What is Availability?

JUDITH E.Y. ROSSEBØ

Judith E.Y. Rossebø is Senior Research Scientist with Telenor R&I. She is also affiliated with The Norwegian University of Science and Technology (NTNU)

Availability has commonly been considered as an atomic property, indicating the average amount of time a system is working as planned. While this notion of availability has served well for describing and analysing voice services in the PSTN/ISDN, it does not fit well with the needs of many next generation network (NGN) [1] services which have been shown to have more bursty usage patterns and more complex correctness and usage requirements than traditional telephony services. For example, in a multimedia service, video resolution requirements and frame rate are important considerations. On the other hand, a telemedicine service should be accessible and usable to authorized users only, at the exact moment that it is needed, and for the duration of the telemedicine session. This article discusses availability in the context of today's and future networks and systems, providing an enhanced notion of availability and a concept model that can be applied to address availability concerns.

1 Introduction

Securing availability of applications and services is increasingly important for service provisioning in (today's and) future networks and systems. In the past, this task was easily managed, as in each country, a single operator delivered telephony services over a dedicated network, and the regulatory authorities set the standards for requirements (eg. availability requirements) and enforced these. The telecommunications services and the network were designed for each other. In this setting, availability has commonly been considered as an atomic property, indicating the average amount of time a system is working as planned.

Over the past 20 years, this situation has evolved. Now, there is a range of different types of services provided over a range of different networks. Today, the PSTN/ISDN and other vertical networks are being replaced by common IP-based networks and services in a distributed environment, the so-called next generation networks (NGN) [1], as shown in Figure 1. There is a range of services provided over a variety of access networks with a common IP core as shown in Figure 2. The NGN brings an increasing demand for new multimedia and networking services such as YouTube, Facebook and interactive IPTV services. These are delivered in a multi-provider, hyper-connected environment. Services are also market driven and users are playing a role in setting requirements (whereas before service provisioning was much more influenced by regulatory and national requirements). Service availability is an important concern in this enhanced, multimedia service environment.

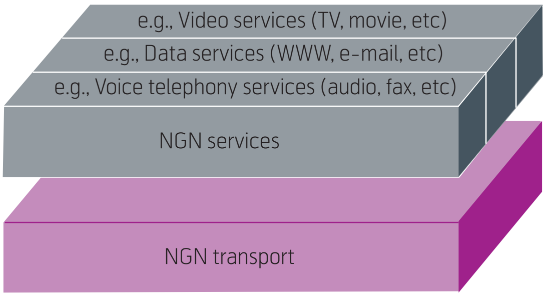
e.g., Video services (TV, movie, etc)
e.g., Data services (WWW, e-mail, etc)
e.g., Voice telephony services (audio, fax, etc)
NGN services
NGN transport

Figure 1 Separation of services from transport in NGN [1]

Motivated by cost savings, more and more telecom services are being migrated from deployment over dedicated networks to deployment over a common IP-based infrastructure. There is a need to ensure that the IP-based infrastructure can support services with acceptable availability characteristics.

This means we need to re-evaluate what must be done to ensure service availability. We need to consider:

- What do users expect?
- What do service providers need to do to ensure the availability of their services?
- What influences service availability?

As services grow in complexity [2], further aspects of availability need to be covered. We see the need to provide a broader notion of availability that addresses these issues. In this article we discuss the issues and present a conceptual model for service availability designed to meet the challenges of securing availability of today's and future services. In section 2 we discuss the user and service provider viewpoints on availability. In section 3 we present and motivate the enhanced notion of service availability. A summary is provided in section 4.

ISSN 0085-7130 © Telenor ASA 2009

2 Viewpoints on Availability

2.1 User Requirements

User requirements depend on the type of service. For example, for online browsing services such as newspapers, timeouts and faults are tolerated. Even for online banking services, timeouts and faults are considered annoying, but are tolerated. However, loss or manipulation of data is not. Only the customer should have access to private information, in particular, account details. For VoIP services, the tolerance is also higher and subsequently, requirements more lax than the requirements specified for the public telephony service. Users are satisfied with the VoIP services available in Norway: quality is not an issue, even though more than half of the users surveyed reported experienced poor voice quality during calls, the VoIP service users reported that they are satisfied with the VoIP service [3].

Service availability requirements may vary over time for the same service. For example, for an online tax return submission service, users may not be overly concerned about availability during the weeks prior to the filing deadline; however, the service must be up on the final evening. For telemedicine applications and multimedia communications services, eg. the emergency telecommunications services, loss of availability may have catastrophic results. For example, if a person is critically ill or wounded and the emergency telephone service (eg., emergency call to 113) is not available, in a worst case situation, the consequences could be fatal. Similarly, it is important to prevent loss of availability of the underlying telecommunication services during a critical surgical operation for which the specialist surgeon at the central hospital is conducting surgery on a patient located remotely.

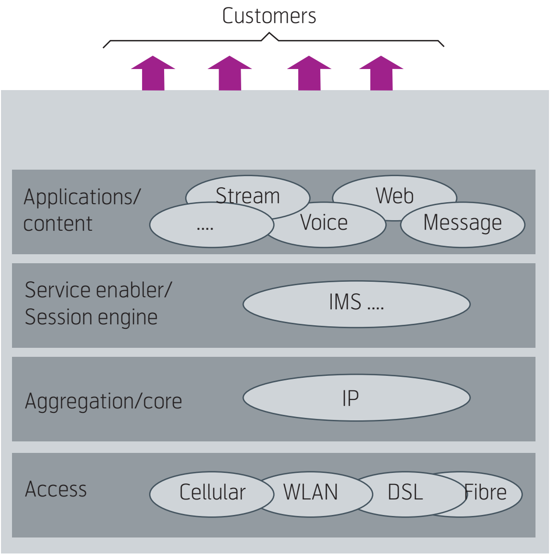

Figure 2 A range of services provided over variety of access networks with a common IP core

Clearly, for fulfilling user expectations; service availability depends more and more on the characteristics and requirements of the services themselves and the different requirements of certain users. How can we classify service availability with such a range of services, each with different requirements and constraints?

2.2 On Understanding Service Availability

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 [4]. This understanding has served well for describing and analyzing availability of services delivered in dedicated networks such as for voice services in the PSTN/ISDN. However, for describing service availability characteristics and analyzing availability of services in the vastly distributed environment in which IP-based services are deployed, an enhanced notion of availability is required.

For example, with the traditional understanding, even with a high mean rate of availability, failure that occurs during peak service request periods will result in high operational loss. One such scenario is a web service with 99.999% average availability that loses connectivity for 5 minutes during peak sales of concert tickets. In a multiple service provider environment, this scenario may be complicated by an attacker/competitor that deliberately attacks or redirects ticket requests from one web-based ticket sales service to the rival provider's service.

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 an important aspect of availability, but, does this sufficiently address how to ensure access for the authorised users? Do we merely over dimension the service and allow unauthorised users to access a service also? If so, can we be sure that unauthorised users do not get in the way of authorised users? In an ideal world we could imagine that this aspect could be ignored, however, there are plenty of examples to the contrary. For example, the Domain Name System (DNS) is vital to the Internet infrastructure. However, DNS in itself is not properly secured and as a result there have been several successful denial of service (DoS) attacks involving spoofing of DNS requests or altering DNS settings [5]. Spoofing of DNS requests can be prevented by a set of countermeasures (means) including requiring authentication of origin of the DNS request. Clearly, there is a need to prevent the denial of legitimate access to systems and services. That is, to focus on prohibiting unauthorised users from interrupting, or preventing users from accessing services.

We summarise the different concerns for analysing availability in Figure 3. The system view is well understood in the dependability field. The Service Availability Forum (SAF) is working on standardising middleware for the open interfaces between the service and system layers [6], as discussed in [7]. From the point of view of the user, a single service component is either available, or it is not. On the other hand, the user may find some level of service degradation acceptable for another type of service component. In both cases, the user may find that the overall QoE for the combined services may be acceptable. In our work on securing availability in service composition, we are analysing availability from the decomposed service viewpoint, according to requirements of the users. For an example of this, please see the next article in this *Telektronikk* volume.

3 Enhanced Service Availability Concept

3.1 General Motivation

The setting for the enhanced service availability concept is derived from the fields of dependability and security. As explained in [7], availability has been treated by the field of dependability and the field of security with different definitions and understandings of what availability is [8], [9], [10], [11].

The definition of availability used as a basis for the enhanced service availability concept is: The property of being accessible and usable on demand by an authorised entity [9], [11]. This definition captures the integral part of securing availability by ensuring access to authorised users while also addressing the aspect of a service being usable in addition to the traditional aspect of readiness for correct service.

The notion of service availability has been further refined using this definition as a basis, to include addressing the *exclusivity* aspect of ensuring that a service is provided to the authorised users *only* [12]. This aspect is important because a system must know how many users are expected to access a service at a given time as well as how long the users are expected to access the service. The number of users accessing at a given time and the session durations can be used to calculate the penetration and usage values. These values could be applied when dimensioning and as a basis for ensured performance levels. If the means to ensure that authorised users *only* are accessing a service is too weak, and unauthorised users are able to access a service, the service availability for authorised users may be affected.

As established in [8], availability is affected by means and threats. The conceptual model of dependability consists of three parts: the attributes of, the threats to, and the means by which dependability is attained [13] and provides a basis for the service availability conceptual model as motivated in [7]. 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 [14], [15].

This is because incidents resulting in loss of availability do not necessarily escalate into faults and therefore classification of means in terms of faults may become 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 ser-

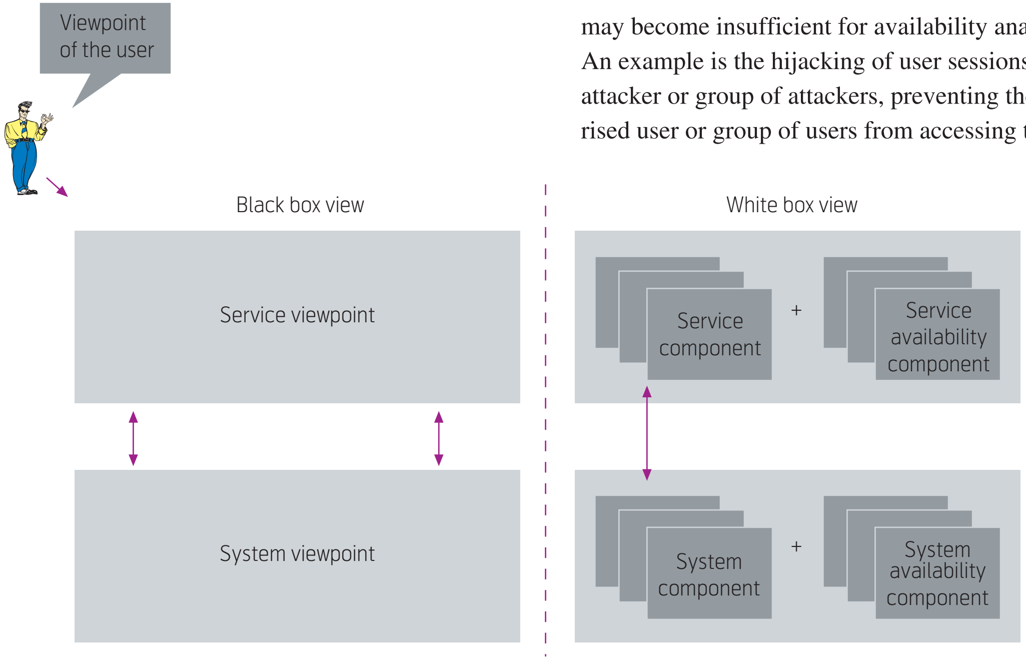

Figure 3 Viewpoints for analysing availability

vice. 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 [16] as an incident such as loss of confidentiality, integrity and/or availability. A fault is an example of an unwanted incident. The service availability conceptual model therefore classifies the means to achieve availability in terms of countering unwanted incidents.

In [17], the threats to dependability are defined as faults, errors and failures, and these are seen as a causal chain of threats to dependability:

fault –> error –> failure

This understanding of threats serves nicely in the dependability model, however, as service availability may be reduced, eg. by a denial of service attack without incurring a fault, error or failure, we apply the definition of threat, as defined in [11]: a threat is a potential cause of an unwanted event, which may result in harm to a system or organisation and its assets.

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. This means that a more fine grained measure of availability is needed than pure up or down.

It should be possible to describe various states of availability in order to specify the extent of which a reduction of service quality may be tolerated. The service availability metric should take into account, for example, measurement of different levels of degradation of services in order to analyze more closely how well user requirements are fulfilled, as well as measuring the ability to adequately provision a service to all of the authorised users requiring the service at a given moment. Such a metric should 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).

3.2 Enhanced Basic Notion

The enhanced notion of 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 usable when needed. Exclusivity involves ensuring that unauthorised users cannot interrupt, hijack, or prevent the authorised users from accessing a service. The focus is on preventing the denial of legitimate access to systems and services by prohibiting unauthorised users from interrupting, or preventing authorised users from accessing services. The aim is to ensure access to users while keeping unauthorised users out. Some of the means to achieve exclusivity address ensuring access for authorised users, and others address techniques for preventing unauthorised users from accessing or interrupting services, eg. by monitoring to discover unwanted traffic and blocking this traffic from unauthorised users.

Accessibility is defined as the quality of being at hand and usable when needed. We divide accessibility properties into three major areas: timeliness, correctness and usability. Timeliness is the ability of a service to perform its required functions and provide its required responses within specified time limits. Usability is concerned with the users' perception of the service, and the ease of use of the service. The measure of correctness of a service may differ widely between different kinds of services.

Consider an online payment 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 (eg. confirmation) or not (eg. timeout). (Over time this can be aggregated to percentages expressing how often one of the two kinds of responses will be given.)

These considerations motivate a notion of service degradation [18]. Service degradation can be defined as reduction of service accessibility. Analogous to accessibility, we divide service degradation into timeliness, usability and correctness degradation. These are mutually dependent on each other. For example,

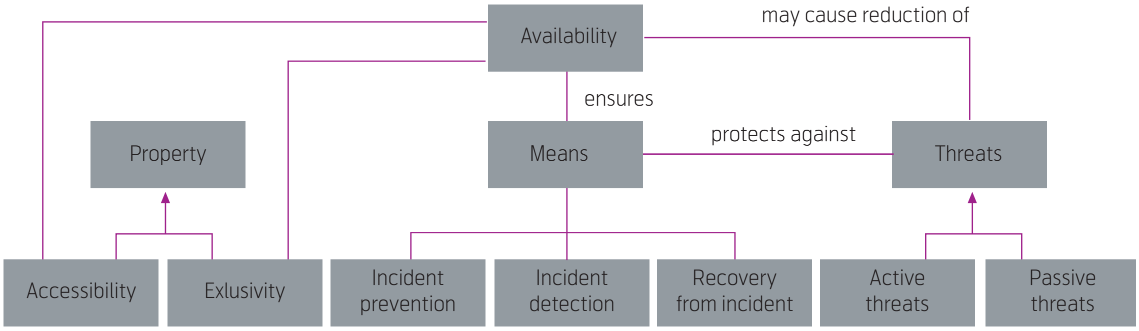

Figure 4 Conceptual model for service availability

graceful degradation in timeliness may be a way of avoiding correctness degradation if resources are limited, or the other way round.

In summary, the overall conceptual model can be depicted as in Figure 4. Availability is affected by means and threats. Means can ensure availability by protecting against threats. Threats may lead to unwanted incidents which may cause reduction of availability.

By means to ensure availability we address protection of the service from incidents leading to a loss of availability. Means are categorised into i) incident prevention: how to prevent incidents causing loss of availability (eg. access control, integrity protection ensuring graceful degradation); ii) incident detection: how to detect incidents leading to loss of availability (eg. traffic inspection, audit logs); and, iii) recovery from incident: the means to recover after an incident has lead to a loss of availability (eg. system adaptability, robustness, maintainability, redundancy).

Threats may originate on the inside (eg. inside attackers) or the outside (eg. 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. For the full motivation and explanation of the model, see [7].

3.3 Decomposing Availability

Based on the conceptual model, the availability of a service can be analyzed with respect to exclusivity and accessibility aspects. On an abstract level, a mathematical representation can be given as follows; 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 $X = (x_1, ..., x_n)$ as an n-tuple where x_i is a measure of an aspect of availability. These include behavioural, preventive and correctness aspects. By this we mean that x_i describes requirements for a particular availability aspect. The minimum requirement for each x_i must be satisfied in order to fulfil the total availability requirement X. Using the 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 describe the degree of accessibility and exclusivity that is sufficient for the user to be able to activate and use the service. The purpose of service availability metrics is to measure how well service availability requirements have been met.

For example, exclusivity metrics could measure how well the following requirements are met:

- 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 should be prohibited from accessing service s when user v is using the service.
- The number of intrusions at a given time t (eg. during a critical moment) 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 given time t.
- The probability that an unauthorised user obtains access to the service at a given time t.
- The probability that unauthorised user u obtains access to service s when user v is using the service.
- The number of intrusions at a given time t.

Similar requirements may be defined for accessibility.

4 Conclusion

We have presented a conceptual model that takes into account a broader spectrum of aspects that influence availability. In order to meet the demands of delivering services in the NGN environment, we define service availability as a composite notion consisting of exclusivity, ie. the property of being able to ensure access to authorised users only, and accessibility, ie. the property of being on hand and usable when needed as defined in [12]. The exclusivity aspect has been generally neglected, in this article we explain where it fits in. The accessibility aspect takes into account concerns of QoS, real time, and dependability. Availability is affected by means and threats. Means are classified in terms of countering unwanted incidents. Means ensure availability and reduce threats. This classification of availability (properties, threats, and means) provides an operational approach that can be applied to service engineering. Securing availability is influenced by many factors. These must be taken into account in order to provision available services. With this model, it is possible to address requirements in a flexible manner, in order to address the different aspects of different services.

The next article demonstrates how the model can be used to analyze compound/composite services. Single provider and multi-provider configurations are exemplified.

Acknowledgements

The research on which this article reports has been funded by the Research council of Norway through the project SARDAS (152952/431).

References

1 ITU. *Next Generation Networks – Frameworks and functional architecture models – General principles and general reference model for Next Generation Networks*. International Telecommunication Union, 2004. (ITU-T Recommendation Y.2011)

2 Arbaugh, W A, Fithen, W L, McHugh, J. Windows of vulnerability: A case study analysis. *IEEE Computer*, 33 (12), 52-59, 2000.

3 *Statistics, 2006 and 2007*. July, 2008 [online] – URL: http://www.npt.no/portal/page/portal/PG_NPTNO_NO/PAG-NPT_NO_HOME/PAG_PUBLIKASJONER_TEKST?p_d_i=-121&p_d_c =&p_d_v=48951. Norwegian Post and Telecommunication Authority.

4 Ross, S M. *Inroduction to probability models*, 6th ed. Academic Press, 1997.

5 Goodin, D. *Users redirected to rogue site*. July, 2008 [online] – URL : http://www.theregister.co.uk/2008/05/29/comcast_domain_hijacked/.

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

7 Rossebø, J E Y, Lund, M S, Husa, K E, Refsdal, A. *A conceptual model for service availability*. Department of Informatics, University of Oslo, 2006. (Research report 337)

8 Laprie, J C (Ed). *Dependability: Basic Concepts and Terminology*. Springer, 1992.

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

10 ISO. *Information technology – Code of practice for information security management*. International Standards Organization, 2000. (ISO/IEC 17799)

11 ISO. *Information technology – Security techniques – Guidelines for the management of IT security*. International Standards Organization, 2001. (ISO/IEC 13335)

12 Rossebø, J E Y, Lund, M S, Husa, K E, Refsdal, A. A conceptual model for service availability. *Quality of Protection: Security Measurements and Metrics*, 23, 2006.

13 Avižienis, A, Laprie, J-C, Randell, B. Fundamental concepts of dependability. **In**: *Third Information Survivability Workshop (ISW)*, 2000.

14 den Braber, F, Lund, M S, Stølen, K, Vraalsen, F. Integrating security in the development process with UML. **In**: *Encyclopedia of Information Science and Technology*. Idea Group, 2005, 1560-1566.

15 Lund, M S, den Braber, F, Stølen, K. Maintaining results from security assessments. **In**: Proc. of the 7th European Conference on Software Maintenance and Reengineering (CSMR). *IEEE Computer Society*, 2003, 341-350.

16 *Risk Management*. Standards Australia, 1999. (AS/NZS 4360:1999)

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

18 Meyer, J F. Performability evaluations: Where it is and what lies ahead. **In**: Proc. of the International Computer Performance and Dependability Symposium. *IEEE Computer Society*, 1995, 334-343.

Judith E.Y. Rossebø is a Senior Research Scientist at Telenor R&I. Prior to joining Telenor in October 2000, she worked three years as a systems engineer at Alcatel Telecom Norway and one year as an assistant professor teaching Mathematics at the University of Tromsø. At Alcatel she worked with IN, and dimensioning, performance, dependability and traffic control in telecommunication networks. She received a Cand.Scient. degree in Mathematics from the University of Oslo in 1994 and is completing a PhD at Norwegian University of Science and Technology (NTNU), Department of Telematics, in the SARDAS project. Since January 2003 she has been the Chairman of ETSI TISPAN WG7 Security. Her research interests include security in general; security issues in multimedia communications services, and in particular securing availability of services.

judith.rossebo@telenor.com