

On Understanding Availability of Services Based on IP Multimedia Subsystem

Judith E. Y. Rossebø, Arlene Pearce, and Terje Jensen

Abstract—Availability has commonly been considered as an atomic property, indicating the average amount of time a system is working as planned. This, however, should be further detailed considering that more compound services are being deployed. Expectedly, such situations are met when federating services based on the IP Multimedia Subsystem. Then, certain parts of the service may work to the complete satisfaction of the user, while other parts are not quite up to the planned behaviour. It is usually also included as the planned behaviour that unauthorized users should not get access to the services or relevant data. Hence, two essential characteristics of availability are accessibility and exclusivity. This paper presents and discusses a conceptual model for service availability intended to capture characteristics of these service types and demonstrates how the model can be used to analyze compound/composite services. Single provider and multi-provider configurations are exemplified.

Index Terms—Availability concept, federated services, IP Multimedia Subsystem, Quality of Experience.

I. INTRODUCTION

AVAILABILITY is a central characteristic in service delivery. In fact, in today's society available telecom services are pivotal for several enterprises, e.g., banking and financial services. The concept of service availability, however, goes beyond telecom. Often, consequences are serious if even parts of telecom or computer systems are unavailable when their services are needed.

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.

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 [1]. Accompanying this interpretation, failure reporting procedures have also been described, e.g. [2] for Public Switched Telephony Network, PSTN. This understanding has served well for describing and analyzing availability

J. E. Y. Rossebø is with Telenor Research and Innovation, NO-1331 Fornebu, Norway. She is also affiliated with The Norwegian University of Science and Technology (NTNU) (e-mail: Judith.rossebø@telenor.com).

A. Pearce is with Telenor Research and Innovation, NO-1331 Fornebu, Norway. She is also pursuing MSc degree at the University of Oslo (e-mail: pearce@pvv.ntnu.no).

T. Jensen is with Telenor Research and Innovation, NO-1331 Fornebu, Norway, and the Norwegian University of Science and Technology (NTNU) (e-mail: terje.jensen1@telenor.com).

This work has partially been funded by the Research council of Norway project SARDAS (152952/431) and partially within the Eureka project Mobile Fixed Convergence in Multiaccess Environment, Mobicome.

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.

Considering the emerging range of IP-based services being delivered in public and private networks today, several challenges follow from the traditional understanding of availability. This paper addresses two challenges. The first challenge is that 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. Such bursty behaviour patterns could be seen for several of the services [3]. The second challenge is that when presented with a set of service components, a user may have different expectations of quality for each component. One example is a buddy list with presence information fed by an IP Multimedia Subsystem, IMS. Different categories of buddies may be defined, say work-related and leisure-related. A user may tolerate lower quality weather forecast or tv-guide services she typically uses in leisure-mode while she may expect near perfection from a stock-ticker service used in work-mode. Similarly, reliable presence information pertaining to her colleagues/employees may be more important to her than reliable presence information pertaining to leisure-related contacts.

In the multi-application environment implied by IMS, several features may contribute to the overall user experience. For example, the user interface may collect parts of presence information, location-dependent data, calendar tasks, and other service components. Different parts of the user interface may be updated by different servers. Hence, the user experience is collated from different sources. Moreover, the different parts of the user interface may have different weights in the experience depending on the user tasks.

We focus on the problem of considering the variability of the contributions of the different parts involved in the services to the overall availability of a service. Issues are the adequacy of the current availability concept, the definition of availability and the availability management. Quality of Experience, QoE, should then be related to these issues as addressed by this paper.

This paper addresses these issues. The service availability concept model motivated and introduced in [4] is presented and exemplified by a set of cases. As services grow in com-

plexity [5], further aspects of availability need to be covered. This paper presents and elaborates on a conceptual model for service availability providing a case study to demonstrate the applicability of the model to service provisioning in a distributed IMS service environment.

Sect. II provides a brief introduction to the enhanced service availability concept. An overview of IMS follows in Sect. III. Sect. IV exemplifies how the enhanced service availability concept can be applied for a federated presence implemented on IMS.

II. ENHANCED SERVICE AVAILABILITY CONCEPT

A. General Motivation

The setting for the enhanced service availability concept is derived from the fields of dependability and security. As explained in [6], availability has been treated by the field of dependability and the field of security with different definitions and understandings of what availability is [7], [8], [9], [10].

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 authorized entity [8], [10]. 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 authorized users *only* [4]. 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 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 basis for ensured performance levels. If the means to ensure that authorized users *only* are accessing a service is too weak, and unauthorized users are able to access a service, the service availability for authorized users may be affected.

As established in [7], 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 [11] and provides a basis for the service availability conceptual model as motivated in [6]. 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 [12], [13].

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 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 [14] 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 [15], the threats to dependability are defined as faults, errors and failures, and these are seen as a causal chain of threats to dependability:

fault \rightarrow error \rightarrow failure

This understanding of threats serves nicely in the dependability model, however, as service availability may be reduced e.g. by a denial of service attack without incurring a fault, error or failure, we apply the definition of threat, as defined in [10]: 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 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). In section IV below, we provide a set of parameters for measuring the availability of an IMS presence service.

B. 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 useable 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, e.g. 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.

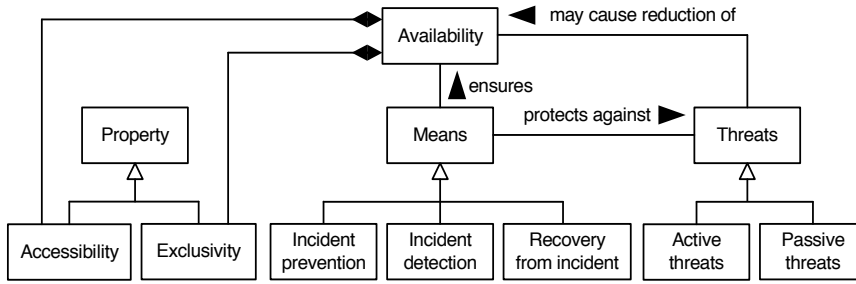


Fig. 1. Conceptual model for service availability

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 (e.g. confirmation) or not (e.g. 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 [16]. 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, graceful degradation in timeliness may be a way of avoiding correctness degradation if resources are limited, or the other way around.

In summary, the overall conceptual model can be depicted as in Fig. 1 (illustrated in UML 2.x format [17]). 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. We have categorized the means into i) incident prevention: how to prevent incidents causing loss of availability (e.g. access control, integrity protection ensuring graceful degradation); ii) incident detection: how to detect incidents leading to loss of availability (e.g. traffic inspection, audit logs); and, iii) recovery from incident: the means to recover after an incident has lead to a loss of availability (e.g. system adaptability, robustness, maintainability, redundancy).

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. For the full motivation and explanation of the model, see [6].

C. 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 $X = (x_1, \dots, x_n)$ 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 of intrusions at a given time t (e.g. 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 of intrusions at a given time t .

Similar requirements may be defined for accessibility.

III. IP MULTIMEDIA SUBSYSTEM (IMS)

IMS has been promoted by several international bodies as a future platform for providing services. It is access agnostic in the sense that services could be provided over any access type and to any device. That is, as long as the device is capable of supporting the proper client behaviour.

A. Layered Architecture

A layered architecture has been applied for defining IMS, see Fig. 2. In the core part, we find common session control and common user data. In IMS terms these are referred to as Call Session Control Function (CSCF) and Home Subscriber Server (HSS), respectively. There are several types of CSCF supporting roaming users, interconnecting between domains and emergency sessions, although these are not depicted in Fig. 2.

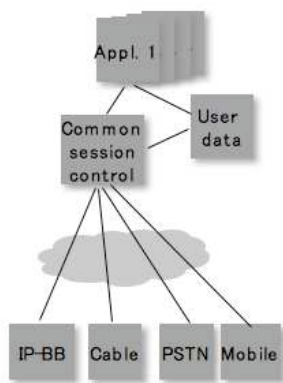


Fig. 2. Layered architecture of IMS

HSS stores user identities and user profiles. Both private and public identities are defined and a user/individual can have a set of user profiles. A profile explains which applications are to be invoked and how services are to be executed.

A range of applications can reside in the common IMS core. Examples of application types are group list managers, Centrex, location, handover support (WLAN 2G/3G).

Main protocols are SIP-based and DIAMETER-based; The former for service/session control and the latter for data access and charging. IMS has, however, defined additional parameters and attributes compared with the original IETF SIP.

The client side is not directly illustrated in Fig. 2. For IP-based terminals, an IMS client would be implemented in the terminal. Then, there will be a session running between the terminal and the CSCF. For traditional circuit-switched networks, such as PSTN and GSM-CS, the client could be said to be implemented in the switching control. That is, the media gateway controller would run session control with the CSCF. In this manner, an IMS installation could be considered to fill similar roles as Intelligent Networks do today.

As shown in Fig. 2, the effect is that multiple applications can be accessed and used through different access types. Or, in other words, it allows a seamless experience over different access and terminal types.

There are several ways in which the key components SIP Application Server, User Agent (UA), CSCF and Home Subscriber Server (HSS) can communicate with each other. This provides flexibility when designing architecture for service creation and orchestration. Fig. 3 shows existing interfaces that can be utilised in designing a new architecture.

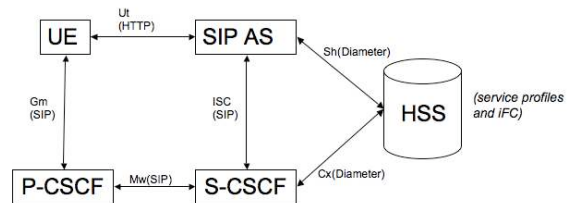


Fig. 3. Interfaces involved in service composition

The SIP AS is an integral part of the IMS specification. Note that it has direct access to the users profiles in the HSS. User profiles can in turn have service profiles. The SIP AS can also take several different roles in the SIP session initiation chain and these roles can be leveraged in service orchestration.

As a set of applications may be invoked for a session, there could be a need for allowing these applications to inter-play. For example, the Centrex application may want to know the location of a user to decide upon further service execution. One example is that different rules should be followed if the user is in the office area compared to when the user is in the car. For this purpose, a Service Capability Interaction Manager has been identified. However, it has not been described in detail, allowing for several implementation options. In particular, there are discussions on whether the Service capability Interaction Manager should be implemented in a centralized, distributed or hybrid manner.

B. Presence Service

One application that we analyze with respect to the service availability concept model presented above, is the presence service. Basically, the presence server collects information about a set of users and presents this information to pre-defined users. Schematically, it can be depicted as in Fig. 5. The users from whom presence information is requested are called Presentities. Presence sources are defined as nodes reporting the presence information. Examples of presence sources are clients in handsets, mail/calendar servers, network elements and indications given through web portals.

This information is collected by the presence server and reported to the pre-defined watchers. Again, different watcher types can be defined, such as clients on handsets and other application servers.

As also shown in Fig. 5, not only users can be reported through this mechanism, but also other types of items, such as stock prices that a user has subscribed to monitor. There are also providers utilising the same mechanism for presenting advertisements, reminders or other offers.

As a result, the list of items provided by the presence server can be grouped into sets that have different expectations as seen by a user. These may also vary during the day, for example as some items may be more related to work, while others are more related to leisure, family or social groups.

The different presence items may be updated from different sources. This means that several different service providers as well as others may be involved. For example, a user may set up buddies that are subscribers of other service providers. This implies that the different providers must interact, and cooperate. An immediate case is to incorporate Facebook friends into an IMS based buddy list as illustrated in Fig. 4.

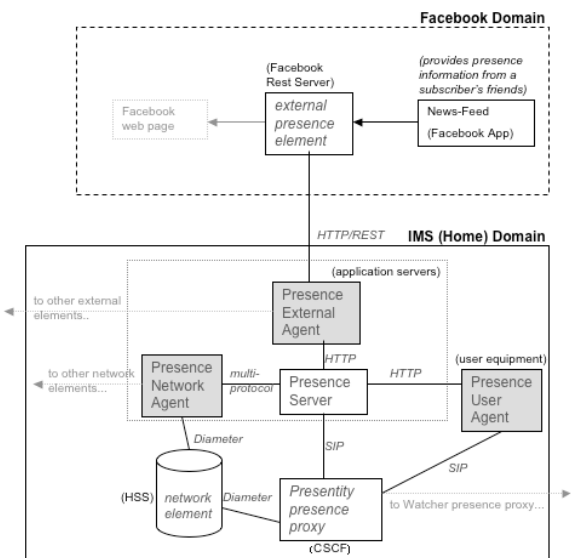


Fig. 4. Facebook as External Presence Element

Presence information, as defined by [18], conveys the ability and willingness of a user to communicate across a set of devices. A presence service is a system that accepts, stores and distributes presence information to interested parties, called watchers. A presence protocol is a protocol for providing a presence services over any network.

The presence server is located in the home network of the user that the presence information is relating to (called presentity). It would commonly include both logic and data storage. Key capabilities are collecting, composing and filtering presence information. Filtering could be used when deciding which presence parameters should be presented to an actual watcher application. This could also be used when only the parameters that have changed since the previous update should be forwarded. Filters can define which tuples are watched

(e.g. all that has contact address equal to tel:user@domain), attributes to be forwarded to a watcher and triggers when notifications should be sent.

Watcher information shall be collected by the presence server, which allows a presentity to obtain that information. Any presence information that the presence service is not able to interpret shall be handled in a transparent manner.

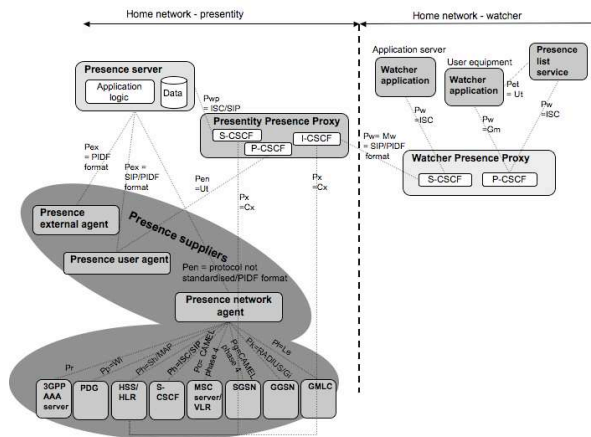


Fig. 5. Functional components for presence (adapted from 3GPP TS 23.141 [19]).

Subscription authorization policy shall be provided by the presence server. This tells which watchers that are allowed to subscribe to which of a presentity's presence information.

A number of suppliers of presence information can be relevant:

- Presence network agent: providing presence information from network elements. A range of network elements and corresponding protocols are shown in Fig. 5. Examples of presence information obtained are i) from GGSN PDP context activation, de-activation, ii) from SGSN attached, not reachable for paging, detached, routing area update, iii) from MSC server attach, detach, location area update, call setup, idle, connected, busy, etc., iv) from 3GPP AAA server WLAN UE attaching/detaching, tunnel establishment/removal.
- Presence user agent: providing presence information on behalf of a principal. This may be located in the terminal or in the network (e.g. when the terminal does not have a proper user client installed). In case it is located in the network, it must be within the presentity's home network.
- Presence external agent: providing presence information by elements outside the providers network. This takes care of any interworking and security issues and resolves location of the presence server. Examples of information supplied by the presence external agent include i) third party services (calendar applications, corporate systems, etc.), ii) internet presence services, iii) other presence services.

The presence proxy can be used to locate the presence server (will make a lookup in the HSS) that is to be used for a given user/presentity.

The watcher entities are divided into watcher presence proxy and watcher applications. The watcher presence proxy carries out functions such as authentication of watchers.

Watcher applications can be located in i) user equipment, ii) application server, and, iii) external to providers domain. The same interface is used both for requesting monitoring and for fetching information. In both cases all or a subset of a presentity's information can be transferred (e.g. referring to a given filter, only parameters changed since last notification, etc.).

A presence list server (an SIP server) can also be involved keeping information regarding grouped lists of watched presentities. This enables a watcher application to subscribe to presence of a group of users/presentities by a single SUBSCRIBE transaction. A presence list server also stores and manages filters associated to presentities in the presence list. Filter shall be attached to individual SUBSCRIBE transactions.

A watcher application sends a SIP SUBSCRIBE to Event:presence addressed to the presentity's SIP URL to subscribe or fetch presentity's presence information. This SUBSCRIBE request will be handled by the IMS core elements reaching the presence server. The presence document is provided by the presence server to the watcher application using SIP NOTIFY (see Fig. 6). The SIP NOTIFY may be triggered by a change of the presentity's status, notified by any of the presence suppliers via the corresponding interfaces and message types.

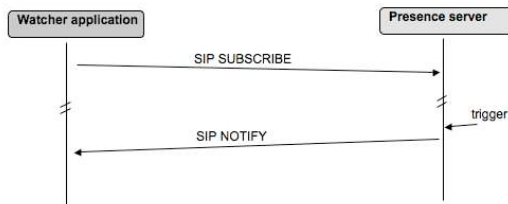


Fig. 6. Schematic message flow between a watcher subscribing to updates of presence information

Both of the accessibility and the exclusivity aspects of service availability discussed in Sect. II are related to the presence item list. For each of the different groups in the item list, the degree of exclusivity achieved and accessibility achieved may differ. For example, during leisure time, buddies related to work may not require to receive updated information on a user's whereabouts. On the other hand, during office hours, it may be considered crucial by the user to receive updated information on other users (buddies) whereabouts.

Level of detail may also differ for the different items. For example, for certain buddies, the user is allowed to see their locations, while for others the locations must not be displayed.

Again, these rights may vary over time.

Hence, the user's desktop can be looked upon as consisting of a set of fields. For each of the fields the user has a certain expectation of correct behaviour, both in terms of that the fields content should be correct and that it should not be revealed to other non-qualified parties. On a conceptual level, each of the items is expected to be up to date with respect to specific timeliness requirements. The overall user's experience might then be divided into expectations for each of the fields. This leads to decomposition into fields with separate availability requirements.

A time stamp field in the message format may be used as a rudimentary security mechanism to prevent replay attacks e.g. launched for the purpose of capturing user credentials for unauthorized access to a service. For example, tuples whose time stamp are older than the time stamp of the most recently received presence document should be discarded.

Attributes shall be mapped to separate tuples that have unique identifiers. In case different attribute values should be shown to different watchers, a set of tuples must be created that contain the same attribute. The subscription authorisation policies give the association of tuples to different watcher groups, that is, which watchers can access which presentity information.

Subscription authorisation lists can be divided into the following categories:

- 1) blocking; watchers not allowed to access any presence information related to the presentity.
- 2) personal; explicitly identifying watchers
- 3) general; groups of watchers who are not necessarily known by the presentity, e.g. all watchers

The list categories are evaluated in the order

- 1) → 2) → 3)

IV. CASE – AVAILABILITY ASPECTS OF THE ENHANCED PRESENCE SERVICE

The enhanced presence service based on IMS is assumed to combine a set of different features, such as location information as well as advertisements, see Fig. 7. There is also support to include buddies that are subscribers of other domains, e.g., other service providers.

A. Single-Provider Case

The first case is depicted in Fig. 8. Here, it is assumed that all users are within the same service provider domain. It is also assumed that group lists are stored by the presence server.

In a straight-forward manner, for the overall service to work, all components have to be in operation. For the configuration shown in Fig. 8, user 2 and user 3 are attached to the same network element. Hence, it suffices that this element is in operation. A block diagram could then be set up in order to analyse the service availability with respect to accessibility.

Suppose, however, that not all users are of equal importance for user N . Then, accessibility estimates can be made for each user (i.e. items) on the presence list, e.g. A_i for user i . Average

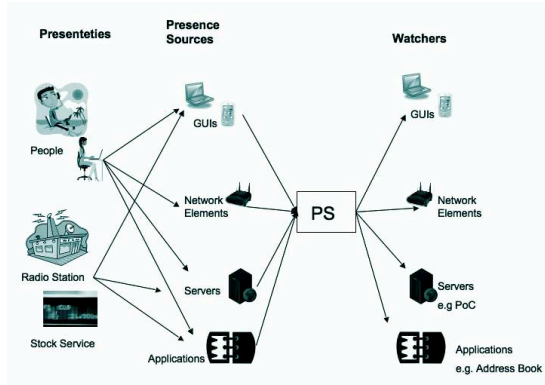


Fig. 7. Presence server conveying information about users to defined watchers

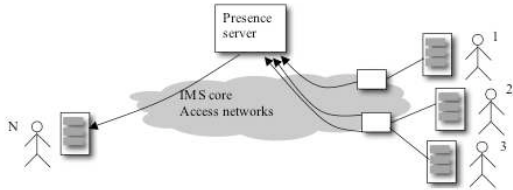


Fig. 8. Case I: user with 3 buddies for presence listing

accessibility is then found as $\sum_i (A_i \cdot \alpha_i) / \sum_i \alpha_i$, where α_i , is the level of importance associated with user i .

B. Multi-Provider Case

Providing different levels when all users are within the same provider domain seems contrived. When the different users are in different domains, different levels may become more relevant. An example is depicted in Fig. 9.

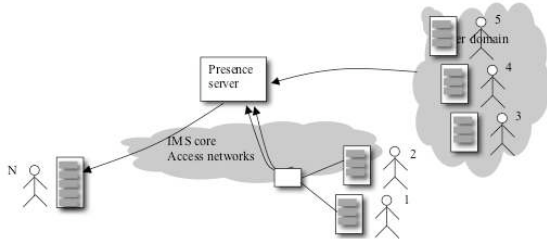


Fig. 9. Case II: user with buddies in different provider domains

Again, a block diagram could be put up for each of the items in the list. For this configuration, however, certain details regarding implementation of presence services in other domains would likely not be available. This implies that accessibility levels would then be part of the Service Level Agreements (SLAs) established between providers. It might also be more suitable to state different levels towards user N

for the different user groups in this configuration. However, this will probably depend on the SLA terms. Similar expression as above follows, although an index j indicates the group of buddies: $A_j = \sum_i (A_{ij} \cdot \alpha_{ij}) / \sum_i \alpha_{ij}$

In some cases, there may not be any pre-established SLA between the providers. It then becomes a business risk evaluation whether a service provider wants to state any performance levels in the service description to user N . Potentially, there may be differentiated levels for the different groups, j .

Level of importance related to a user, α_{ij} , may vary depending on the role of the user, N . For example, during working hours, it is more important to follow work colleagues, for example, involved in the same project, than outside working hours. As projects come and go, the colleagues involved will also vary, requiring that this information is easily updated frequently.

C. Parameters Included in Enhanced Presence

The parameters given for a presence item may include nickname, mode, location, as well as others. Typically, location could be given with different level of granularity. In some cases, e.g. during roaming, the location may also be unknown. On the other hand, there are certain usages where location is considered as very important, such as following children in kindergarten and following some mental patients.

For user N , different levels of importance could then be attached to the different parameters. These importance weights may also differ for the different presence items as shown in Fig. 10.

As these parameters may be pushed or pulled from different sources, different response or delivery times would result. In some respects, this is similar to the design of a web page consisting of a set of objects. In order to improve QoS and network performance, presence parameters should be delivered in appropriate sub-groups. That is, waiting for the last presence parameters before sending an update to the users buddy list (watcher) would likely result in too long response times and degraded QoE.

This sub-grouping is particularly a critical aspect when information is collected from vastly different sources, some residing in semi-real-time environments while others within best-effort environment. In effect, this balances the different aspects of availability (correctness, timeliness and usability) as described in Sect. II.

D. Use of Presence Items for Other Purposes

Having configured means for controlling presence items on a terminal, one could utilise this for other purposes as well, such as advertising and time-related special offers. One such implementation has been tested in a real user environment in Finland, in the SmartRotuuaari project [20]. The service implemented included highly personalized direct marketing to customers mobile phones.

When users accept such commercial activities, a service provider would then likely have an agreement with a of companies for delivering the advertisements/offers. Then, there

will also be requirements on accessibility for providing this feature. Again, block diagrams could be made assisting the accessibility evaluation.

E. Exclusivity

From the outset, overall exclusivity could be analysed in a similar manner as for accessibility. That is, either looking at the overall average or looking at each presence item individually. However, the analysis approaches would likely differ as there are often other types of threats affecting the exclusivity aspect.

The aim with respect to the exclusivity aspect is to ensure that authorised users only should have access to presence items that they subscribe to. Allowing unauthorised users to access presence items may have a negative effect on accessibility, but also, may be in violation of privacy directives in the jurisdiction.

Commonly, for the presence service there will be stricter requirements regarding exclusivity than for accessibility. This is mainly due to privacy aspects, avoiding any third party being able to follow a users actions. However, it is also important to ensure that the authorised users are not prevented or interrupted from accessing the presence items. Activity initiated by unauthorised users can adversely affect the accessibility aspects.

It may also be important to measure whether a rogue service provider is interfering with the availability of presence items delivered from another service provider.

It would also be a business decision on whether to provide exclusivity measures for the average or for groups of items. Some statistical aggregation measures could also be defined. How the exclusivity aspects are measured may also depend on the means that are deployed for ensuring exclusivity.

F. Threats and Corresponding Means

Considering the different servers, network elements, protocols, data, etc. that are involved, various threats would be relevant. Examples of threat agents are malicious users, rogue service providers, unauthorised users masquerading as authorised users to obtain information on another users whereabouts. The vulnerability of the presence service to distributed denial of service attacks should be evaluated.

Hence, combinations of means to address the threats, covering protection, detection and recovery will be required. Preventative means will involve access control measures, for ensuring that authorised users only have access to the items. Monitoring of activity to detect malicious user behaviour could also be deployed. Essentially, preventative mechanisms will try to eliminate the possibility of attacks by threat agents or to enable the presence service nodes to be able to endure attacks without denying service access to authorised users. Detection and recovery will involve detecting attacks on nodes of the service, or against specific users, and responding immediately to restrict impact.

ETSI TISPAN¹ has developed a threat vulnerability and risk assessment (eTVRA) method and tool that may be used to analyse the presence service [22]. Using the eTVRA method and tool, the threats to availability of the presence service can be analyzed and a set of recommended countermeasures can be identified that when implemented will reduce the overall risk.

G. User Experience

As mentioned above, expectations from users may vary for the different presence items and the different parameters for each item. Moreover, these weights may vary over time. Schematically, the actual user experience could be estimated by multiplying the weights and the actual obtained accessibility and exclusivity measures as depicted in Fig. 10.

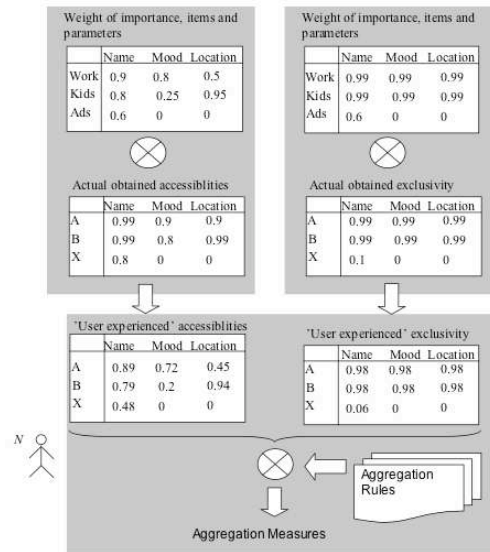


Fig. 10. Schematic procedure for estimating actual user experience (for illustration only)

Rules for aggregating experience parameters would likely be adapted to different purposes, such as statistical performance levels stated in SLAs, for Key Performance Indicators, and so forth.

It still remains to figure out how aggregated measures should be calculated and how many measures should be defined. Theoretically, based on questionnaires, a single parameters could suffice, say providing an estimate for the question, "On a scale of 1 – 0, how satisfied are you with the presence service?"

¹The European Telecommunications Standards Institute (ETSI) is an independent, non-profit organization, with a mandate from the European Commission to undertake standardisation of Information and Communication Technologies (ICT) within Europe alongside its partners CEN/ISSS (European Committee for Standardization/Information Society Standardization System) and CENELEC (European Committee for Electrotechnical Standardization). ETSI TISPAN [21] is taking a leading role in developing standards for fixed mobile convergence (FMC).

Additional requests for following the performance levels are also needed. One example is monitoring key quality indicators related to selected aspects of the business. Another example is accessing performance levels when corresponding parameters are specified in the SLA between actors.

V. CONCLUSION

This paper elaborates and exemplifies a conceptual model for availability. A key point is to include both the accessibility and the exclusivity aspects of the service availability measure. Hence, only the authorized users should be ensured access to the service, and with the proper service levels. So far, it seems that exclusivity is an aspect of availability that has rarely been included in the literature. However, the concept presented here shows where exclusivity fits in with an enhanced notion of service availability.

The enhanced notion seems even more important when considering collaboration between providers, and also between different roles of the same user. Proper service composition, also referred to as orchestration, then becomes even more important and challenging.

A schematic example is provided through the enhanced presence service realised by IMS. Considering the federated nature of the presence service, a range of challenging aspects need to be addressed, including differentiation of presence items and parameters for an item while also handling multiple sources of presence data.

Among items for potential further work is the task of conducting user experiments to estimate QoE related to different items of the presence service. Other items are to assess weights and further details for the schematic procedure outlined in Fig. 10.

REFERENCES

- [1] S. M. Ross, *Introduction to probability models*, 6th ed. Academic Press, 1997.
- [2] 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.
- [3] 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.
- [4] J. E. Y. Rossebø, M. S. Lund, K. E. Husa, and A. Refsdal, "A conceptual model for service availability," *Quality of Protection: Security Measurements and Metrics*, vol. 23, 2006.
- [5] W. A. Arbaugh, W. L. Fithen, and J. McHugh, "Windows of vulnerability: A case study analysis," *IEEE Computer*, vol. 33, no. 12, pp. 52–59, 2000.
- [6] J. E. Y. Rossebø, M. S. Lund, K. E. Husa, and A. Refsdal, "A conceptual model for service availability," Research report 337, Department of Informatics, University of Oslo, 2006.
- [7] J. C. Laprie, Ed., *Dependability: Basic Concepts and Terminology*. Springer, 1992.
- [8] *ISO 7498-2, Information Processing Systems – Interconnection Reference Model – Part 2: Security Architecture*, International Standards Organization, 1989.
- [9] *ISO/IEC 17799, Information technology – Code of practice for information security management*, International Standards Organization, 2000.
- [10] *ISO/IEC 13335, Information technology – Security techniques – Guidelines for the management of IT security*, International Standards Organization, 2001.
- [11] A. Avižienis, J.-C. Laprie, and B. Randell, "Fundamental concepts of dependability," in *Third Information Survivability Workshop (ISW)*, 2000.
- [12] F. den Braber, M. S. Lund, K. Stølen, and F. Vraalsen, "Integrating security in the development process with UML," in *Encyclopedia of Information Science and Technology*. Idea Group, 2005, pp. 1560–1566.
- [13] M. S. Lund, F. den Braber, and K. Stølen, "Maintaining results from security assessments," in *Proc. of the 7th European Conference on Software Maintenance and Reengineering (CSMR)*. IEEE Computer Society, 2003, pp. 341–350.
- [14] *AS/NZS 4360:1999, Risk Management*, Standards Australia, 1999.
- [15] 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*, vol. 1, no. 1, pp. 11–33, 2004.
- [16] J. F. Meyer, "Performability evaluations: Where it is and what lies ahead," in *Proc. of the International Computer Performance and Dependability Symposium*. IEEE Computer Society, 1995, pp. 334–343.
- [17] *UML 2.0 Superstructure Specification, formal/05-07-04*, Object Management Group, 2006.
- [18] J. Rosenberg, *A Presence Event Package for the Session Initiation Protocol SIP*, RFC 3856, 2004.
- [19] *Presence Service; Architecture and functional description, Stage 2*, Third Generation Partnership Project, Technical Specification Universal Mobile Telecommunications System (UMTS), 3GPP, TS 23.141 V 7.3.0 (2007-10), Release 7, 2007.
- [20] T. Ojala, J. Korhonen, M. Aittola, M. Ollila, T. Koivumäki, J. Tähtinen, and H. Karjaluoto, "SmartRotuuri context-aware mobile multimedia services," in *Proc. 2nd International Conference on Mobile and Ubiquitous Multimedia*. Washington, DC, USA: IEEE Computer Society, 2003, pp. 9–18.
- [21] European Telecommunication Standardisation Institute, "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN)," http://portal.etsi.org/tispan/TISPAN_ToR.asp, 2008.
- [22] J. E. Y. Rossebø, S. Cadzow, and P. Sijben, "eTVRA, a threat, vulnerability and risk assessment method and tool for eEurope," in *ARES*. IEEE Computer Society, 2007, pp. 925–933.