Service Level Indication

A proposal for QoS monitoring in SLA-based multidomain networks

S. D'Antonio[2], M. D'Arienzo[1], M. Gargiulo[2], S. P. Romano[1], and G. Ventre[1,2]
1: Dipartimento di Informatica e Sistemistica - Università degli Studi di Napoli "Federico II" – Via Claudio, 21 – 80125 Napoli – ITALY
e-mail: {maudarie, spromano, giorgio}@unina.it

2: Laboratorio ITEM - Consorzio Interuniversitario Nazionale per l'Informatica (C.I.N.I.) – Via Diocleziano, 328 – 80124 Napoli – ITALY
e-mail: {salvatore.dantonio, mauro.gargiulo}@napoli.consorzio-cini.it

Abstract: The offering of QoS based communication services has to face several challenges. Among these, the provisioning of an open and formalised framework for the collection and interchange of monitoring and performance data is felt as one of the most important issues to be solved. Indeed, this is true in scenarios where multiple providers are teaming (intentionally or not) for the construction of a complex service to be sold to a final user, like in the case of the creation of a virtual private network spanning multiple network operators and infrastructures. In this case, failures in providing certain required levels in the quality parameters should be dealt with an immediate attribution of responsibility across the different entities involved in the end-to-end provisioning of the service. But also in cases apparently much simpler, for example when an user requires a video streaming service across a single operator network infrastructure, there is a demand for mechanisms for the measurement of the received quality of service across all the elements involved in the service provisioning: the server system, the network infrastructure, the client terminal and the user application. It is clear that this is a complex problem, involving different technologies, disciplines and research areas. In this paper, starting from the ongoing work in the definition of standard interfaces for the Quality of Service negotiation (Service Level Agreements) and control (Service Level Specifications), as well as from the work ongoing in the IPFIX and IPPM working groups from the IETF, we introduce a new document specifically for delivering monitoring information to user applications. We called such a document Service Level Indication. We here aim at sketching a possible starting point for a research discussion.

C. McDonald (ed.), *Converged Networking*

Key words: Quality of Service, Performance Monitoring, Service Level Agreements, Service Level Indication, Service Negotiation

1. INTRODUCTION

In the last few years, a new approach to QoS issues has aimed to the definition of new network architectures, in which the nodes and elements are capable of providing deterministic or statistical guarantees for the transmission of data. Such an approach is based on the definition of mechanisms and algorithms for admission control and resource management (*Pro-active QoS*) and, albeit very complex, it is producing interesting results in terms of definition of realistic, deployable models of novel network architectures.

We expect that the number and the diffusion of architectures with Pro-active QoS, will grow in the near future, as also shown by the several research groups working on these issues.

In such a scenario, users can choose among different levels of Quality of Service in order to best meet their application and pricing constraints. The service and its delivery quality are negotiated through a contract, named Service Level Agreement (SLA), between the user and a service provider. Such a service provider is intended as an entity capable to assemble service contents, network and server side resources. These resources could either be located in different domains or belong to different providers [6].

One of the issues that such a scenario can bring out is related to two aspects, only apparently un-related: first, the auditing of the actual satisfaction of current SLA with the service provider; second, the dynamic re-negotiation of the service agreements themselves.

As far as the first aspect, we can reasonably forecast that as soon as users start to pay for communication services with QoS-related guarantees (e.g. service availability), it will be required to verify whether or not the conditions specified in the SLA are effectively met by the provider. With reference to the second aspect, indeed, re-negotiation of QoS has been always accepted as an important service in performance guaranteed communications. Nevertheless, this aspect has been considered in practice as marginal since it is generally linked to user-expectations, related to changes in the perceptive satisfaction for the delivered service. However, re-negotiation is also related to different issues, strongly connected to critical problems such as the efficiency of network resource allocation, the end-to-end application level performance, and the reduction of communication costs. For example we might consider a scenario where the quality of service received by a distributed client-server application can be seen as influenced by several factors: the network performance, the server load and the client

computational capability. Since those factors can be varying in time, it is logical to allow applications to modify Service Level Agreements on the basis of the QoS effectively achievable and perceivable at the application layer. We therefore believe that the possibility to modify the existing QoS based agreements between a service provider and the final user will assume an important role in Premium IP networks [5]. Such networks provides users with a portfolio of services thanks to their intrinsic capability to perform a service creation process while relying on a QoS-enabled infrastructure. Furthermore, applications can impose additional requirements to Premium IP networks because of this possible dynamic re-negotiation of SLAs.

Currently, applications willing to monitor the received QoS, have to be capable to collect performance data by themselves, for example by on-line measurements of the communication performance achieved during data exchange. This can be done by inserting software probes in the application code, or by using feed-back information made available by specific protocols such as RTP (*Real Time Protocol)* [2]. This approach, however, is quite problematic, since i) it requires specific development of code dedicated to such on-line measurement; ii) it is capable of only a high-level identification of possible end-to-end performance problems, with no detailed information on their original causes.

In this document, we illustrate a proposal that we believe might represent the basis for a novel monitoring and control architecture for the collection and distribution of performance data, linked to the deployment of a distributed computing infrastructure. The idea which paves the ground to our proposal is mainly based on the definition of an information document, that we defined *Service Level Indication* (SLI). As this document will disclose, Service Level Indication is to be produced with the cooperation of all distributed components in order to obtain a detailed picture of the level of service that is currently offered. The SLI-based monitoring architecture is quite simple in its formulation; nonetheless it brings in a number of issues, related to its practical implementation, to its deployment in real-life scenarios, and to its scalability in complex and heterogeneous networked infrastructures. Some of these issues will be highlighted in the following, where we will also sketch some possible guidelines for deployment, together with some pointers to potential innovative approaches to this complex task.

The paper is organized as follows. In section 2 we introduce QoS monitoring issues in SLA-based infrastructures, underlining business aspects of such a service. A possible monitoring framework is presented in section 3. In section 4 we explain the data export process. Finally, section 5 provides some concluding remarks to the presented work.

2. MOTIVATION

Computer networks are evolving to support services with diverse performance requirements. To provide quality of service (QoS) guarantees to these services and assure that the agreed QoS is sustained, it is not sufficient to just commit resources since QoS degradation is often unavoidable. Any fault or weakening of the performance of a network element may result in the degradation of the contracted QoS. Thus, QoS monitoring is required to track the ongoing QoS, compare the monitored QoS against the expected performance, detect possible QoS degradation, and then tune network resources accordingly to sustain the delivered QoS [3].

The adoption of a pro-active behaviour with reference to the provisioning of QoS communication, introduces several issues that we believe are worth of investigation for their impact on future SLA-based Premium IP networks. Among them, a major role is played by dynamic negotiation between applications and networks for the selection of an adequate quality of service. In fact, when an user contacts, for example, the provider of a video-delivery service, he will expect to negotiate the access to the service and its price. In Premium IP networks such price will be influenced not only by the multimedia content of interest to the user but also by the QoS level that will be required in the delivery of the content itself across the network. In this context, we have a negotiation for a service where the user will ask for a content and for a certain quality of its delivery, and the service provider will answer with a price. Actually, we might also expect that such initial negotiation will be performed automatically between the client application on one side and the provider on the other one. Additionally, such a negotiation might happen at service subscription time rather than upon service invocation.

2.1 Requirements

In an architecture capable to dynamically negotiate a rich portfolio of services, it becomes of primary importance the availability of mechanisms for the monitoring of service performance parameters related to a specified service instance. This capability is of interest both to the end-users, as the entities that 'use' the service, and to the service providers, as the entities that create, configure and deliver the service.

In the case of SLA-based Premium IP networks, QoS monitoring information should be provided by the network to the user application, by collecting and appropriately combining performance measures in a document which is linked to the SLA itself and which is conceived following the same philosophy that inspired the SLA design: i) clear

differentiation of user-level, service-level and network-level issues; ii) definition of lean and mean interfaces between neighbouring roles/components; iii) definition of rules/protocols to appropriately combine and export information available at different levels of the architecture.

Summarizing the above considerations, we can state that monitoring seems to offer excellent novel opportunities to service providers who do have the possibility to effectively engineer their network infrastructures. By this way, they can exploit at its best the unprecedented potential disclosed by dynamic service creation and delivery. In this case, "SLA monitoring" may be seen in the light of a more general activity related to "Network monitoring", and this is mainly due to the following objectives:

- **usage-based accounting**: existing business models for selling IP-based services exploit accounting mechanisms based on time or volume. Therefore, data and measures related to different flows are necessary in order to define service costs as well as prices to be presented to the end-user. Accounting may be performed per user or per user group and related criteria can take into account several parameters, such as time of day, used (label switched) path, class of service. Moreover, accounting may be applied to IP basic or advanced services.
- **traffic profiling**: traffic profiling is a process of characterizing IP flows and flow aggregates by using a model based on some key parameters of the flow, such as flow duration, volume, time and used protocols. Results provided by traffic profiling activity become a fundamental starting point for network planning, network dimensioning, trend analysis, business models synthesis. Furthermore, measurement statistics and accuracy heavily depend on the particular traffic profiling goal. Typical input data needed for traffic profiling are the distribution of used services and protocols in the network and the amount of packets of a specific type. Since objectives for traffic profiling can vary, this activity requires a highly flexible measurement infrastructure, especially concerning the options for measurement configuration and packet classification.
- **traffic engineering**: the goal of traffic engineering is the optimization of network resource utilization and traffic performance. Such an objective is achieved through methods for measurement, modeling, characterization and control of a network. Parameters, such as link utilization, load between specific network nodes, number, size and entry/exit points of the active flows are monitored and measured because collected data analysis can improve the traffic engineering effectiveness. For this reason, the measurement infrastructure has to be able to adapt to network topology changes and provide the end user or application with information either on-line or off-line, depending on both the activities to be performed and the network actions to be taken.

- **attack/intrusion detection**: in this case we refer to a generic network monitoring activity rather then traffic measurement. Flow information capturing plays an important role for network security, both in the detection of security violations, and in the subsequent defense counteractions. In case of a Denial of Service (DoS) attack, flow monitoring can allow detection of unusual load situations or suspicious flows. In a second step, flow analysis can be performed in order to gather information about the attacking flows, and for deriving a defense strategy. Intrusion detection is a potentially more demanding application which would not only look at specific characteristics of flows, but that may also use a stateful packet flow analysis for detecting specific, suspicious activities, or unusually frequent activities.

2.2 Monitoring as a service

In a Premium IP network, the service provisioning is the result of an agreement between the user and the service provider, and it is regulated by a contract. The SLA is the document resulting from the negotiation process and establishes the kind of service and its delivery quality. The service definition stated in the SLA is understood from both the user and the service provider, and it represents the service expectation which the user can refer to. Such SLA is not useful to give a technical description of the service, functional to its deployment. Therefore, a new and more technical document is needed. The Service Level Specification document, as described in [1], derives from the SLA and provides a set of technical parameters with the corresponding semantics, so that the service may be appropriately modelled and processed, possibly in an automated fashion.

The SLS can also be used by different providers, in order to cooperate in the fulfilment of the service: this issue, which is mainly related to the interdomain scenario, requires that a thorough definition of the protocols and mechanisms involved in the exchanging of information between each pair of peering entities along the service delivery chain is provided.

In order to evaluate the service conformance to specifications reported in SLA and SLS documents, we introduce a new kind of document, the *Service Level Indication* (SLI).

By considering the different abstraction levels (user, service, network), it is possible to distinguish among three kinds of SLIs (Figure 1):

- *Template SLI*, which provides a general template for the creation of the documents containing the monitoring data associated to a specified service;
- *Technical SLI*, which contains detailed information about the resource utilization and/or a technical report based on the SLS requirements. This

document, which pertains to the same level of abstraction as the SLS, is built up by the resource owners;

- *User SLI*, i.e. the final document forwarded to the user; it contains, in a friendly fashion, information about the service conformance to the negotiated SLA. The User SLI is built up, by the service provider, on the basis of both the SLS, the Template SLI and the Technical SLI.

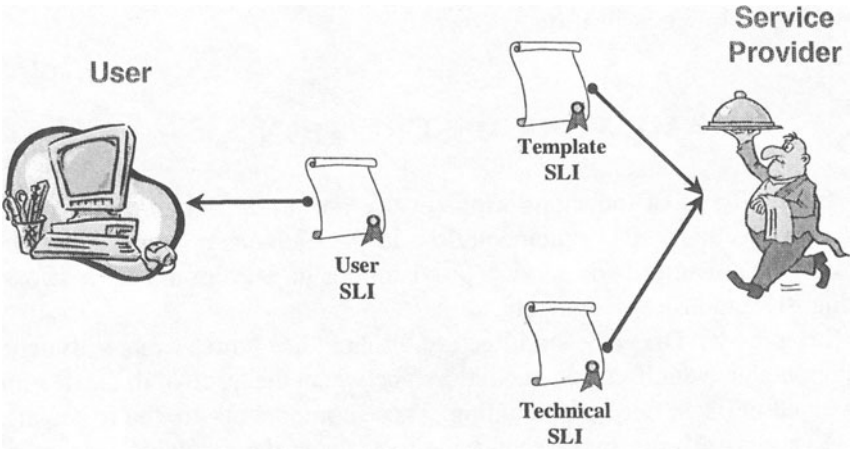

Figure 1. The three SLI documents and their scope

The service monitoring has to be finalized to the delivery of one or more SLI documents, either User SLI or Technical SLI. In the SLI issue, multiple entities are involved, as network elements, content servers, and user terminals. Involving all these elements has a cost: it is due to the usage of both computational and network resources, needed for information collection, analysis and distribution. This cost depends on both the number of elements involved and the information granularity.

For example, we can consider a VPN service spanning dozens of hosts located all over the world. For such a service, the monitoring might be able to report detailed information about throughput, delay, packet loss, availability, etc.

From this point of view, monitoring may be under all perspectives considered as a service, for which ad hoc defined pricing policies have to be specified and instantiated. More precisely, drawing inspiration from the concept of *metadata,* we might define the monitoring as a *metaservice,* i.e. a 'service about a service'. This definition is mainly due to the fact that it is hardly conceivable a monitoring service per se: monitoring is strictly linked to a pre-existing service category, for which it provides some value-added information. Therefore we won't consider a standalone monitoring service,

but we will rather look at it as an optional clause of a traditional service, thus taking it into account in the SLA negotiation phase.

A similar consideration can be made for the SLA re-negotiation, which is a mandatory requirement whenever discrimination between static and dynamic SLAs occurs. Starting from the monitoring results, in fact, a user can establish whether the negotiated QoS is appropriate or it has to be refined, thus possibly triggering the re-negotiation of the current SLA. Similarly, the service provider can ask for SLA re-negotiation in order to optimize resource utilization.

3. FRAMEWORK DESCRIPTION

On the basis of the monitoring service definition, we aim at describing the roles needed to its implementation. In the Cadenus project [7], we find a possible architecture for service provisioning in a Premium IP network by using SLA and SLS documents.

In the CADENUS architecture, there are three main functional components, which act as mediators, between the user and the resources involved in the service provisioning. These components are the following:

- **Resource Mediator(s)**: they have to manage the available resources, by configuring the involved nodes. Each service can concern different domains and then different Resource Mediators. Now, the Resource Mediator also have to gather basic monitoring information, and produce the Technical SLI.
- **Service Mediator(s)**: they are in charge to create the service as required from the user, using the resources made available by one or more Resource Mediators. Each SLA (and related SLS) refers to a single Service Mediator. Now, by using information stated in SLS and Technical SLI, they also have to evaluate SLA fulfilment and produce the User SLI.
- **Access Mediator(s)**: they act as service brokers, then they have to provide the user interface, with AAA functionality. Now, the Access Mediator receives the User SLI from the Service Mediator and returns this information to the user according to its profile.

4. DATA EXPORT

Till now we have made a bird's eye view analysis of current definitions and issues related to QoS measurement and monitoring. However, in the context of SLA-based services the following innovative aspect has to be

considered: in order to allow users, service providers and network operators to have information about QoS parameters and network performance the need arises to export data collected by measuring devices. For this reason, the concept of data model has to be introduced. Such model describes how information is represented in flow records. As stated in [4], the model used for exporting measurement data should be flexible with respect to the flow attributes contained inside reports. Such reports can be obtained in two possible ways: push mode and pull mode. In push mode, the measuring device decides without an external trigger on when to send a report on measured flows. In pull mode, report sending is triggered by an explicit request from a data collector or some other receiver of flow records. Furthermore, the measuring device should be able to report measured traffic data regularly according to a given interval length and when a specific event occurs.

Since the service and its quality are perceived in a different fashion depending on involved actors (end user, service provider, network operator), there is a need to define a number of documents, each pertaining to a specific layer of the architecture, suitable to report information about currently offered service level. As far as data reports, we have defined a set of new objects aiming at indicating whether measured data, related to a specific service instance, are in accordance with the QoS level specified in the current SLA.

Having in mind the Cadenus architecture, it is possible to identify the components responsible for the creation of each of the monitoring documents (Figure 2). Such documents are then exchanged among the components as described in the following, where we choose to adopt a bottom-up approach:

1. Upon Service Mediator request, the Resource Mediator builds up the Technical SLI document on the basis of data collected by the measuring devices. The fields it contains are directly derived from those belonging to the SLS and are filled with the actual values reached by the running service. The resulting document is sent to the Service Mediator.
2. By means of the Technical SLI received from the Resource Mediator, the Service Mediator is capable to evaluate the received service quality conformance with respect to the requests formulated through the related SLS. It can be interested in such information both for its own business and in order to gather data for the compilation of a complete report in case of user request.
3. Upon user request, the Service Mediator, exploiting data contained in a Technical SLI, produces a further report indicating the QoS level as it is perceived by the end user. The document it is going to prepare is derived from a service specific template (the so-called SLI Template), which provides an abstraction for the measurement results in the same way as

the SLA Template does with respect to the service parameters. Such a document, hereby called User SLI, is ready for delivery to the end-user.

4. The Access Mediator receives the User SLI from the Service Mediator, puts it in a format that is compliant with both the user's preferences and the user's terminal capabilities and forwards it to the end-user.

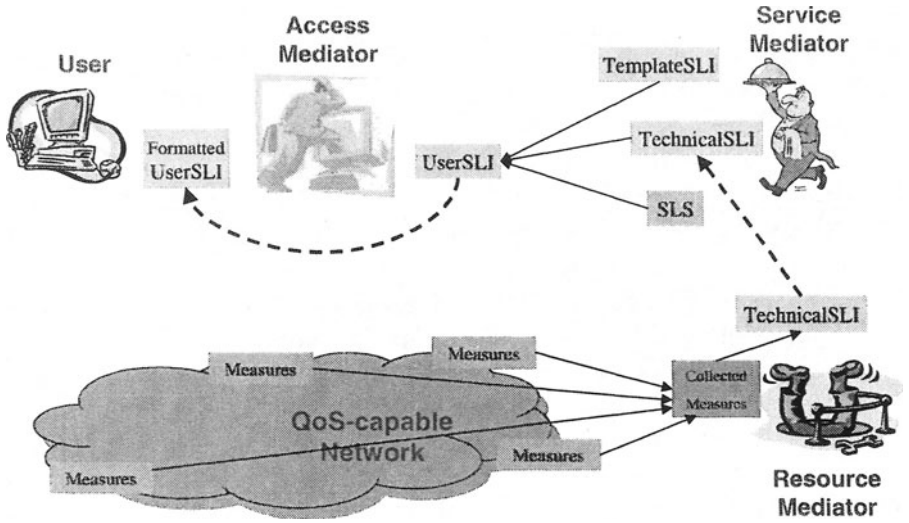

Figure 2. Information passing in a Cadenus-like Monitoring Framework

5. DISCUSSION AND CONCLUSIONS

The need for technologies and architectures for the provisioning of seamless monitoring and performance data in future QoS based networks is under everybody's eyes. It is a need related to the concrete requirements of a wide number of different players: application designers, content providers, service providers, network operators, end users and last but not least, third party monitoring agencies. However, it is also clear that the provisioning of such feature is particularly complex and critical, since it involves the coordination and orchestrated operation of a large number of elements, separately owned and managed along what we have called the provisioning chain from the service location to the end user. We therefore foresee a number of issues to be faced before this problem can be solved. For some of them we provided a possible solution, while for the others the discussion is still wide open.

We briefly mention here the main facets of the general issue of QoS monitoring, focusing on the networking infrastructure.

First of all, the collection of monitoring data from the network elements. This issue is clearly related to both technical and business aspects.

As far as the first ones, the work ongoing in the area of policy based management of network elements is providing a technical framework in which the control and configuration of network nodes will be much more straightforward than that currently achievable through the traditional SNMP based approach. However, it is clear that for global communication infrastructures, involving large number of nodes with a huge number of active connections, we do have a problem of scalability with respect to the collection and delivery of performance data. In spite of this, we believe that there are features in the existing network architectures that might be exploited to reduce at least this problem. For example, in DiffServ based network architectures monitoring of Service Level Agreements can be performed usually per traffic class and not per single traffic flow, and could be normally limited to the ingress and egress points of a domain. More detailed performance data collections (in terms of specific flows or network elements) could be triggered only in the presence of specific demands from the involved parties or in the case of anomalies. We could therefore imagine a scenario where a network operator could regularly broadcast information (in our model a Technical SLI) related to the average performance behaviour of its network infrastructure.

As far as the business aspects, i.e. those related to the business nature of the provisioning of communication services, we can mention here the one we believe is the most important: trust. In global networks, large scale infrastructures will be managed by a multitude of different operators, each of them managing a separate network domain. Quality of Service will therefore be an issue involving a number of parties, each of them responsible only for the service provided in the domain that it directly manages. Such parties will be obliged, at the same time, to compete and to cooperate with peering entities. Can we foresee a scenario where such performance data will be openly (albeit in a controlled way) available? We believe that rather than being an obstacle to the deployment of a common framework for SLA monitoring, trust will be an important trigger for it, if not a prerequisite. In fact, we can expect that no operator will start charging for premium services involving infrastructures owned by others without a formal, standardized way for exchanging performance data about the communication services offered to and received from other operators.

A further issue is related to the devising of a common quality of service measurement framework. It is clear that performance data should be provided in a way that is independent of both the network architecture offering the service and the application service demanding it. Our proposal is a first attempt in this direction.

6. ACKNOWLEDGMENTS

This work has been performed under the partial financial support of the Italian Ministry of Education, University and Research (MIUR) under grants "LABNET2" and "Web Systems with QoS Capabilities". We also gratefully acknowledge access to the results of IST project CADENUS IST-1999-11017 "Creation and Deployment of End-User Services in Premium IP Networks", http://www.cadenus.org.

7. REFERENCES

[1] D. Goderis et al., Internet Draft « Service Level Specification Semantics and Parameters », draft-tequila-sls-02.txt, February 2002.

[2] IETF Audio-Video Transport Working Group, « RTP: A Transport Protocol for Real-Time Applications », IETF proposed standard RFC 1889, January 1996.

[3] Y. Jiang, C.-K. Tham and C.-C. Ko, « Challenges and approaches in providing QoS monitoring », International Journal of Network Management, 2000, 10:323-334.

[4] J. Quittek, T. Zseby and B. Claise, Internet Draft « Requirements for IP Flow Information Export », draft-ietf-ipfix-reqs-02.txt, March 2002.

[5] M. Smirnov et al., « SLA Networks in Premium IP », Deliverable 1.1, IST Project CADENUS, IST 11017, March 2001, www.cadenus.org

[6] Vladimir Smtlacha, « QoS Oriented Measurement in IP Networks », CESNET technical report number 17/2001, December 2001.

[7] G. Ventre et al., « Quality of Service control in Premium IP networks », Deliverable 2.1, IST Project CADENUS, IST 11017, March 2001, www.cadenus.org