Encyclopedia of Internet Technologies and Applications

Mario Freire
University of Beira Interior, Portugal

Manuela Pereira
University of Beira Interior, Portugal

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Acquisitions Editor: Kristin Klinger
Development Editor: Kristin Roth
Senior Managing Editor: Jennifer Neidig
Managing Editor: Sara Reed
Copy Editor: Larissa Vinci and Mike Goldberg
Typesetter: Amanda Appicello and Jeffrey Ash
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanonline.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of Internet technologies and applications / Mario Freire and Manuela Pereira, editors.
p. cm.
Summary: "This book is the single source for information on the world's greatest network, and provides a wealth of information for the average Internet consumer, as well as for experts in the field of networking and Internet technologies. It provides the most thorough examination of Internet technologies and applications for researchers in a variety of related fields"--Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-59140-993-9 (hardcover) -- ISBN 978-1-59140-994-6 (ebook)
1. Internet--Encyclopedias. I. Freire, Mário Marques, 1969- II. Pereira, Manuela.
TK5105.875.I57E476 2007
004.67'803--dc22
2007024949

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Quality of Service and Service Level Agreements

Christos Bouras
Research Academic Computer Technology Institute and University of Patras, Greece

Apostolos Gkamas
Research Academic Computer Technology Institute and University of Patras, Greece

Dimitris Primpas
Research Academic Computer Technology Institute and University of Patras, Greece

Kostas Stamos
Research Academic Computer Technology Institute and University of Patras, Greece

INTRODUCTION

In order for advanced applications in modern computer networks to function satisfactorily, there is often the need for a guaranteed network performance and guaranteed values for several network parameters. When the provisioning and usage of network services is agreed, relevant specifications for the level of the services are also defined.

The definition of suitable specification for the offered services is very important when these services are provided by external parties. Taking Internet connectivity as an example of a provided service, it is very important that the specifications define the minimum guaranteed capacity of the connection (which affects the speed of the network), the availability of the connection, and other parameters that affect the response time of network applications, so that users do not have to face constant and annoying problems. By specifying the required level of service, the client of the service specifies the required results, but does not specify the way the provider is going to implement the service. A client who tries to specify the way that the provider is going to implement the service cancels the advantage of the provider's specialized technical knowledge.

The specifications guarantee the responsibility on the part of the provider and define the fee for the provided service.

In cases where the buyer assigns the operation of some activities to an external provider (outsourcing), the only way to guarantee a satisfying level of service quality for the buyer is to define the required level of service and then constantly measure the performance of the provider (with regard to the service) in order to decide whether the service is adequately provided and to guarantee that the buyer receives the level of service that is being paid for.

Specifications also affect the cost of the service. Clients who desire very high quality levels create greater resource requirements for the provider and therefore increase the cost of the service. Therefore, proper definition of specifications is vital.

BACKGROUND

The scope of service level agreements is wider than the provisioning of network related services by external providers to clients. It can generally be applied to any service provided by an external party. The term has nevertheless gained acceptance in the Information Technology (IT) field because of the nature of the provided services (Bouras et al., 2002, 2004; Heinanen et al., 1999; Jacobson et al., 1999; Xiao et al., 1999). More specifically, the complexity of the provided services, the difficulty in finding acceptable measurement procedures, and the novel nature of many networking services can easily lead to misunderstandings and disagreements between the parties (Czajkowski, 2002). Therefore, the need for a clear, precise, and unambiguous document detailing such issues has naturally arisen. The follow-

ing paragraphs of this article briefly explain how this purpose can be reached.

SLA CREATION

A service level agreement (SLA) is a legal document that is included or attached to a more general contract for an agreement of outsourcing activities to an external party. An SLA includes a description of the provided services and determines the specifications for the performance level and the results that the user of the service expects to receive from the provider (Salsano et al., 2000; Singh et al., 2002). The provider has to achieve or exceed these levels of performance.

Although SLAs can be defined for a wide range of services, they are extensively used in the area of network services provisioning.

SLAs can be categorized according to their objective as:

- Technical support
- Networking
- Applications
- Systems infrastructure
- Development environment
- Content
- Process support
- Process execution

The first three categories are the most common ones. A proper SLA has to deal with the following issues:

- What the provider promises
- How the provider will fulfill his or her promises
- Who and how is going to measure the performance
- What happens when the provider fails to reach the promised performance
- How the SLA is modified over time (if that is possible and acceptable)
- What exactly is the service that will be provided
- What is the level of service that the provider expects to reach
- What are the responsibilities of both parties
- How the quality of the service will be measured
- How will the achieved performance be reported
- What corrective measures the provider will take if the desired level of service is not achieved
- What is going to be paid for the service
- What financial compensations are going to be given if the desired levels of service are not reached

The above issues have to be dealt with and agreed upon, so that the SLA can contain the following:

- **Definition of the service:** A description of the service with exact definition of its extent and type.
- **Definition of terms related to the service:** Definition of the terms used in order to avoid misconceptions, especially if these terms are not clearly defined and generally accepted.
- **Specifications of the service and level of quality of the service:** Detailed specification of the service that will allow the verification of the level achieved, and also descriptions related to the service, such as availability and response time.
- **Method that will be used to measure the service, tools to be used, and values calculation:** Description of the way measurements are going to be made, the installations used, the calculation method for statistical quantities, and the method of results interpretation.
- **Compensations and other penalties:** The types and weights of penalties for failure to achieve the desired level of service, and possibly other more direct measures when the level is very low or failures very common.
- **Responsibilities, exceptions, and limits in accountability:** Responsibilities of the parties and exceptions for specific cases, and also compensations that may be paid.
- **Reports and types of documents:** Structure and format of all standardized documents that will be used for the documentation and management of the service.
- **Communication procedures and issue resolving:** Description of all communication procedures between the parties and definition of their relationships. Conflict resolution methods should also be described.
- **Prices and cost:** The cost may be provided as a whole or in parts, or a combination of the two (Wang et al., 2001).

The above issues will be organized in a document that will constitute the SLA.

Preparation of an SLA

Description of the Service

Before an organization defines the requirements for the level of service that will be required and who will be the provider, it first has to decide and make clear what the desired results would be, and also the general goal (such as cost reduction, performance increase, flexibility increase, and more choices for the users). The reason is that the metrics that will be defined vary depending on the desired goal. Furthermore, the organization has to clearly define the scope and the limits of the service, which means that the organization has to define not only what will be included in the SLA, but also what will not be.

Each of the above decisions influences the specifications that will be set and therefore the cost of the service. The description of the service will have to focus on what the organization expects to receive from the provider and not how the service will be implemented. The provider is the party that in using its specialized knowledge will implement the service in the most appropriate way. Interference by the organization might lead to non-optimal implementation and might increase the cost.

As an example, if the service is to interconnect two points, the organization has to define, depending on its needs, parameters such as the link capacity, the maximum number of errors, and the round trip time. The organization is not responsible for defining how the interconnection will be achieved, such as whether it will be wireless or wired, whether there will be intermediate nodes, what low level protocols are going to be used, etc.

Nevertheless, the implementation affects the maximum level of service that can be achieved (for example, in order to achieve very high availability, it is necessary to have excess equipment). The organization is therefore justified during the evaluation of offers to require implementation details during the query phase and to examine whether the offered implementation can support the required level of service. It should not, however, generally impose a specific implementation of the service.

Metrics

After the desired service has been described, the parties have to specify what is going to be measured for the duration of the service. The most basic characteristics of the service are determined based on the requirements of the organization. If, for example, the goal is to carry out some procedures quickly and properly, it is wise to study parameters such as speed and quality of the result. A common phenomenon is that the service is provided according to the specification but the users are not satisfied. This means that the parameters measured are not representative of what is acceptable and what is important. A solution to this problem is to define more than one (at least three) parameters for each part of the service and therefore to measure more **metrics**. Even if one of the metrics has not been chosen successfully, the rest of the metrics are likely to capture the "quality" perceived by the users.

The accuracy of the measurements also has to be determined, because too much accuracy is often undesirable because of its cost or complexity. Sometimes a parameter cannot be measured in all cases (or it is not desirable to do so). In such cases there can be a sample measurement and the metrics defined for the service can be obtained from these samples.

Boundaries for the Metrics

After the metrics have been defined, appropriate boundaries for these metrics have to be specified. This can be achieved by studying relevant data in the literature or measuring comparable data from the current situation (the service is provided internally by the organization). These values are then compared to the levels offered by the market, and it is judged whether improvement is necessary, and in what degree. If some of the metrics are statistical in nature, the time period over which the measurement is made has to be taken into account. In order to avoid situations where the service is provided according to the statistical requirements, but in specific cases the service is out of bounds, there have to be additional (more flexible) boundaries for short periods.

There are two kinds of boundaries set for the metrics of the service:

- A level below which the service is no longer considered acceptable; such a violation leads to measures such as compensation of the organiza-

tion and obligation of the provider to improve the service.

- A level that the service should never fall below, and in which case it is violated, leads to the assumption that the provider is no longer capable of acceptably providing the service and may bring about drastic measures such as termination of the contract or revision of the financial agreement.

Therefore, the boundaries defined should be close to what is acceptable and what is not, and should be realistic. Unnecessarily strict boundaries are costly and can often cause violations that might incur unnecessary burden to both the provider and the organization.

Compensations

It is improbable that a service will always be offered at the specified levels of quality, and therefore cases of violation should be expected. The SLA should provision compensations both because the organization did not receive the proper service (and therefore should not be charged for it) and because the provider should be punished for breaching the SLA. Compensations should be reasonable, because unreasonably high compensations lead to reduced interest by the provider market, while unreasonably low compensations give little motive for the provider to make available the proper levels of service.

Usually, compensations are defined as a percentage of the provider's fee. Continuous or multiple violations of the SLA are undesirable, even if compensated. In order to avoid such violations, the SLA can set limits on the number or duration of violations and provision drastic measures (such as contract termination) if these limits are exceeded.

Measurements

The next step is to determine how the metrics are going to be measured, who is going to perform the measurements, and how they are going to be reported. This is an issue of how the measurements are going to take place, and how the parties are going to agree on the measured quantities. The main problem is that by assigning the service to the provider, the organization often does not have the necessary technical knowledge in order to set up a measurement system. Furthermore, the measurement system should be recognized by the provider and the measurements have to be accepted as realistic.

In order to deal with such complexities, it is possible that:

- The organization preserves part of the technical knowledge necessary for the implementation and maintenance of a measurement system.
- The measurement system is implemented and used by both the organization and the provider.
- The organization and the provider maintain independent measurement systems, in order to verify each other's credibility. However, this approach introduces the issue of resolving conflicting measurements.
- Measurements are performed by a third party that is credible with both parties. The system is automated and produces results without interference.
- Measurements are performed by the provider, who is, however, supervised by an independent auditor, authorized by the organization.

Regardless of the measurement method chosen and the source of the data, the party responsible for collecting them and the party processing them have to be decided upon and thoroughly described. These data will be used for generating reports containing the measurements of the metrics, violations of the SLA (if any), consequences, measures for avoiding future violations, and any other suggestions. The source, the periodicity, and the scope of circulation for these reports have also to be decided.

Communication Procedures

Defining the **communication** procedures between the organization and the provider is important so that no misunderstandings are created, and in order to have a clear procedure for reporting the appearance and fixing of problems. Furthermore, archived communication data can be used for measuring metrics such as response time and service restoration time.

Other Issues

Finally, the possibility that the SLA should cover issues specific to each case also has to be investigated. In case such issues exist, they should be recorded, along with each party's responsibility for each.

Sample Metrics for Various Services

The metrics presented below are relatively general and can be used for a variety of services. In any case, their usefulness is determined according to the criteria mentioned previously and the levels of service quality should be set according to need.

Some indicative metrics are the following:

- **Availability:** This metric is defined as the percentage of time that the service can be accessed. Usually acceptable values of the availability metric are between 98-100 percent. Common bounds under which drastic measures are taken are 2-3 points lower than the above values (95- 98 percent). In order to avoid prolonged periods of non-availability, the calculation of availability should take place at short intervals (such as daily) or limits set on each period of unavailability (such as half an hour). **Availability** can be measured using an automated system that periodically checks access to the service (for example, for a Web service the check would be the fetching of several pages, for an interconnection service the check would be the transmission and reception of a small message). If the measurement system is outside the provider's network, it might not be able to access the service because of a problem that is not the provider's responsibility. If the measurement system is inside the provider's network, it might not be able to identify connectivity problems for the whole of the provider's network. Therefore, the measurement system should be located at the edge of the provider's network.
- **Capacity:** Capacity is the size of the connection used for some interconnection service (Machiraju, et al., 2002). It influences, along with other parameters, the speed that can be achieved at a network link. Also, in cases of interconnection services to the Internet, capacity only refers to the provider's network. The overall speed of communication with other network nodes might be slow because of reasons outside the provider's responsibility.
- **Number of updates (such as the changes at a Web site's pages):** It defines the responsibility of the provider to update the content related to the service. Alternatively, the rate of updates can be used.
- **Service response time:** Service response time is the time from the moment that a user makes a request to the server until the result of the request comes back. Usually the response time cannot be measured for each individual request, and therefore this metric is the average response time, probably accompanied by its variation. Alternatively, service response time can be defined as the response time for a large number of requests. Service response time can be measured using an automated system that sends requests to the server and calculates the average response time.
- **Number of requests to the server:** This is the number of requests to the server (such as a Web server) in a specific time period, and which have to be served according to the quality level defined by the rest of the metrics. Measurement can be made using the logging service of the server.

FUTURE TRENDS

Networking services have become ever more important in the economy and the everyday operation of most businesses. Together with increased complexity come also increased requirements from the users of these services, as they rely on them for their productivity. These trends make the existence of a properly designed SLA document very important, both for the users and providers of the service, who want to avoid excessive and unpredictable liability. The critical nature of networking services in modern business environment requires strict and unambiguously defined rules and agreements. This evolution is certain to continue in the future, which will result in the greater need for complex but clearly written, usable, and real-world functional SLAs.

CONCLUSION

As SLAs are complex documents, the creation of a proper SLA can be a demanding task. In order to become more manageable, it is common to use a template and other tools for generating the basic document. Because SLAs are usually part of a contract and are considered confidential business information, examples of real-world SLAs are not very easy to find.

A good SLA cannot guarantee customer satisfaction only by having specific objectives and metrics. A good SLA should also include compensations and remedies for situations when the specified service level goals are missed. These compensations should motivate the provider into achieving the objectives, and the remedies should guarantee that the service quality will return to acceptable levels quickly.

REFERENCES

Bouras, C., Campanella, M., & Sevasti, A. (2002), SLA definition for the provision of an EF-based service In *Proceedings of the 16th International Workshop on Communications Quality and Reliability (CQR 2002)* (pp. 17-21).

Bouras, C., & Sevasti, A. (2004). SLA-based QoS pricing in DiffServ networks. In F. Huebner & R. D. van der Mei (Eds.), *Computer Communications Journal, Special Issue, Performance and control of next generation communications networks, 27*(18), 1868-1880.

Coollawyer.com. *Services agreements and services level agreements*. Retrieved 27 July, 2006 from http://www.coollawyer.com/webfront/internet_law_library/articles/law_library_service_level_article.php

Czajkowski, K., Foster, I., Kesselman, C., Sander V., & Tuecke S. (2002, July). SNAP: A protocol for negotiating service level agreements and coordinating resource management in distributed systems. In *Proceedings of 8th Workshop on Job Scheduling Strategies for Parallel Processing* (pp. 153-183). Edinburgh, Scotland.

Darwin Executive Guides. Service level agreements. Retrieved from http://guide.darwinmag.com/technology/outsourcing/sla/

Heinanen, J., Baker, F., Weiss, W., & Wroclawski, J. (1999). *Assured forwarding PHB group.* RFC 2597. Accessed 27 July, 2006.

IEC.org. Visual networks, inc. & telechoice: Carrier service-level agreements (SLAs). Retrieved March, 2006 from, http://www.iec.org/online/tutorials/carrier_sla/index.html

Jacobson, V., Nichols, K., & Poduri, K. (1999). *An expedited forwarding PHB.*RFC 2598. Accessed 27 July 2006.

Machiraju, S., Seshadri, M., & Stoica, I. (2002). A scalable and robust solution for bandwidth allocation. In *Proceedings 10th International Workshop on Quality of Service (IWQoS),* Miami (pp. 148-157).

Salsano, S., et al. (November 2000). *Definition and usage of SLSs in the AQUILA consortium, draft-salsano-aquila-sls-00.txt.* Internet Draft.

Singh, D. M., Garg, R., Singh, R. R., & Saran, H. (2002). A SLA framework for QoS provisioning and dynamic capacity allocation. In *Proceedings of Tenth International Workshop on Quality of Service (IWQoS),* Miami Beach (pp. 129-137).

The SLA Toolkit. http://www.service-level-agreement.net/verizonbusiness.com. ervices agreements and services level agreements. Retrieved 27 July, 2006 from http://www.verizonbusiness.com/terms/us/products/internet/sla/

Wang, X., & Schulzrinne, H. (2001). Pricing network resources for adaptive applications in a differentiated services network. *INFOCOM*, 943-952.

Xiao, X., & Ni, L. M. (1999). Internet QoS: A big picture. *IEEE Network*, 13(2), 8-18.

KEY TERMS

Availability: In the context of this article, availability is the percentage of time that a specific service is operating satisfactorily. The exact definition of the availability (for example, under which circumstances the service is no longer considered available) is a subject of the SLA for the service.

Capacity: In the context of this article, capacity is the bandwidth available for a specific service. The exact definition of the capacity (for example, at which network layer it should be measured) is a subject of the SLA for the service.

Compensations: The possible actions (such as payments or forfeiture of a fee) against a service provider that fails to meet the level of service defined in the SLA.

Metrics: The characteristics that are properly well-defined so that they can be measured and used to evaluate the service and its compliance to the SLA.

Quality of Service (QoS): The ability to provide specific guarantees to traffic flows regarding the network characteristics such as packet loss, delay, and jitter experienced by the flows.

SLA: Service level agreement—a formal agreement between a Service Provider and a Service Client to provide a service with well-defined service levels, accompanied by possible penalty clauses if the SLA is not met.

Service Response Time: Service response time is the time from the moment that a user makes a request to the server, until the result of the request comes back.