline by 24 hours [21]. In the traditional sense, the service was still "up and running", and the hardware and software were still functioning correctly. Yet, a large number of users were being refused by the server, so that it was not available to a significant number of authorised users. The situation was exacerbated by the fact that the new users had much longer holding times than users filing web-based returns in 2004 due to the filling out of different forms in order to complete the changes to the tax return online. Up until 2005, only single form returns could be filed electronically. More complicated returns that require the user to fill out supplementary forms could not be filed electronically and had to be submitted on paper returns in the traditional way. In 2005, the online submission service allowed users with more complicated returns to file electronically. The result was that the number of users filing electronically increased, and many of the new users completing returns online had much longer holding times in order to fill out the additional forms as well as the main form. The increase in both penetration and usage parameters that was not foreseen

[1] Public Switched Telephone Network/Integrated Services Digital Network

resulted in loss of availability for a large number of users.

Indeed, as the environment where services are deployed becomes more and more complex [1] a more "fine-grained" view on "what is availability" is needed. Several global virus attacks have recently showed that availability is indeed affected by security breaches, e.g., when e-mail servers are flooded by infected e-mails, the availability for "real" e-mails decreases. Another example is the so called denial of service (DoS) attack, for which a service is overloaded with requests with the only purpose of making the service unavailable for other users.

In this report we motivate and introduce an augmented notion of service availability. In the heart of the resulting conceptual model lies a characterisation of availability as aspects of accessibility and exclusivity. Further, we seek to preserve well-established definitions from our main sources of inspiration to the extent possible: security, dependability, real-time systems, and quality of service (QoS). The report shows how the conceptual model may be used as a basis for specifying service availability requirements in a practical setting.

In Sect. 2 we provide the basis for our analysis of availability including our analysis of different viewpoints and approaches on availability and other aspects in the fields of security and dependability. Motivated by this discussion on related work in the fields of dependability and security research, we identify the requirements a conceptual model of service availability should satisfy. These requirements are summed up in Sect. 3. In Sect. 4 the properties of service availability are discussed, in Sect. 5 the means to achieve service availability are classified, and in Sect. 6 we present some of the threats to service availability. In Sect. 7 the overall conceptual model including a service availability measure is presented. Summary and conclusions are provided in Sect. 8. A list of definitions is provided in Appendix A as well as a list of acronyms and abbreviations in Appendix B.

2 Requirements to a Refined Notion of Service Availability

The setting for our availability analysis is derived from the fields of dependability and security, and we therefore strive to conform to the well-established concepts and definitions from these fields where there is a consensus. We also look to different approaches and viewpoints in dependability and security research to motivate and derive a set of requirements for a service availability concept model which enables an augmented treatment of availability that is more suited to securing availability in today's and future services.

2.1 Classifying Availability

Availability has been treated by the field of dependability and the field of security. The definitions of availability commonly used in these fields are:

1. Readiness for correct service [3].

2. Ensuring that authorised users have access to information and associated assets when required [11].

3. The property of being accessible and usable on demand by an authorised entity [9, 12].

We find the first of these definitions to be insufficiently constraining for practical application to design of services with high availability requirements. An integral part of securing availability is ensuring that that the service is provided to authorised users ; this is not addressed by the first definition. It is, however, addressed by the second, but neither the first nor the second captures the aspect of a service being *usable*. The third definition captures all of these aspects, and therefore is the basis for our analysis of availability and development of a more refined availability model.

In order to ensure service availability, it is essential to refine this notion to include addressing the aspect of ensuring that the service is provided to the authorised users *only*. The example of the on-line service for delivering tax returns given in Sect. 1 illustrates the importance of this aspect. Anyone may browse the Norwegian tax authorities information pages, although it is mainly for the use of Norwegian tax payers. However, access to the online submission service is for authorized taxpayers only. More importantly, a particular taxpayers forms must be available to that user only. The system must know how many authorised users are expected to access the service at the critical time, and the users holding times must be correctly estimated. For example, in order to calculate penetration and usage parameters, the total number of authorised users that are expected to access the service at a given time must be known. This is important to prevent the service from being overloaded. Additionally, it is also important to ensure that an individual tax form with details about a particular user's tax return is available to that particular user only.

The emergency telecommunications service (ETS) is an example that clearly shows the need to guarantee that authorised users only (in this case authorised emergency services personnel) can access and user the service during a disaster situation.

As already argued, there is a need to provide an enhanced classification and model of service availability in order to thoroughly analyse and enable the rigourous treatment of availability throughout the design process depending on the requirements of the individual services. *Our refined availability model should therefore characterise the properties/attributes of service availability including that services should be provided to the authorised users only.*

2.2 Classification of Threats and Means

The IFIP WG 10.4 view on dependability is elaborated in [3]. Fig. 1 shows the concept model of dependability as shown in [3].

This conceptual model of dependability consists of three parts: the *attributes* of, the *threats* to and the *means* by which dependability is attained [3]. This is a nice approach which motivates us to use a similar approach in our classification of service availability. *Clearly, threats to availability such as denial of service, and means to availability such as applying redundancy dimensioning techniques, have an important place in our availability model.*

In [3], the *means* by which dependability can be attained are fault prevention, fault tolerance, fault removal and fault forecasting. *Fault prevention*: how to prevent introduction of faults. *Fault tolerance*: how to deliver correct service in

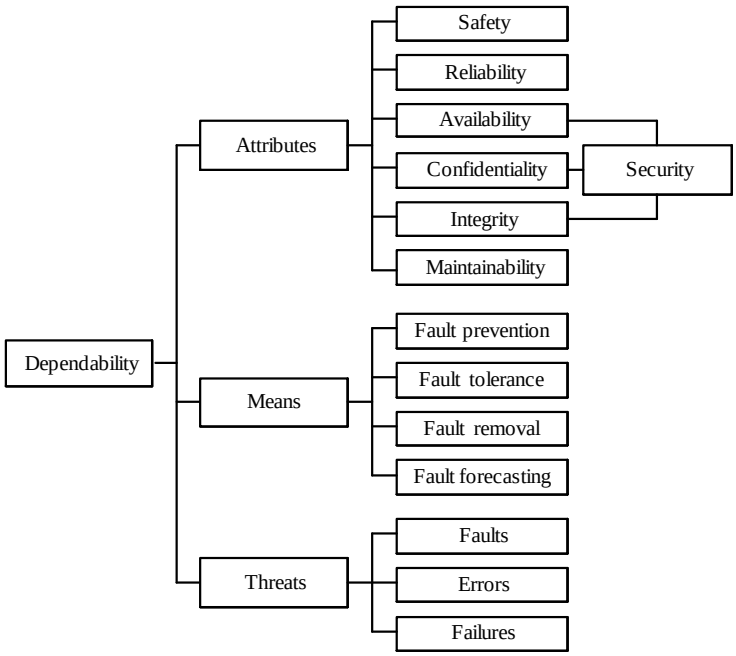

Figure 1: Conceptual model of dependability [3]

the presence of faults. *Fault removal*: how to reduce the number of faults, and finally *fault forecasting*: how to estimate the present number, future incidents and likely consequences of faults.

This approach does not address all of the means by which service availability can be obtained. This is because, incidents resulting in loss of service availability do not necessarily transpire due to faults and therefore classification of means in terms of faults as in [3] is, in our view, insufficient for availability analysis. An example is the hijacking of user sessions by an attacker or group of attackers, preventing the authorised user or group of users from accessing the service. This incident results in loss of service availability for a set of users, without incurring a fault in the system.

An *unwanted incident* is defined in [25] as an incident such as loss of confidentiality, integrity and/or availability. A fault is an example of an unwanted incident. Therefore, in order to classify threats to availability and means to achieve availability in a security setting, we are also motivated by the approach used in the security field of risk analysis and risk management as in [7, 15]. *The availability model should classify the means to achieve availability in terms of countering unwanted incidents.*

In [3], the *threats* to dependability are defined as faults, errors and failures, and these are seen as a causal chain of threats to dependability:

$$\text{fault} \longrightarrow \text{error} \longrightarrow \text{failure}$$

This understanding of threats serves nicely in the dependability model, however, we use the definition of threat, as defined in [12]: a *threat* is a potential cause of an unwanted event, which may result in harm to a system or organisation

and its assets. Unlike [3], we do not consider such a causal chain alone as the sole threats to availability, as service availability may be reduced by e.g. a denial of service (DoS) attack which reduces the service availability without causing a fault, error, or failure to the actual service itself. *The conceptual model of service availability should classify known threats to availability while conforming to existing literature on the classification of security threats.*

2.3 Viewpoints for Analysing Availability

For our availability analysis, it is appropriate to evaluate whether we should consider a system from a black box or white box perspective. In [14], E. Jonsson provides a conceptual model for security/dependability with a black box view as shown in Fig. 2.

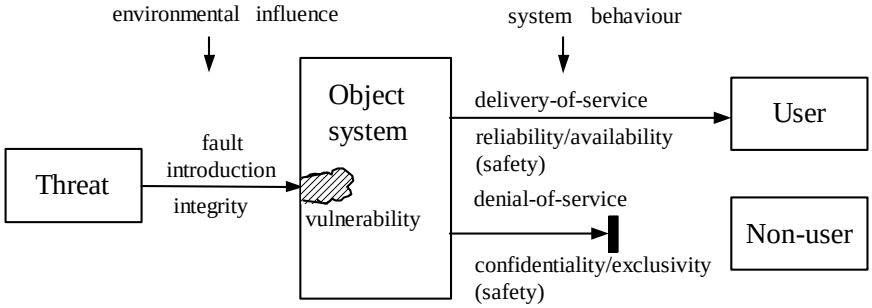

Figure 2: Jonsson's conceptual model [14]

In this system model view, Jonsson considers availability to be a purely behavioural aspect related to the outputs of the system, solely with respect to the users. As can been deduced from Fig. 2, exclusivity is a means to ensure availability. This viewpoint is valid and useful for some aspects of availability analysis; however, we see the need for evaluating availability from other viewpoints as well. Availability aspects of the internal components of the system must also be analysed.

We claim that aspects of availability must indeed be observed from both the input and output sides as well as the internal components of the system. For example, denial of service attacks can be observed as malicious input to a system to either flood the system and render it unavailable, or in order to alter the integrity of the system, e.g., by deleting a group of users from the database of authorised users. In the latter case, the input messages of the intruder can be observed, and the changes to the internal database, resulting in a loss of availability for those users that were deleted, will also be registered.

It is also important to observe and analyse the internal behaviour in the system in order to analyse the availability aspects of components, in particular service components which collaborate to deliver the service. Motivated by a service-oriented system view, only a whitebox view allows and facilitates the specification of the internal means to achieve availability and the examination of internal causes that affect availability. *The conceptual model should therefore address internal as well as external concerns of availability.*

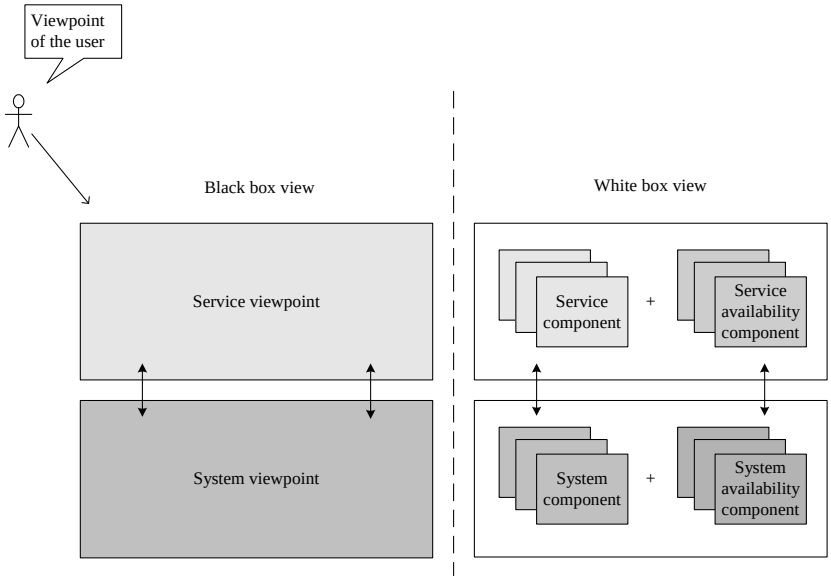

Figure 3: Viewpoints for analysing availability

Fig. 3 summarizes the different concerns for analysing availability. From the point of view of the user, the service is either available, or it is not. The system view is well understood in the dependability field, and as discussed above, Johnson provides an evaluation from a system viewpoint and with a security point of view. The Service Availability Forum (SAF) is working on standardising middleware for the open interfaces between the layers [22], as shown in Fig. 4 and discussed in Sect. 2.4. In our work on securing availability in service composition, we are analysing availability from the decomposed service viewpoint, according to requirements of the users.

2.4 Requirements of Different Services

In the current and future telecommunications market, there are many different types of services each of which may have different requirements with respect to availability. Telephony services, and in particular, emergency services, are examples of services with stringent availability requirements. Internet-based services, however, have somewhat different requirements. Requirements for what may be tolerated of delays or timing out of services are rather lax currently for e.g., online newspaper services. Yet, a citizen who leaves the tax return to the last minute before the deadline for filing requires urgently that the online tax return submission service is available at that particular moment [21].

For traditional telecommunications services, the availability requirement of $99,999\%$ availability is still valid, however, it does not sufficiently address all of the differentiated requirements with respect to service availability. More precisely, as advocated by the Service Availability Forum (SAF) [22], there is also a need for a customer centric approach to defining availability requirements. The availability concern of the Service Availability Forum is readiness for correct service and in particular continuity of service, with a focus on the demands of

the customers.

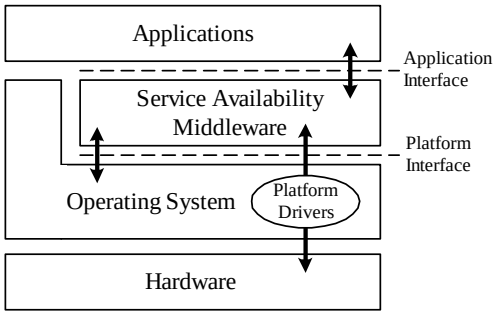

Figure 4: The SAF framework [22]

Service availability as defined by the SAF aims to meet the following demands:

- Customers demand and expect continuous service availability.
- Customers want always-on services and connections that are maintained without disruption-regardless of hardware, software, or operator-caused system faults or failures.

The availability concern of the SAF is readiness for correct service and in particular continuity of service, with a focus on the demands of the customers. The SAF is concerned with availability of today's systems from the dependability perspective providing a transition from the application of dependability to traditional telecommunications systems to current systems which are distributed.

We intend to incorporate the ideas of the SAF in our model, to enable customer oriented availability requirements, however, extending these to include the aspects of ensuring that unauthorised users cannot interrupt, hijack, or prevent the authorised users from accessing a service. *The model must address the service availability requirements in a flexible manner, in order to address the different aspects of availability.*

2.5 Measuring Availability

As discussed in the introduction, we need a more fine grained measure of availability than pure "up" or "down". Services can exist in numerous degraded but operational/usable/functional states between "up" and "down" or "correct" and "incorrect". For example, an online newspaper may behave erratically with slow response times for displaying articles browsed without going down or becoming completely unavailable. It should be possible to describe various states of availability in order to specify just how much a reduction of service quality may be tolerated.

While both the Common Criteria [10] and Johnson [14] define security measures and provide techniques for measuring security in general, there is a need for a more fine grained metric for measuring service availability that takes into account, for example, measurement of how well user requirements are fulfilled, as well as a need for measuring the ability to adequately provision a service

to all of the authorised users requiring the service at a given moment. Such a metric needs to take into account the appropriate set of parameters, not just the usual average based on the Mean Time To Failure (MTTF) and the Mean Time To Repair (MTTR). *Our aim is to incorporate techniques from the existing initiatives in the fields of security and dependability in order to arrive at a more complete composite measure of service availability.*

3 The Requirements Summed Up

Based on the above discussion we arrive at a set of requirements for the conceptual model.

1. *The model should characterise the properties of service availability.* This is to enable the rigourous treatment of service availability depending on the requirements of the individual services.
2. *The model should characterise the means to achieve service availability in terms of countering unwanted incidents.*
3. *The model should classify known threats to service availability while conforming to existing literature on the classification of security threats.*
4. *The model should address internal and external concerns of availability.* With a black box view only, only the externally observable properties of availability can be studied. Using a white box view the internal means to achieve availability can be specified and internal causes that affect service availability can be examined.
5. *The model should facilitate specification of service availability requirements in a flexible manner, in order to address the different aspects of availability.* There are many different types of services, and they may have different requirements with respect to availability. Availability requirements should be flexible enough to address the different services consistently.
6. *The model should provide a basis for defining a service availability metric.* Our aim is to incorporate techniques from the existing initiatives in the fields of security and dependability in order to arrive at a more complete composite measure of availability.

4 Properties of Service Availability

We claim that service availability encompasses both exclusivity, the property of being able to ensure access to authorised users only, and accessibility, the property of being at hand and useable when needed. As such, contrary to, e.g., [3], which treats availability as an atomic property, we see service availability as a composite notion consisting of the following aspects:

- Exclusivity
- Accessibility

We elaborate on these two properties in Sect. 4.1 and Sect. 4.2.

4.1 Exclusivity

By *exclusivity* we mean the ability to ensure access for authorised users only. More specifically, this involves ensuring that unauthorised users cannot interrupt, hijack, or prevent the authorised users from accessing a service. This aspect is essential to prevent the denial of legitimate access to systems and services. That is, to focus on prohibiting unauthorised users from interrupting, or preventing authorised users from accessing services. Our definition of exclusivity involves both users and non-users, i.e., ensuring access to users while keeping unauthorised users out. This is in order to properly address means of achieving exclusivity some of which will address ensuring access for authorised users and others will address techniques for preventing unauthorised users from accessing or interrupting services.

The goal with respect to exclusivity is to secure access to services for authorised users in the best possible way. Essentially this means:

- Secure access to services for the authorised users.
- Provide denial of service defence mechanisms. Here we focus on prohibiting unauthorised users from interrupting, or preventing users from accessing services.
- Ensure that unauthorised users do not gain access to services.

Note that attacks via covert channels or by eavesdropping can lead to loss of confidentiality without loss of exclusivity as the attacker is not accessing the service, but passively listening in on service activity. Confidentiality, however, consists of exclusivity and absence of unauthorised disclosure of information.

4.2 Accessibility

We define *accessibility* as the quality of being at hand and usable when needed. The notion of "service" is rather general, and what defines the correctness of a service may differ widely between different kinds of services. Accessibility is related to QoS [4, 18, 26], but what is considered relevant qualities vary from one domain to another. Furthermore, QoS parameters tend to be technology dependent. An example of this is properties like video resolution and frame rates [26], which are clearly relevant for IP-based multimedia services and clearly not relevant in other service domains, such as SMS or instant messaging services.

What all services do seem to have in common is the requirement of being timely; for a service to be accessible it must give the required response within reasonable time. In addition to being timely, a service will be required to perform with some quality to be usable. Hence, we divide accessibility properties into two major classes of properties: *timeliness* properties and *quality* properties. Timeliness is the ability of a service to perform its required functions and provide its required responses within specified time limits. A service's quality is a measure of its correctness and/or how usable it is.

Consider an online booking service. From the viewpoint of a user at a given point in time, we could say that the quality of the service is either 1 or 0 depending on whether the user gets a useful reply (e.g. confirmation) or unuseful reply (e.g. timeout). (Over time this can be aggregated to percentages expressing how often one of the two kinds of responses will be given.)

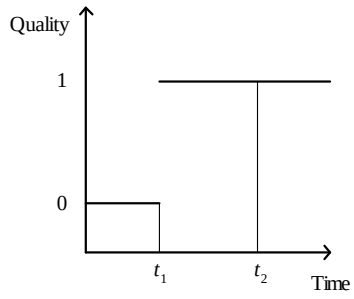

Figure 5: Quality vs. timeliness

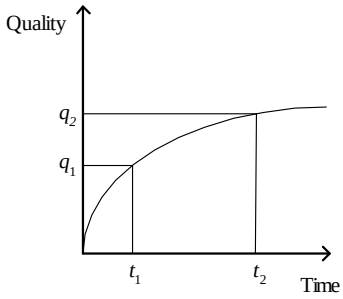

Figure 6: Quality vs. timeliness

In a multimedia service like video streaming, the frame rate may be seen as a timeliness property (each frame should be timely) while the resolution of each frame and the colour depth are quality properties.

In both these examples we may see a dependency between timeliness and quality. In the first example (Fig. 5) we may assume a deadline t_2 for the response to the user for the service to be accessible. However, we must also assume some processing time t_1 for the service to be able to produce an answer. This means that the quality requirement enforces a lower bound on the timeliness; if the deadline is too short the user will always receive the timeout message. In other words we must have that $t_1 < t_2$ for the service to be accessible.

In the other example (Fig. 6) we may assume that higher quality requires more processing time per frame. This means that a required quality q_1 provides a lower limit t_1 on the processing time of each frame. Further, to get the required frame rate there must be a deadline t_2 for each frame, which provide an upper bound q_2 on the quality. This means the service must stay between this lower and upper bound to be accessible. This approach may be seen as an elaboration of Meyer's concept of *performability evaluation* [16].

These considerations motivates a notion of *service degradation*. We define service degradation to be reduction of service accessibility. Analogous to accessibility we decompose service degradation into timeliness degradation and quality degradation, and see that these are quantities mutually dependent on each other. For example, graceful degradation in timeliness may be a way of avoiding quality degradation if resources are limited, or the other way around. A combination of graceful degradation in timeliness and graceful degradation in quality may also be applied. Related to QoS, accessibility may actually be considered a QoS tolerance cut-off, i.e., the point at which the QoS deteriorates to a level where the service is deemed no longer usable, so that the service is considered unavailable.

5 Means to Ensure Service Availability

Traditionally, the approach to meeting availability requirements has primarily focused on ensuring accessibility aspects of availability such as by introducing redundancy, and by service replication. This is a valid approach to availability, but it does not ensure, e.g., that the service is accessible to authorised users only. There are costs involved in introducing redundancy and replication, which need

to be justified. The goal should be to obtain more comprehensive, more cost-effective means to achieving availability, and to specify, design, and implement a set of measures that enable delivery of services and/or systems according to availability requirements.

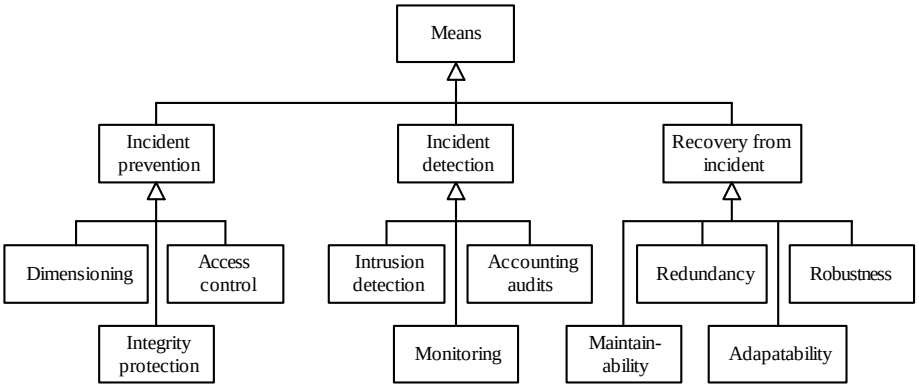

Figure 7: Means to ensure availability

By means to ensure availability we address *protection* of the service from incidents leading to a loss of availability. Therefore, in our model as shown in Fig. 7 (and represented in UML 2.0 [17]), we categorise the means into the following three groups: *incident prevention*: how to prevent incidents causing loss of availability; *incident detection*: how to detect incidents leading to loss of availability; and *recovery from incident*: the means to recover after an incident has lead to a loss of availability. We do not attempt to create an exhaustive list of all such measures, but do provide examples that illustrate the different aspects of securing service availability.

5.1 Incident Prevention

Preventative means are defined as the internal aspects of a system that are designed to prevent, stop or mitigate intrusions, faults, errors, or other incidents which have a negative effect on the availability of a system.

Access control is an important preventative means for achieving the exclusivity aspect of service availability. Access control is the prevention of unauthorised use of a resource, including the prevention of use of a resource in an unauthorised manner [9].

Providing *integrity protection* mechanisms is important for example, in order to protect against manipulation and redirection of messages resulting in denial of service for the authorised user. For example, without message integrity protection, an unauthorised user may manipulate messages in a man-in-the-middle attack and redirect all messages to her/him instead of to the authorised user, resulting in denial of service for the authorised user.

On the other hand, it is important to ensure that the required resources e.g. in the network that an authorised user has permission to use during a session are indeed allocated to the user to ensure that the service is delivered according to the user availability requirements. We have grouped this in under *dimensioning*. The purpose of dimensioning is to ensure that expected needs

will be met in an economical way, for subscribers, service providers and network operators [19]. Dimensioning techniques involve ensuring that processors are not overloaded, and that enough resources are provided. Correct resource allocation is an essential problem of denial of service protection. A means to controlling this is a combination of policy and access control functions. Another aspect of dimensioning is scalability, ensuring that the solution adapts easily so that the service may be available to a large number of users at a reasonable cost.

Another example of a means for avoiding loss of service availabilty is *graceful degradation* [23], that is degradation of a system in such a manner that it continues to operate, but provides a reduced level of service rather than failing completely. By applying graceful degradation schemes a complete loss of availability can be prevented.

5.2 Incident Detection

Incident detection consists of means to discover incidents such as denial of service attacks, faults, errors or failures, which lead to a loss or reduction of availability.

Detective measures will commonly be coordinated with recovery aspects of the system in order to adapt and restore system availability. Fault detection, traffic flow monitoring, intrusion detection systems (IDS), and security audits are all examples of detective measures.

Fault detection encompasses identifying and locating faults in the system, this includes detection of faults that may not be detected by the external behaviour and may not cause an immediate incident, but have the potential to cause an incident in another situation. Traffic flow monitoring is important to detect anomalies in the traffic flows, as well as for ensuring the QoS guarantees can be met. Active traffic monitoring for which traffic flows are actively generated to test the capacity, is used to specifically measure performance capacity and is a useful tool for defining dimension rules. Passive monitoring of the real traffic is useful as a detective means in identifying performance problems. Passive monitoring provide information on the actual behaviour of the communications traffic and is useful in understanding communications requirements. Traffic flow measurements can be used to optimise network usage as well as to analyse traffic under congestion conditions with respect to the type of traffic, the origin of the traffic, and the dynamic behaviour of the traffic e.g., its burstiness. Intrusion detection systems provide key components used to obtain information about unwanted activity on the network or within computer systems. With an intrusion detection system, it is possible to forensically collect records on an attack or break-in even though an attacker has deleted service logs. By accounting audits we mean mechanisms for service usage accounting and resource usage accounting. This is useful for e.g., analysing unauthorised use of the system or services. Accountability is the property that ensures that the actions of an entity may be traced uniquely to the entity.

For an efficient approach to unwanted incident detection, it is wise to combine monitoring, fault detection and IDS techniques along with audit logs generated and process the information and data collected in real time or close to real time in order to detect and thwart attacks or incidents that have the potential to result in loss or reduction of availability.

5.3 Recovery from Incident

Recovery from incident consists of the means to recover from incidents leading to loss or reduction of availability. This includes techniques for adapting the service, e.g. in the case that anomalies are detected by the IDS so that major unwanted incidents of loss of availability are avoided. Recovery means may entail, e.g., making changes to the internal aspects of the system, such as correction of faults or removal of system vulnerabilities. Additionally, external filters may be implemented to filter away the discovered cause of the incident such as malicious traffic or traffic from unauthorised users. Recovery addresses the *adaptability*, *robustness*, *maintainability* and *redundancy* aspects of the system.

In a modular *redundancy* configuration, the original system is multiplied into a number of identical subsystems which are simultaneously active. In a standby redundant system there are two or more copies of the original systems. Only one of the copies is active at a time. The above-mentioned redundancy techniques are most appropriate for achieving fault tolerance with respect to hardware failures. In case of software fault tolerance the concept of N-version programming is often applied. Using a common software specification the system is developed in a number of different versions by separate teams. These versions are executed on separate computers and inconsistent outputs are rejected. However, these techniques were not developed to address, for example, damages to software due to exploitation of vulnerability in an attack on the system, such as a denial of service attack. In this case additional techniques need to be applied to return the system to normal operation as well as enabling the system to adapt and remove vulnerabilities. *Robustness* is the degree to which a system or component can function correctly in the presence of invalid or conflicting inputs. Based on the discussion above it is clear that techniques such as software and hardware redundancy contribute to the robustness of systems. *Maintainability* is the ability to undergo modifications and repairs and is also important for the recovery aspect of accessibility. Maintenance is carried out either in a preventive way with the intention to reduce probability of failure or in a deferred way, which is to perform maintenance after failures have occurred. The latter approach represents the opposite of preventive maintenance and is often motivated by reduced maintenance costs. *Adaptability* in the event of reduction or loss of service is also an important means of restoring availability, at least partially. Routing mechanisms in the Internet Protocol is an example of a means to ensure that the Internet adapts if a segment or segments are down to ensure that service is restored as soon as possible and possibly without the users being aware of the reduction of service.

6 Threats to Service Availability

As we stated in the discussion above in Sect. 2.2 for the basis of analysis of availability we use the definition of threat, as defined in [12]. A *threat* is a potential cause of an unwanted event, which may result in harm to a system or organisation and its assets. In our case the asset of interest is the availability of the service.

It is common to distinguish between active and passive threats. An active threat is when something or someone attempts to alter system resources or

affect the operation of a system, while a passive threat is when something or someone tries to get hold of information from a system without affecting the system resources. [24]

It is obvious that a pure passive threat alone does not affect the availability of a service, but passive threats may of course be part of a larger threat to the availability of a system. For this reason we concentrate on active threats and do not go into passive threats in this report.

6.1 Active Threats

The most explicit threats to service availability are *DoS* attacks. *Replay*, *masquerade*, *modification of messages*, *man-in-the-middle attack* and *misuse of service*, as shown in Fig. 8 (and represented in UML 2.0 [17]), are examples of other kind of active threats that may affect availability. Threats may originate on the inside (inside attackers) or the outside (outside attackers) of the system. The impact of threats varies with the nature of the threats; some threats may result in degradation of the service, others in complete loss of service. Going into detail on this issue is outside the scope of this report, but below we give some examples on how some of these threats may affect service availability.

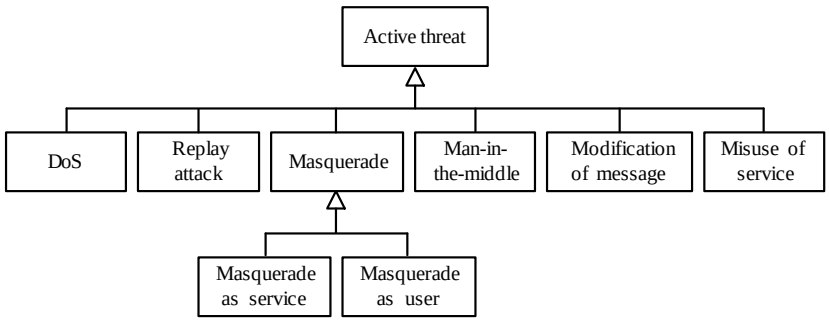

Figure 8: Active threats

DoS attacks may lead to loss of use due to unauthorised use of the service preventing authorised users from accessing the service. Unauthorised use may also create over-usage problems having an overload effect and in this way degrading the quality of the service for the authorised users. Examples of DoS attacks on access to network resources are e.g. “ping of death” and “Smurf” attack [5].

In a replay attack, the attacker captures the authentication credentials of an authorised user and replays the authentication message at a later time to obtain access to a service.

In a masquerade, an attacker steals the identity of a real user and obtains fraudulent access by masquerading as the real user while preventing the valid user from accessing services. Or, the other way around, an attacker replaying or masquerading as a service may deceive the user, and the service the user intended to access is then not available.

7 Conceptual Model for Service Availability

Based on the requirements from Sect. 2 and our discussion above we propose the overall model presented in Fig. 9 (represented in UML 2.0 [17]) and further explained in the following text.

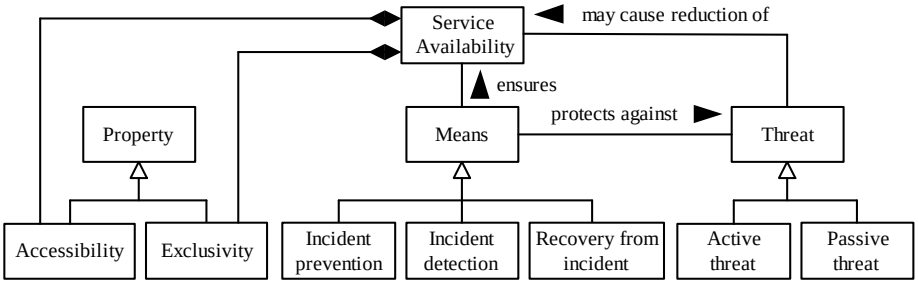

Figure 9: The overall picture

In the figure the relationships between availability, threats and means are shown. Availability is affected by means and threats. Means ensures availability by preventing and countering unwanted incidents and protects against threats. Threats may lead to unwanted incidents causing reduction of availability.

There are many different types of services, and they may have different requirements with respect to availability. Availability requirements should be flexible enough to address the different services consistently. We propose that availability is specified by the means of availability policies and predicates over measurable properties of services, and that these policies and predicates are decomposed in accordance with the decomposition of availability in the conceptual model. An availability policy consists of an accessibility policy (e.g., required resources) and an exclusivity policy (e.g., which entities have permissions to use the service or system).

The predicates place conditions on the allowed behaviour of the service. In order to express these predicates, there is a need to describe rules for allowed or prohibited behaviour and to provide a means for measuring the availability properties of a service. Figure 10 illustrates how availability properties are related to services, i.e., as part of the relation between the service and the user entity using the service.

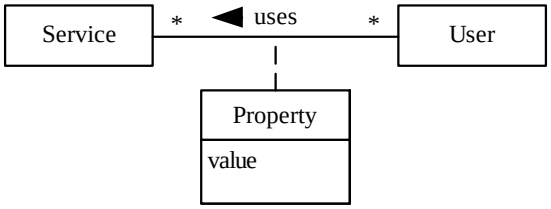

Figure 10: Service availability

Our conceptual model provides the foundation for an availability metric in

that it provides decomposition of availability properties that may be mapped to measurable quantities. This metric will include behavioural measures, preventative measures, and correctness measures such as the measurement of degree of degradation.

The following is the mathematical representation of the availability metric for a service. Let A denote a service with an availability property for a user group U, and let X denote the availability metric for service A. We represent X as an n-tuple $X = (x_1, \ldots, x_n)$ where x_i is a measure of an aspect of service availability. Using our conceptual model this idea can be refined as follows: We represent X as a tuple $X = (X_1, X_2)$ where X_1 measures the exclusivity properties, and X_2 measures the accessibility properties.

Essentially, the aim is to measure and determine the degree of accessibility and exclusivity that is sufficient for the authorised user to be able to activate and use the service. The purpose of measurements is to establish that service availability requirements have been met. For example, in order to address how well the system keeps users while still granting access to authorised users we have the following exclusivity requirements:

- The probability that an authorised user is denied access to the service at a given time t should be less than x.
- The probability that an unauthorised user obtains access to the service at a given time t should be less than y.
- User u shall be prohibited from accessing service s when user v is using the service.
- The number of intrusions at a given time t should be less than z.

Based on these requirements, we have the following measures of aspects of exclusivity:

- The probability that an authorised user is denied access to the service at a time t.
- The probability that an unauthorised user obtains access to the service at a given time t.
- The probability that user u obtains access to service s when user v is using the service.
- The number of intrusions at a given time t.

Similar measures may be defined for accessibility. These may be defined with a basis in measures for service degradation, timeliness, performance, and quality.

Service availability metrics can be derived in this manner to measure the ability to meet each of the exclusivity and accessibility requirements. Service availability measurements can then be designed for observing the specific parameters identified. For example, for a voice over IP (VoIP) call service, an accessibility requirement may be that call-set up time is required to be less than x ms, and the measurement accumulates the amount of time that the system is in this (call-set up) state. Similarly, regarding a call-blocking requirement for the

VoIP service, the percentage of calls by authorised users blocked may be measured. Additionally, fine-grained measurements may also be derived regarding establishing the causes of the call-blocking, e.g. observing and measuring the presence of intrusive/DoS behaviour resulting in the blocking of calls.

In order to apply the model, the availability requirements must be determined. Threats must then be analysed to understand what affects availability and means for ensuring availability need to be identified to meet requirements and counter threats. Measurements of the different aspects are then used to evaluate how well the availability requirements are met. We are currently applying the model to our work on ensuring availability in service composition.

8 Conclusions

The contribution of this report is a conceptual model for service availability that takes into account a much broader spectrum of aspects that influence availability than previously addressed by work in this area. We have argued that exclusivity is an aspect of availability that has been generally neglected in the literature, and shown where it fits in an enhanced notion of service availability. Further we have shown how QoS, real time and dependability considerations may be integrated in the model and treated as accessibility properties.

We have established that there is a need for a more fine grained metric for measuring availability and have provided a representation of the availability metric for a service that allows specification of the measurable requirements for exclusivity and accessibility properties.

Our conceptual model for availability embraces both a white box view as well as a black box view of availability and, hence, addresses both internal and external concerns of availability. The need for this is apparent in our current work on ensuring availability in service composition that encompasses a collaboration of roles, which are slices of behaviour across distributed systems. These must be composed correctly in order to achieve a service with the required availability.

The model also contains a classification of threats to availability and means to ensure availability, and establishes the relationship between threats, means and availability properties. Together these elements provide a framework in which all relevant views and considerations of availability may be integrated, and a complete picture of service availability may be drawn.

References

[1] W. A. Arbaugh, W. L. Fithen, and J. McHugh. Windows of vulnerability: A case study analysis. *IEEE Computer*, 33(12):52–59, December 2000.

[2] C. Asker. Billettkaos for U2-konserten. February 5, 2005 [online] – URL : http://www.aftenposten.no/kul_und/musikk/article963497.ece. (In Norwegian).

[3] A. Avižienis, J.-C. Laprie, B. Randel, and C. Landwehr. Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing*, 1(1):11–33, January-March 2004.

[4] M. Barbacci, M. H. Klein, T. A. Longstaff, and C. B. Weinstock. Quality attributes. Technical report, Software Engineering Institute, Carnegie Mellon University, December 1995.

[5] CERT Advisory CA-1998-01 Smurf IP Denial-of-Service Attacks. March 9, 2004 [online] – URL : http://www.cert.org/advisories/CA-1998-01.html.

[6] D. Clark, W. Lehr, and I. Liu. Provisioning for bursty Internet traffic: Implications for industry and Internet structure. MIT ITC Workshop on Internet Quality of Service, 1999.

[7] F. den Braber, M. Soldal Lund, K. Stølen, and F. Vraalsen. Integrating security in the development process with UML. In *Encyclopedia of Information Science and Technology*, pages 1560–1566. Idea Group, 2005.

[8] P. Enriquez, A. B. Brown, and D. A. Patterson. Lessons from the PSTN for dependable computing. Workshop on Self-Healing, Adaptive and self-MANaged Systems (SHAMAN), 2002.

[9] International Standards Organization. *ISO 7498-2, Information Processing Systems – Interconnection Reference Model – Part 2: Security Architecture*, 1989.

[10] International Standards Organization. *ISO/IEC 15408, Information technology – Security techniques – Evaluation criteria for IT security*, 1999.

[11] International Standards Organization. *ISO/IEC 17799, Information technology – Code of practice for information security management*, 2000.

[12] International Standards Organization. *ISO/IEC 13335, Information technology – Security techniques – Guidelines for the management of IT security*, 2001.

[13] International Telecommunication Union. *ITU-T E.800, Quality of Telecommunication Services – Terms and Definitions Related to Quality of Service and Network Performance Including Dependability*, 1994.

[14] E. Jonsson. Towards an integrated conceptual model of security and dependability. In *Proc. The First International Conference on Availability, Reliability and Security*, pages 646–653. IEEE Computer Society, 2006.

[15] M. S. Lund, F. den Braber, and K. Stølen. Maintaining results from security assessments. In *Proc. Seventh European Conference on Software Maintenance and Reengineering (CSMR)*, pages 341–350. IEEE Computer Society, 2003.

[16] J. F. Meyer. Performability evaluations: Where it is and what lies ahead. In *Proc. International Computer Performance and Dependability Symposium*, pages 334–343. IEEE, 1995.

[17] Object Management Group. *UML 2.0 Superstructure Specification, document: ptc/04-10-02 edition*, 2004.

[18] Object Management Group. *UML Profile for Modeling Quality of Service and Fault Tolerance Characteristics and Mechanisms*, 2005.

[19] A. Olsson. *Understanding Telecommunications: Building a Successful Telecom Business*. Studentlitteratur, 2004.

[20] S. M. Ross. *Inroduction to probability models*. Academic Press, 6th edition, 1997.

[21] E. Ryvarden. Skatte-servere tålte ikke trykket. April 30, 2005 [online] – URL : http://www.digi.no/php/art.php?id=213094&utskrift=1. (In Norwegian).

[22] Service Availability Forum. SAF Backgrounder. March 6, 2004 [online] – URL : http://www.saforum.org/home.

[23] K. G. Shin and C. L. Meissner. Adaption and graceful degradation of control system performance by task reallocation and period adjustment. In *Proc. 11th Euromicro Conference on Real-Time Systems*, pages 29–36. IEEE, 1999.

[24] R. Shirey. *Internet Security Glossary. RFC 2828*. Network Working Group, 2000.

[25] Standards Australia. *AS/NZS 4360:1999, Risk Management*, 1999.

[26] A. Vogel, B. Kerherve, G. von Bochmann, and J. Gecsei. Distributed multimedia and QoS: A survey. *IEEE Multimedia*, 2(1):10–18, 1995.

A Definitions

This appendix contains a list of definitions of terms used in this report. The definitions are obtained from international standards to the extent possible, and from established sources in the literature. For terms that are defined differently in the standards, the order of prioritization is as follows: [9] first, then [12], [25], and [24].

Access control: The prevention of unauthorised use of a resource, including the prevention of use of a resource in an unauthorised manner [9].

Accessibility: The quality of being at hand and usable when needed.

Accounting audit: Mechanism for service usage accounting and resource usage accounting. This is useful for e.g., analysing.

Accountability: The property that ensures that the actions of an entity may be traced uniquely to the entity [9].

Active threat: When something or someone attempts to alter system resources or affect the operation of a system [24].

Asset: Anything that has value to an organisation [12].

Adaptability: The ability to change or be changed to fit changed circumstances

Attack: An assault on system security that derives from an intelligent threat, i.e., an intelligent act that is a deliberate attempt (especially in the sense of a method or technique) to evade security services and violate the security policy of a system [24].

Authorisation: The granting of permission based on authenticated identification [9].

Authorised: Granted rights or permissions [24].

Availability: The property of being accessible and usable on demand by an authorised entity [9, 12].

Confidentiality: The property that information is not made available or disclosed to unauthorised individuals, entities, or processes [12].

Data integrity: The property that information has not been altered or destroyed in an unauthorised manner [12].

Denial of service: The prevention of authorised access to resources or the delaying of time critical operations [9].

Dependability: The ability to deliver service that can justifiably be trusted [3].

Dimensioning: The purpose of dimensioning is to ensure that expected needs will be met in an economical way, for subscribers, service providers and network operators [19]

Eavesdropper: A person that does a passive wiretapping done secretly, i.e., without the knowledge of the originator or the intended recipients of the communication [24].

Exclusivity: The ability to ensure access for authorised users only.

Failure: A termination of the ability of a functional unit to perform a required function [25].

Fault: Abnormal condition that may cause a reduction in, or loss of, the capability of a functional unit to perform a required function [25].

Fault forecasting: How to estimate the present number, future incidents and likely consequences of faults [3].

Fault prevention: How to prevent introduction of faults [3].

Fault removal: How to reduce the number of faults [3].

Fault tolerance: How to deliver correct service in the presence of faults [3].

Graceful degradation: Degradation of a system in such a manner that it continues to operate, but provides a reduced level of service rather than failing completely [23].

Incident detection: How to detect incidents leading to loss of availability.

Incident prevention: How to prevent incidents causing loss of availability.

Integrity: See data integrity and system integrity [12].

Integrity protection: Protection of from unauthorised modification. This applies to both data integrity and system integrity.

Intrusion detection: A security service that monitors and analyzes system events for the purpose of finding, and providing real-time or near real-time warning of, attempts to access system resources in an unauthorized manner [24].

Man-in-the-middle attack: A form of active wiretapping attack in which the attacker intercepts and selectively modifies communicated data in order to masquerade as one or more of the entities involved in a communication association [24].

Maintainability: The ability to undergo modifications and repairs [3].

Masquerade: A type of attack in which one system entity illegitimately poses as (assumes the identity of) another system entity [24].

Misuse: A threat action that causes a system component to perform a function or service that is detrimental to system security [24].

Modification of message: Altering message contents.

Monitor: To check, supervise, observe critically, or record the progress of an activity, action or system on a regular basis in order to identify change [25].

Non-repudiation: The ability to prove an action or event has taken place, so that this event or action cannot be repudiated later [9, 12]

Passive threat: When something or someone tries to get hold of information from a system without affecting the system resources [24].

Performability: A system's ability to perform when performance degrades as a consequence of faults [16].

Preventative means: Internal aspects of a system that are designed to prevent, stop or mitigate intrusions, faults, errors, or other incidents which have a negative effect on the system.

Privacy: The right of individuals to control or influence what information related to them may be collected and stored and by whom and to whom that information may be disclosed [9].

Quality: A measure of a service's correctness and/or how usable it is.

Quality of service (QoS): The collective effect of service perfomances, which determine the degree of satisfaction of a user of the service [13].

Recovery from incident: The means to recover after an incident has lead to a loss of availability.

Replay attack: An attack in which a valid data transmission is maliciously or fraudulently repeated, either by the originator or an adversary who intercepts the data and retransmits it, possibly as part of a masquerade attack [24].

Redundancy: Replication of the original systems.

Reliability: The property of consistent intended behaviour and results [12].

Residual risk: The risk that remains after safeguards have been implemented [25].

Risk: The potential that a given threat will exploit vulnerabilities of an asset or group of assets and thereby cause harm to the organization [12].

Risk analysis: A systematic use of available information to determine how often specified events may occur and the magnitude of their consequences.

Risk management: The culture, processes and structures that are directed towards effective management of potential opportunities and adverse effects.

Robustness: The degree to which a system or component can function correctly in the presence of invalid or conflicting inputs.

Safeguard: A practice, procedure or mechanism that reduces risk [12].

Service degradation: Reduction of service accessibility.

System integrity: The property that a system performs its intended function in an unimpaired manner, free from deliberate or accidental unauthorised manipulation of the system [12].

Security audit: An independent review and examination of system records and activities in order to test for adequacy of system controls, to ensure compliance with established policy and operational procedures, to detect breaches in security, and to recommend any indicated changes in control, policy and procedures [9].

Threat: A potential cause of an unwanted event, which may result in harm to a system or organisation and its assets [12].

Timeliness: The ability of a service to perform its required functions and provide its required responses within specified time limits.

Timeliness degradation: Reduction of a service's timeliness.

Traffic analysis: The inference of information from observation of traffic flows (presence, absence, amount, direction, and frequency) [9].

Unwanted incident: Incident such as loss of confidentiality, integrity and/or availability [25].

Usable: Capable of being used.

Vulnerability: A weakness of an asset group or group of assets, which can be exploited by one or more threats [12].

B Abbreviations

DoS Denial of Service

ETS Emergency Telecommunications Service

IP Internet Protocol

PSTN/ISDN Public Switched Telephone Network/Integrated Services Digital Network

QoS Quality of Service

SAF Service Availability Forum

SMS Short Message Service

UML Unified Modelling Language

VoIP Voice over Internet Protocol