Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 537–540

ISBN 978-83-60810-22-4

Violation of Service Availability Targets in Service Level Agreements

Loretta Mastroeni
Department of Economics
University of Rome Tre
Via Silvio D'Amico 77, 00145 Roma, Italy
E-mail: mastroen@uniroma3.it

Maurizio Naldi
Department of Computer Science
University of Rome at Tor Vergata
Via del Politecnico 1, 00133 Roma, Italy
E-mail: naldi@disp.uniroma2.it

***Abstract*—Targets on availability are generally included in any Service Level Agreement (SLA). When those targets are not met, the service provider has to compensate the customer. The obligation to compensate may represent a significant risk for the service provider, if the SLA is repeatedly violated. In this paper we evaluate the probability that a SLA commitment on the service availability is violated, when the service restoration time follows an exponential, Weibull, or lognormal distribution. For a two state model, where the service alternates between availability and unavailability periods, we show that such probability decreases as the variance of the restoration time grows, and that lengthening the time interval over which the service availability is evaluated reduces the risk for the service provider just if the compensation grows quite less than the length of that time interval.**

I. Introduction

SERVICE level agreements (SLAs) define the contractual obligations of the service provider towards the customer, the mechanisms to enforce the delivery of the committed service quality, and the obligations of the service provider if the service level falls below the committed value [1], [2].

A key parameter in SLAs is the service availability, for which a target figure is declared. If the service is unavailable, the customer is not provided what it has paid for (an economical loss in itself), but suffers an additional larger loss due to the discontinuity in its business operations or social relationships. The latter category of losses may reach values of the order of 100-200 k$ per minute of service interruption (see [3]).

The obligations of the service provider generally consist in the payment of a sum if the target availability figure is not met. If the service availability targets are not met repeatedly, the service provider is bound to suffer large losses, especially if that happens on a massive scale rather than for a few individual customers.

The service provider has to be able to evaluate the risk it incurs because of SLA violations, which in turn requires to evaluate the probability that the target figures are not met. Many SLA monitoring tools exist: HP OpenView Firehunter, CiscoWorks2000 Service Management Solution, and Lucent's CyberService, among those marketed in the recent past. However, SLA monitoring tools allow to perform an *ex-post* evaluation of the quality of service delivered, rather than the predictive evaluation that the service provider needs to properly set its commitment and negotiate a sustainable SLA.

In [4] the probability of violating availability SLAs has been evaluated by simulation, when the service restoration time follows an exponential distribution, but two more complex distributions are envisaged for the service restoration time, namely the Weibull and the lognormal. The same Weibull distribution is suggested in [5] as the best-fit model in the cases of grid computing services.

In this paper we provide a thorough examination of the risk of violating the SLA obligations, considering three models (exponential, Weibull, and lognormal) for the service restoration time. We provide an analytical expression for the probability of violating an availability SLA commitment for the exponential case, while previous results were based on simulation only. We provide simulation results for the same probability of violation, when the service restoration time follows instead a Weibull or lognormal distribution, which had not been dealt with in the literature. We show that the probability of SLA violation decreases as the variance of restoration times grows, and that lengthening the time interval over which the availability targets are examined is convenient for the service provider just if the compensation amount for each violation grows quite less than proportionally with the length of that time interval.

The paper is organized as follows. In Section II, we define the service model we adopt for our analysis. In Section III, we provide a formal definition of the service level agreement for availability and of the compensation policy. Finally, we provide in Section IV the results for the three models considered.

II. Service model

In order to assess the violations of SLA obligations, we need a model for the service provided to the customer. In this section, we describe a simple model based on the alternation of ON and OFF states, and provide the definition of availability.

We consider the service to be either available or not. Though the customer could experience a graceful performance degradation, SLA commitments are sharp [6]. At time t, the state S_t of the service equals 1 if the service is available, and 0 otherwise. The service undergoes a sequence of alternating availability and unavailability (ON and OFF) states, whose average durations are respectively the *Mean Time To Failure* (MTTF) and *Mean Time To Repair* (MTTR). The durations of the OFF periods are represented by the sequence of positive

i.i.d. random variables $\{B_1, B_2, B_3, \ldots\}$. We assume that the service starts in the ON state. The variable $N_T \in \mathbb{N}_0$ represents the number of failures in the period $(0, T]$ ($N_T = 0$ means that the service works uninterruptedly in $(0, T]$).

The service model is fully specified when we define the probability distribution for the duration of the ON and OFF periods. Here we assume that the duration of the ON period follows an exponential distribution, and the duration of the OFF period follows either an exponential distribution, or a Weibull, or a lognormal distribution. Two-parameter models (lognormal and Weibull) are used for more complex repair scenarios such as with significant travel time [4].

As the key service performance parameter for our model, we consider the availability. For our two-state model, the steady-state availability Φ is defined as the expected value of the state variable [7], or, equivalently, as the probability that the service is ON, and can also be expressed through MTTF and MTTR:

$$\Phi = \mathbb{E}[S_t] = \mathbb{P}[S_t = 1] = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}}. \tag{1}$$

III. Service Level Agreements and Compensation Policies

In SLAs, the service provider commits itself to provide an adequate quality of service, and compensate the customer if that commitment is not honored. In this section, we review the definition of service availability targets, and describe the compensation policy considered in the following.

In SLAs, target values are indicated for service availability [8], [9]. In the basic definition (1), we must state what the object is whose availability we consider, and how we declare that object to be available or not. For example, in [10] a list of services and the associated definitions of availability are provided.

In order to check if the SLA obligations are met in an operational context, we set an observation interval T and measure the availability through the ratio of the cumulative outage duration X_T during the observation interval and the length of the observation interval itself:

$$\hat{\Phi} = \frac{T - X_T}{T} = 1 - \frac{\sum_{i=0}^{N_T} B_i}{T}. \tag{2}$$

If we fix the length of the observation interval, the SLA obligation $\hat{\Phi} > z$ (the threshold z being a positive quantity) can be expressed as the constraint $X_T < W = (1-z)T$ on the cumulative outage duration X_T over the observation interval. In particular, we can set a threshold $z = \Phi$ equal to the declared steady-state availability. However, due to the random nature of the failure process, there is a non-zero probability that the SLA obligation is violated.

The compensation policy states what the service provider is to pay its customer when the service fails. We assume that the compensation is paid out for failures occurring over a period of time of extension T (the observation period), rather than on each single failure. In this paper, we consider a simple compensation policy based on the steady-state availability: a fixed amount of money is paid when $X_T > W = (1 - \Phi)T$.

IV. Probability of Violation of Availability Targets

The risk for service providers, deriving from the unfulfilled commitments, depends on the probability of violating the SLA targets. In this section, we provide results for the cases where the service restoration times follow an exponential, a Weibull, or a lognormal distribution. For the exponential case, we obtain an approximate analytical expression. Instead, for the Weibull and lognormal cases, we resort to simulation.

Exponential restoration times. In services with high availability, we have MTTR $\ll$ MTTF. In that case, the process of failure occurrences can be approximated by a Poisson process. Over a finite horizon T, the number N_T of failures follows approximately a Poisson distribution with average value λT, where λ is the failure rate. By the total probability theorem, and recognizing that the number of failures and the duration of the outages are independent of each other, we can express the probability of violation as

$$\begin{aligned}\mathbb{P}[X_T > W] &\simeq \mathbb{P}\left[\sum_{i=0}^{N_T} B_i > W\right] \\ &= \sum_{k=0}^{\infty} \mathbb{P}[N_T = k] \cdot \mathbb{P}\left[\sum_{i=0}^{N_T} B_i > W | N_T = k\right] \\ &= \sum_{k=1}^{\infty} \mathbb{P}[N_T = k] \cdot \mathbb{P}\left[\sum_{i=1}^{k} B_i > W\right]\end{aligned} \tag{3}$$

Since the durations of outages are i.i.d. random variables with an exponential distribution, their sum follows an Erlang distribution. When there are k failures, the probability distribution of the cumulative outage duration is

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{k} B_i \le x\right] &= \int_0^x \frac{\mu^k e^{-\mu v} v^{k-1}}{(k-1)!} dv \qquad x > 0 \\ &= \frac{1}{(k-1)!}\gamma(k, \mu x),\end{aligned} \tag{4}$$

where $\gamma(k, x)$ is the lower incomplete Gamma function [11].

By replacing the expression of the Poisson distribution and the result (4) in the probability of violation (3), we obtain the final expression

$$\begin{aligned}\mathbb{P}[X_T > W] &= \sum_{k=1}^{\infty} \frac{(\lambda T)^k}{k!} e^{-\lambda T}\left[1 - \frac{1}{(k-1)!}\gamma(k, \mu W)\right] \\ &= 1 - e^{-\lambda T} - \sum_{k=1}^{\infty} \frac{(\lambda T)^k}{k!(k-1)!} e^{-\lambda T}\gamma(k, \mu W)\end{aligned} \tag{5}$$

The resulting probability of violating the SLA depends on the failure rate λ, the observation interval T, and the threshold W for the overall duration of outages $W = (1 - \Phi)T$.

We report in Fig. 1 the probability of SLA violation for several values of the steady-state availability, from 95% to the excellent five nines case. We assume MTTR=4 hours, a value typically adopted (see Chapter 2.2 in [12]). The range of values

TABLE I
EXPECTED NUMBER OF TARGET VIOLATIONS OVER ONE YEAR

Obs. int. T [months]	Viol. prob. over T	Expected violations
1	0.140	1.679
2	0.224	1.345
3	0.277	1.109
4	0.312	0.936
6	0.353	0.706
12	0.401	0.401

for the observation interval goes up to 1 year, corresponding to 8640 hours. The probability of violating the targets grows

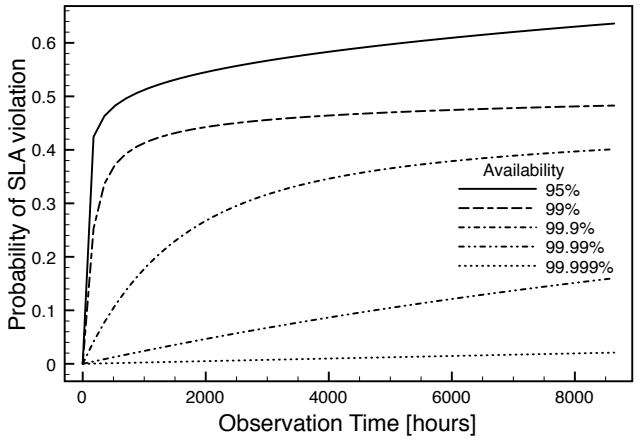

Fig. 1. Probabiity of violation under exponential restoration times

with the observation interval. This would seem contradictory with the generally held view that adopting long measurement intervals favors the service provider, since it allows to smooth out single events of protracted unavailability. Actually, we have to assess the overall number of violations taking place over an evaluation period of fixed length. If we reduce the length of the observation interval T, the probability of violation over T reduces, but the number of observation intervals included in the evaluation period increases. For example, for the case where $A = 99.9\%$ and MTTR $= 4$ hours, we see in Table I that the expected number of violations over a year actually decreases as we lengthen the observation interval: service providers may reduce their risk by lengthening the observation interval. However, we should consider the economical loss deriving from the application of the compensation policy. We expect the compensation sum to increase as the observation interval lengthens. We should therefore multiply the expected number of violations (third column in Table I) by the compensation paid for each violation. The data in the table show that lengthening the observation interval from one month to one year reduces the overall expected loss if $C_{12}/C_1 < 1.679/0.401 \simeq 4.19$, i.e., quite less than 12 times, the increase in the observation interval.

If the service provider adopts a threshold on the measured availability larger than the steady state availability, it may bring the probability of violation down to acceptable values. For example, in [13] it is envisaged that the service provider may revise the performance objectives (e.g., by relaxing the constraint on the availability), if the SLA obligations are not being met. In Fig. 2, we see that the probability of violating the SLA obligation decreases when we raise the threshold over the steady-state availability (43 minutes for $T = 1$ month and 2 hours 19 minutes for 6 months, when the steady-state availability is 99.9% and MTTR=4 hours).

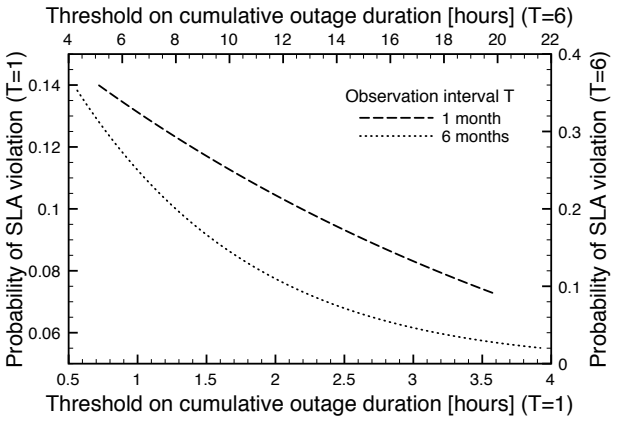

Fig. 2. Impact of limit outage duration on the probability of SLA violation

Weibull restoration times. The probability distribution of restoration times may differ from the exponential. For example, in [5] a Weibull distribution is proposed to model the duration of outages in grid computing.

Under the Weibull hypothesis, the probability distribution for duration of the generic i-th outage is

$$\mathbb{P}[B_i < x] = 1 - e^{-(x/\sigma)^\theta} \qquad x \geq 0, \tag{6}$$

where σ is the scale factor, and θ is the shape factor. When $\theta = 1$ the Weibull distribution becomes the exponential one. For the case of grid computing, the shape factor should lie in the $[0.6, 1]$ range [5]. When $\theta < 1$, the variance of the service restoration time increases with respect to the exponential case.

We determine the probability of violation by simulation, since no closed form exists for the distribution of the sum of i.i.d. Weibull random variables. We consider 10^5 instances of the observation interval, generating the number of failures according to a Poisson process, and the duration of each outage through a Weibull-distributed random number.

In Fig. 3 we show the probability of violating the SLA, when MTTR=4 hours and $\theta = 0.7$ (a standard deviation slightly lower than six hours, against the four hours of the exponential case). Despite the larger variance of the restoration time in the Weibull case, the violation probability is slightly lower than in the exponential case.

In Fig. 4, we examine in greater detail the impact of the shape factor, when the steady-state availability is 99.9%. Since the variance of the service restoration time grows as θ decreases, the probability of SLA violation decreases as the

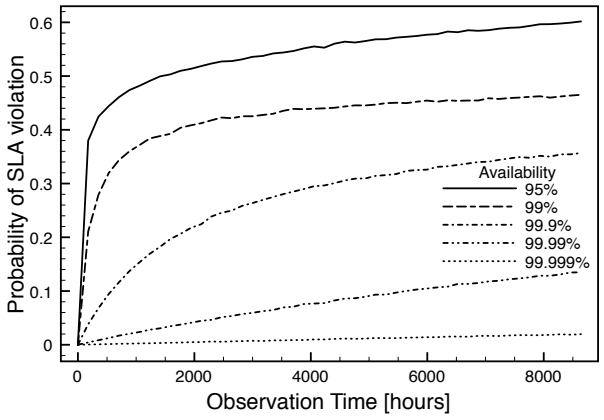

Fig. 3. Probability of violation under Weibull repair times ($\theta = 0.7$)

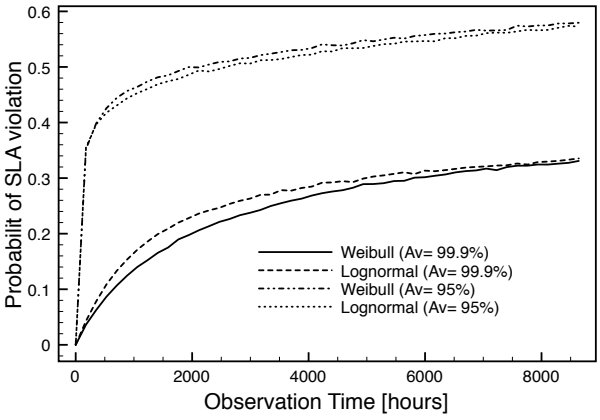

Fig. 5. Comparison between the Weibull and the lognormal case

variance grows. When $\theta = 0.6$, the violation probability is roughly 17.5% lower than in the exponential case.

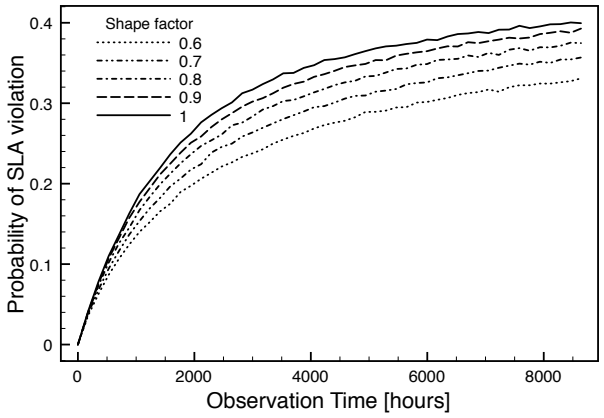

Fig. 4. Probability of violation under Weibull repair times ($\Phi = 99.9\%$)

Lognormal restoration times. In addition to the exponential and the Weibull case, the lognormal model has been proposed in [4] for the service restoration times.

Again, no closed form exists for the probability distribution of the sum of lognormal random variables, and we resort to simulation. We adopt a restoration time with mean value MTTR=4 hours, and a standard deviation ranging from 4 hours (as in the exponential case) to 7 hours (as in the Weibull case with $\theta = 0.6$), and 10^5 simulation instances. In Fig. 5, we compare with the Weibull case for the largest standard deviation of the service restoration time (7 hours): the differences are larger for the high availability case ($\Phi = 99.9\%$), but quite negligible when the steady-state availability is not very large, and smooth out as the observation interval lengthens.

V. Conclusion

We have evaluated the probability that the availability commitments included in a Service Level Agreement are not met. The analysis has been conducted for a two-state service model, with alternating periods of service availability and service restoration. Three probability models have been considered for the service restoration times: exponential, Weibull, and lognormal. We have shown that the availability target values are less likely to be violated as the variance of the service restoration time gets larger. If the service providers opts for longer evaluation intervals (for the assessment of SLA commitments), it must set compensations quite less than proportional to the length of the observation interval itself.

References

[1] A. Keller and H. Ludwig, "The WSLA Framework: Specifying and Monitoring Service Level Agreements for Web Services," *J. Network Syst. Manage.*, vol. 11, no. 1, pp. 57–81, 2003.

[2] H. Ludwig, A. Keller, A. Dan, R. P. King, and R. Franck, "A Service Level Agreement Language for Dynamic Electronic Services," *Electronic Commerce Research*, vol. 3, no. 1-2, pp. 43–59, 2003.

[3] M. Pesola, "Network protection is a key stroke," Financial Times, FT Business Continuity, March 9, 2004.

[4] A. P. Snow and G. R. Weckman, "What Are the Chances an Availability SLA will be Violated ?" in *Sixth International Conference on Networking (ICN 2007)*, 2007, p. 35.

[5] R. Alsoghayer and K. Djemame, "Probabilistic risk assessment for resource provision in grids," in *Proceedings of the 25th UK Performance Engineering Workshop*, Leeds, 6-7 July 2009, pp. 99–110.

[6] A. Michlmayr, F. Rosenberg, P. Leitner, and S. Dustdar, "Comprehensive QoS Monitoring of Web Services and Event-Based SLA Violation Detection," in *MW4SOC 09*, Urbana Champaign, Illinois, USA, 30 November 2009.

[7] T. Aven and U. Jensen, *Stochastic Models in Reliability*. Springer, 1999.

[8] M. Vogt, R. Martens, and T. Andvaag, "Availability modeling of services in IP networks," in *Design of Reliable Communication Networks, 2003. (DRCN 2003). Proceedings. Fourth International Workshop on*, October 2003, pp. 167 – 172.

[9] E. Bouillet, D. Mitra, and K. Ramakrishnan, "The structure and management of service level agreements in networks," *Selected Areas in Communications, IEEE Journal on*, vol. 20, no. 4, pp. 691 –699, May 2002.

[10] P. Cholda, J. Tapolcai, T. Cinkler, K. Wajda, and A. Jajszczyk, "Quality of resilience as a network reliability characterization tool," *IEEE Network*, vol. 23, no. 2, pp. 11–19, 2009.

[11] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, Eds., *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[12] E. Bauer, *Practical System Reliability*. J. Wiley-IEEE Press, 2009.

[13] D. Verma, "Service level agreements on IP networks," *Proceedings of the IEEE*, vol. 92, no. 9, pp. 1382 – 1388, September 2004.