The Cumulus Assessment Module as General SLA Evaluation Mechanism for Telecommunication Services

Peter Reichl[1] and Wolfgang Haidegger[2]
[1]Telecommunications Research Center (FTW) Vienna, Donau-City-Str. 1, A-1220 Wien
[2]Siemens AG Österreich, Siemensstrasse 88-92, A-1211 Wien
reichl@ftw.at; wolfgang.haidegger@siemens.at

Keywords:
Service Level Agreements, Operations Systems and Business Systems Support, Sequential Probability Ratio Test, Network Management

Abstract:
Service Level Agreements (SLAs) form the heart of every business relation between service users and service providers. Their contents feed the network management systems as well as operation support and business support systems and may vary widely depending on the type of service provided. Therefore, a general mechanism to assess their fulfilment is extremely helpful for both contract partners, but by no means easy to be realized. This paper proposes the Cumulus Assessment Module (CAM) as an efficient general solution for this problem. After introducing its fundamental statistical functionality based on Wald's Sequential Probability Ratio Test, we describe how CAM fits into a general management architecture and define its abstract interfaces. Finally, the validity of our approach is demonstrated with two diverse examples taken from security management and the management of Storage Area Networks.

1. INTRODUCTION

While the evolution of telecommunications had taken a relatively linear course from the analog over the digital towards the intelligent network, the start of the exponential Internet growth together with the rising propagation of mobile networks has led to drastic consequences for both telecommunications services and underlying networks. The services come both from the IP-world (like "Click-to-Dial") and from the old telecom world (e.g. "Hunting") and are provided end-to-end, i.e. from end-user to end-user without any regard to the underlying infrastructure. On the other hand, the communication networks on which these services are based can be highly heterogeneous. Especially big network providers will comprise circuit switched domains, IP-domains with various Quality of Service mechanisms, perhaps an optical core and Frame Relay or ATM networks as legacy networks. These network collections are mostly integrated using various gateways both on the transport and on the control layer.

Providers of transport and application services provide individual packages of subsets of their services to customers according to the latter ones' business models. The contract which fixes the services and the quality with which they have to be provided is widely known as Service Level Agreement (SLA). Today it is still a widespread standard procedure to agree on these SLAs by telephone and fax, followed by the provider taking some time to provision the network accordingly. This approach makes mutual control of the fulfilment difficult to realize (and from the customer side virtually impossible).

One of the central problems with SLA supervision concerns the huge amount of data generated within the network, especially in the case of IP based communication networks. This can be solved by devising an efficient SLA method to assess the fulfillment of SLAs pretty accurately without requiring all available data, maybe in line with the well-known general 80/20 paradigm (reaching 80% of correctness with 20% of effort).

Another problem concerns the heterogeneous information, which can be elicited from the network in an end-to-end scenario. The homogenisation of QoS data pertaining to one specific service across the network is still

a challenge. In addition it is neither elegant nor practical to implement individual assessment mechanisms for every type of service provided.

In this paper we propose a generic mechanism, the Cumulus Assessment, which allows controlling the conformance of all contract partners with respect to any kind of SLA. It is a generalisation of the Cumulus Pricing Scheme (CPS) introduced in [5], which represents one specific application area of SLA assessment.

In this paper we start with an introduction of CPS and present a general version of the statistical mechanism the assessment is based on. Next, we discuss the interfaces of the resulting Cumulus Assessment Module (CAM) along with architectural issues. Two practical examples illustrate the validity of the approach, before a couple of summarizing conclusions finish the paper.

2. CUMULUS PRICING AND THE REVERSE SEQUENTIAL PROBABILITY RATIO TEST

2.1. The Cumulus Pricing Scheme (CPS)

The Cumulus Pricing Scheme CPS has been proposed in the context of Internet pricing to allow for a technically feasible flat-rate based charging scheme which nevertheless allows to react in a delayed manner to misbehaving end-users. The basic idea consists of using statistical probes of network parameters in order to determine whether the previously agreed contract between customer and provider is still fulfilled. If not, so-called “Cumulus Points” are sent out to the customer as an early warning mechanism and are kept over time until their accumulation exceeds a certain threshold and thus makes a renegotiation of the contract necessary. For a more detailed discussion of this scheme we refer to [6] and references theirin.

One of the enticing features of this mechanism is its capability of delivering pretty good results, even in case of rather sparse data. This allows the pulling of statistical probes from the network instead of collecting data all the time throughout the networks.

The second attractive property of the mechanism concerns the fact that it can be used for such diverse services as providing resource usage or security properties. This is closely related to the robustness of the basic statistical mechanism even in case of spare data, as will be discussed later on in this section.

The third advantage of this mechanism is the ease with which it can be fit into a general network management architecture, where is represents the connection between data pulled directly from the network, Operation systems Support (OSS) functions and Business Systems Support (BSS) functions. It can be put into one module, the Cumulus Assessment Module (CAM), and used centrally for varying input.

2.2. The Reversed Sequential Probability Ratio Test

The rest of the section describes the statistical mechanism CAM uses for assessing the SLA fulfillment. To this end, we start with describing the fundamental problem of SLA assessment in a formal way. W.l.o.g. we assume that the SLA consists of just one parameter (random variable) X for which a target value μ_0 and an upper bound ξ is given (note that this case can be easily generalized to more than one parameter and to a corridor with upper and lower bounds ξ and ϑ, resp). The SLA is considered to be out of balance if the realizations $X = x_1, x_2 ...$ are on average $\geq \xi$, whereas the SLA is fulfilled otherwise. To assess the SLA fulfilment from parameter measurements taken over the run time of the contract, one could for instance use a (weighted or exponential) averaging procedure with a well-chosen window size. However, we propose to use a much more sensitive approach based on the Sequential Probability Ratio Test (SPRT) due to Wald, which has the additional advantage of taking explicitly into account the probabilities α for errors of the first kind (SLA is considered to be in balance but is not) and β for errors of the second kind (SLA is considered to be out of balance but is in balance).

To this end, we assume the realizations $X = x_1, x_2 ...$ are normally distributed with standard deviation equal σ which is known to the CAM. In order to assess whether the SLA is out of balance, we have to find out which of the following hypotheses is correct:

H_0: X is $N(\mu_0, \sigma)$-distributed

H_1: X is $N(\mu_1, \sigma)$-distributed with $\mu_1 \geq \xi$.

In any case, it is sufficient to assume $\mu_1 = \xi$. The standard SPRT is based on calculating the series of so-called likelihood ratios

$$\Lambda_{i,j} = \frac{f(x_i, \mu_1)}{f(x_i, \mu_0)} \cdot \frac{f(x_{i+1}, \mu_1)}{f(x_{i+1}, \mu_0)} \cdot \ldots \cdot \frac{f(x_j, \mu_1)}{f(x_j, \mu_0)},$$

where $f(x_i, \mu)$ is calculated as the value of the density function $f(., \mu)$ of the normal distribution $N(\mu, \sigma)$.

Depending on α and β, we determine two magic numbers A and B (see below) and proceed as follows: Starting from realization x_i, we determine $\Lambda_{i,i} = \frac{f(x_i, \mu_1)}{f(x_i, \mu_0)}$. If $\Lambda_{i,i} > A$, then we reject H_0 in favour of H_1, whereas for $\Lambda_{i,i} < B$, we accept H_0. If however $B \le \Lambda_{i,i} \le A$, we cannot make an immediate decision, but continue with sampling x_{i+1}, calculate $\Lambda_{i,i+1} = \frac{f(x_{i+1}, \mu_1)}{f(x_{i+1}, \mu_0)} \cdot \Lambda_{i,i}$, compare it with A and B and decide about accepting one of the hypotheses in the same way as we did with $\Lambda_{i,i}$ etc. until we come to a decision.

Determining the numbers A and B may be quite tricky, therefore we only state Wald's result suggesting that $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$ provide an approximation sufficient for all practical purposes.

There is a huge literature discussing the robustness and stability of the resulting decisions, for details see e.g. [8]. Note however that we need the following trick to make this algorithm useful for our SLA assessment problem: as we continuously progress with our measurements in time, in order to determine whether the SLA is currently fulfilled or not, we have to use the most recent measurements in reverse order. Hence, having taken probe x_i we start with calculating $\Lambda_{i,i}$, and in case we cannot make a definitive decision we continue with $\Lambda_{i,i-1} = \frac{f(x_{i-1}, \mu_1)}{f(x_{i-1}, \mu_0)} \cdot \Lambda_{i,i}$ etc.

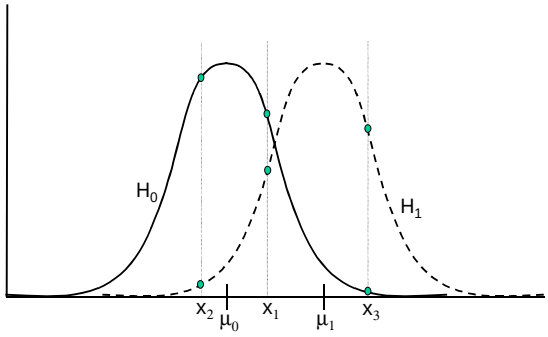

Figure 1: Illustration of SPRT

Figure 1 illustrates the fundamental principle of SPRT. The density functions of the normal distributions corresponding to both hypotheses H_0 and H_1 are the well-known Gaussian bells with means μ_0 and μ_1 and identical standard deviations σ. Three realizations x_1, x_2, x_3 of the random variable X are indicated together with the respective values of the probability density functions (small circles). From this figure we can observe easily that $\Lambda_{1,1} = \frac{f(x_1, \mu_1)}{f(x_1, \mu_0)} \approx 1$, $\Lambda_{2,2} << 1$ and $\Lambda_{3,3} >> 1$. Thus, for any reasonable values A and B and x_1 as our first measurement, we will end up with no decision and further probing. Our next probe x_2 might already lead to $\Lambda_{2,2} < B$, i.e. the acceptance of H_0 (as one would expect reasonably from a sample being smaller than μ_0). If we assume, however, that $\Lambda_{2,2}$ is a value just above B, according to our reverted SPRT version we calculate $\Lambda_{2,1} = \Lambda_{2,2} \cdot \Lambda_{1,1}$, which is smaller than $\Lambda_{2,2}$ due to $\Lambda_{1,1}$ being slightly smaller than 1 (in our example) and thus might already lead to $\Lambda_{2,1} < B$. If this is still not the case, we will have to continue probing (of course, the sketched x_3 will rather feed the acceptance of H_1).

Finally note that the size of σ is of significant influence for the speed of convergence of the algorithm. In fact, it has been proved for a similar application of the reverse

SPRT [9] that $\ln \Lambda_{i,j}$ is decreasing/increasing in direct proportion to $\frac{\mu_1 - \mu_0}{\sigma^2}$. This allows the immediate conclusion that the speed of convergence is directly proportional to the variance of the normal distribution. Hence, if the raw network data are already preprocessed, e.g. using some IPPM (IP Performance Metric) mechanism, this leads to smaller σ (remember the well-known statistical fact that averaging over n i.i.d. random variables decreases the standard deviation by a factor of $1/\sqrt{n}$) and hence may significantly increase the speed of deciding whether the SLA is out of balance or not.

3. CUMULUS ASSESSMENT MODULE DESCRIPTION AND ARCHITECTURAL ISSUES

Figure 2 shows the principal architecture of the Cumulus Assessment Module CAM with its in- and outputs. It is based on the following assumptions concerning the contents of SLAs:

- For each service there will be scalar entities (in most cases real numbers) which describe the quantitative properties of the service (amount of agreed usage of a resource, number of times a resource may be accessed, length of time a service may be used/must be provided, …). In the figure this is generically described as "parameter".
- In most cases there will be an agreement about the deviation of the parameter, which will not be penalized. In the figure this is described as "corridor". As already mentioned in the previous section, this may include upper or lower bounds or both of them.
- If the customer (or provider) behaviour deviates too much from the agreed "parameter" target (i.e. leaves the corridor), in one or the other way a "penalty function" will be activated to deal with this misbehaviour.

In addition CAM will also depend on the measurements collected from the network. Essentially, these can have three forms:

- A full set of measurements over a certain period of time: These may include different timescales, from very short ones (e.g. bits/second) to quite long ones (e.g. number of downtimes of a network link per month).
- Statistical Probes: If a network management system is not able to cope with the amount of data delivered from the network, which might be the case, if one measures usage within an IP network, then CAM, whose heart consists of an SPRT function block, can also deal with statistical probes.
- Statistically prepared values: It might be that the network already delivers a mean value of a specific series of measurements. In principle this would be preferable if the situation of bullet two is true. In reference [4] IETF provides the means to define performance metrics via IPPM, describing the exact conditions under which the respective data have to be collected.

The three variants of data delivery warrant an additional input block, namely data distribution parameters (see figure 2), telling the CAM how a specific set of data is to be interpreted.

Finally, as output, CAM delivers the Cumulus Points as the actual balance over a certain time interval. They are the abstract evaluation of the way the SLA under investigation has been fulfilled.

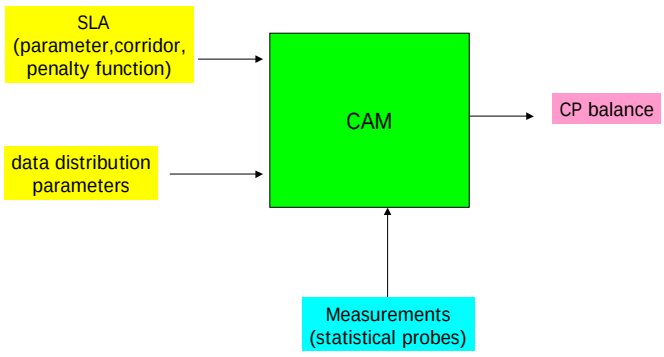

Figure 2: Architecture of the CAM

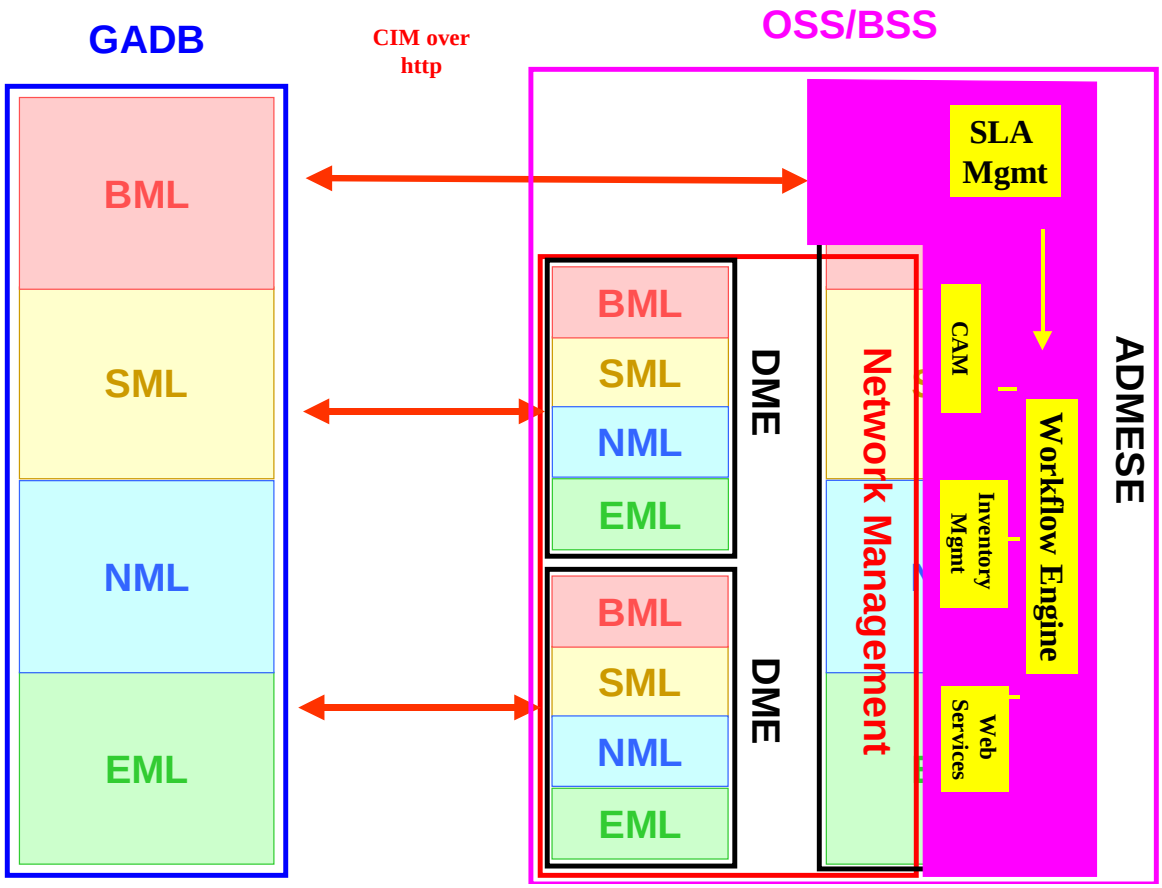

Figure 3: Placement of the Cumulus Assessment Module within the extended TMN-Model according to [2].

Figure 3 shows the positioning of CAM within the enhanced Telecommunication Management Network (TMN) architecture [2] we have developed based on the original ITU-T recommendation [1].

In [2] the following situation is described: In addition to the original management layers described in [1], the Element Management Layer (EML), the Network Management Layer (NML), the Service Management Layer (SML) and the Business Management Layer (BML), Domain Management Entities (DMEs) and an All Domain Management Entities Supervisor Entity (ADMESE) were defined. Whereas the DMEs are the management units responsible for the management of one technical domain (MPLS, ATM, DiffServ, …), the ADMESE is the entity, which has an end-to-end view over the complete managed network. In [2] it is shown that because of this complete overview the ADMESE is in the unique position to provide both operational support in the sense of OSS and business support in the sense of BSS. Figure 3 shows that the SLA Management block using the Workflow Engine as the task distributor could start the CAM. Either the Workflow Engine already passes on the SLA parameters for CAM or it just hands down a pointer into the General Application Database (GADB), which contains the relevant information. CAM also interfaces with the network management part, which allows it to access the necessary measurement data. The results, id est the CP balance, will be stored in the BML section of the GADB.

The following chapter will show two possible applications, one from the OSS area and one from the BSS area.

4. EXAMPLES

One of the main advantages of CAM is its ability to deal with spare as well as with mass data. This section argues that this is a prerequisite for any general SLA evaluation mechanism, moreover we demonstrate the huge variety of different areas the CAM can be used in.

4.1 Storage Area Network

The first example is concerned with the usage of a storage area network (SAN). For the purposes of this chapter a SAN can be regarded as a high-speed special-purpose network (or sub-network) that interconnects different kinds of data storage devices with associated data servers on behalf of a larger network of users. Suppose that the provider of a SAN has Service Level Agreements, which state at which times what amount of data is allowed to be stored by a specific customer. In addition it is agreed what happens, if the frequency of storage or the capacity of the stored data are higher than agreed. Then measurement of the storage frequency and the amount of capacity used may yield two different flavours of CPs, if either measurement is over the agreed limit. One of the reactions might be the putting in place of access control mechanisms, another one the charging of additional money. This application would certainly be part of a business support system.

4.2 Security Management

The second example is concerned with security issues. Suppose that a big company has a networking group, which negotiates SLAs with all other groups within the company (law, finance, service, …) concerning the usage of the information infrastructure. Suppose further that one specific part is the security clearance and the associated behavioural procedures. In this respect a member of the company might have different security roles for different areas of his or her activity. Depending on the role the user assumes, security breaches might be worth one or more CPs per misbehaviour instance. The evaluation of CPs per role might result in the downgrading of the security clearance of the respective person and perhaps the loss of certain responsibilities. In this case we are clearly looking at an operations systems support application.

5. SUMMARY AND OUTLOOK

This paper has dealt with a rather general approach for SLA monitoring, which has been validated by two examples from areas as different as security management and management of Storage Area Networks. The basic idea is to apply a specific statistical test, i.e. the reverse formulation of the Sequential Probability Ratio Test (SPRT), to come to a firm conclusion about the fulfillment of an agreed SLA, independent of the quality and granularity of available measurement data. It has been shown that the resulting Cumulus Assessment Module CAM may be easily integrated into any network management architecture and provides a robust and efficient mechanism to assess a wide variety of contracts between provider and end-user. Note finally that the use of CAM is not restricted to SLA assessment, but may include further useful evaluations like user profiling and service pricing. Apart from that, current and future work also includes a variant of CAM which allows the symmetric handling of both user and provider misbehaviour.

REFERENCES

[1] ITU-T Recommendation M.3010: *Principles for a Telecommunications Management Network*, 1996.

[2] [2] W. Haidegger: *„ The All Domain Management Entities Supervisor Entity as Model for Operations System Support";* Applied Telecommunications Symposium, 04-2004, Washington, FL, USA.

[3] W. Haidegger: „A Model for the Usage of Synergies Between Edge Pricing and End-to-End Performance Management"; CCCT, 08-2003, Orlando

[4] Paxon et al: *Framework IP Performance Metrics*, RFC 2330, May 1998

[5] P. Reichl, P. Flury, J. Gerke, B. Stiller: "How to Overcome the Feasibility Problem for Tariffing Internet Services: The Cumulus Pricing Scheme". *Proc. of IEEE International Conference on Communications (ICC 2001)*, Helsinki, Finland, pp. 2079-2083, June 2001.

[6] P. Reichl, D. Hausheer, B. Stiller: "The Cumulus Pricing Model as an Adaptive Framework for Feasible, Efficient and User-friendly Tariffing of Internet Services". Journal of Computer Networks, Elsevier, 2003.

[7] A. Wald: Sequential Analysis. New York (Wiley), 1947.

[8] M. Fisz: Wahrscheinlichkeitsrechnung und mathematische Statistik. Berlin 1989.

[9] P. Reichl: Contraception Management by STORCH: The Sequential Test for Ovulation Reckoning and Contraception Handling. Unpublished manuscript, 1995. Available at URL: http://userver.ftw.at/~reichl.

Peter Reichl has studied mathematics and philosophy in Munich and Cambridge (UK). His past affiliations include Aachen University of Technology, Bell Labs (Murray Hill) and ETH Zurich, where he has finished his Ph.D. thesis on Internet pricing. Currently he is working as a key researcher at the Telecommunications Research Centre Vienna (ftw.). His current research interests include pricing and QoS in the Internet, Internet Economics, QoS in heterogeneous mobile environments and user aspects of telecommunication services.

Wolfgang Haidegger has graduated in mathematics from University of Vienna and holds a Ph.D. in theoretical physics from the same university. He has been with Siemens AG Österreich for the last thirteen years, participating and leading projects in the areas of broadband telecommunications. His current research interests include end-to-end management systems, traffic engineering and mathematical models for long-term performance analysis of high-speed networks.