Automated SLA Monitoring for Web Services

Akhil Sahai, Vijay Machiraju, Mehmet Sayal, Li Jie Jin, Fabio Casati
HP Laboratories, 1501 Page Mill Road, Palo-Alto, CA 94034
{firstname_lastname}@hpl.hp.com

Keywords: Web Services, SLA, Contracts, specification, measurement, monitoring, modeling, instrumentation

Abstract: Automating SLA monitoring involves minimizing human involvement in the over-all monitoring process. SLA monitoring is difficult to automate as it would need precise and unambiguous specification and a customizable engine that collects the right measurement, models the data and evaluates the SLA at certain times or when certain events happen. Also most of the SLA neglect client side measurement or restrict SLAs to measurements based only on server side. In a cross-enerprise scenario like web services it will be important to obtain measurements at multiple sites and to guarantee SLAs on them. In this article we propose an automated and distributed SLA monitoring engine.

1 Introduction

A web service can be described broadly as a service available via the Internet that conducts transactions. E-businesses set up Web Services for clients and other Web Services to access. They have a Uniform Resource Locator at which they can be accessed and have a set of Interfaces that can be utilized to access them. Web services that are capable of intelligent interaction would be able to discover and negotiate with each other, mediate on behalf of their users and compose themselves into more complex services. This composition could be static or dynamic. Emerging standards such as SOAP, UDDI, and WSDL[1] are steps in this direction. As these web services interact and delegate jobs to each other they would need to create and manage *Service Level Agreements* amongst each other. Service Level Agreements (SLA)s are signed between two parties for satisfying clients, managing expectations, regulating resources and controlling costs. SLA management involves the procedure of signing SLAs thus creating binding *contracts*, monitoring their compliance and taking control actions to enable compliance.

Web Services are being designed, so as to automate e-business on the web. Just as little human intervention is desirable in day-to-day functioning of web services, the same is true for monitoring of service level agreements on these web services. However, SLA monitoring is difficult to automate as it would need a precise and unambiguous definition of the SLA as well as a customizable engine that understands the specification, customizes instrumentation, collects the necessary data, models it in a logical manner and evaluates the SLA at certain times or when certain event happen.

---

[1] SOAP: Simple Object Access Protocol (Microsoft, W3C); UDDI: Universal Discovery, Description, and Integration (Consortia, includes HP); WSDL: Web Services Description Language (IBM, Microsoft, W3C)

Also most of the SLAs are about measurement located at a particular location. There are however two aspects that are specific to web services. The first one being the fact that the web services are being designed so as to work over the internet. The internet is inherently unreliable and even though its alright for document retrieval/dissemination it poses a problem when real business has to be undertaken on it. It is necessary to ensure that the consumer perceives that the provider is adhering to its promised service level agreements. In other words to provide Quality of Experience to the consumer. Also a guarantee that is true at the server side may not be true at the consumer side because of the unreliable nature of the Internet. So, often server side measurements may not hold for client side. These reasons may necessitate client-side measurement. The second aspect that is specific to web services is that they are inherently multi-party in nature. A typical web service will use other web services to perform its task. These web services will have service level agreements with each other. However, a consumer orders to only one of the web service. The other web services work together to fulfill the consumer's order (as shown in Figure 1). An analogy on the Internet is that of an Internet based Bookseller service using well known shipping companies to ship goods physically to clients. As multiple parties are involved service level agreements may have to be guaranteed over activities that span multiple web services (for example a guarantee that a client after she orders the book online will receive a book at her home within 3 days). In such cases, unless the measurements are obtained from multiple locations and aggregated, SLA monitoring cannot be done. In this paper we propose an automated and distributed SLA monitoring engine that enables the above functionality.

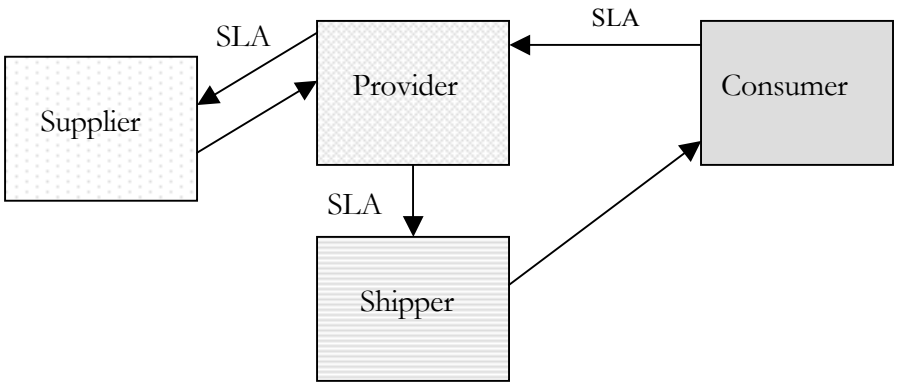


Fig1: Multiple web services cooperating with each other to accomplish a task

The messages that web services exchange with each other in order to execute an end-to-end business goal creates a logical network between these services. As these web services become more and more prevalent, in addition to the job of measuring end-to-end metrics and enforcing end-to-end objectives surrounding a business goal will require another set of messages and protocols to be defined. We envisage management agents installed at an enterprise site managing the relationships of one or more web services it offers with other management agents responsible for the web services of other enetrprises. These management agents are termed business management platform (BMP) agents in our case. For the purpose of this article they will be shown to monitor SLAs between web services and to exchange measurements and protocols for achieving the same.

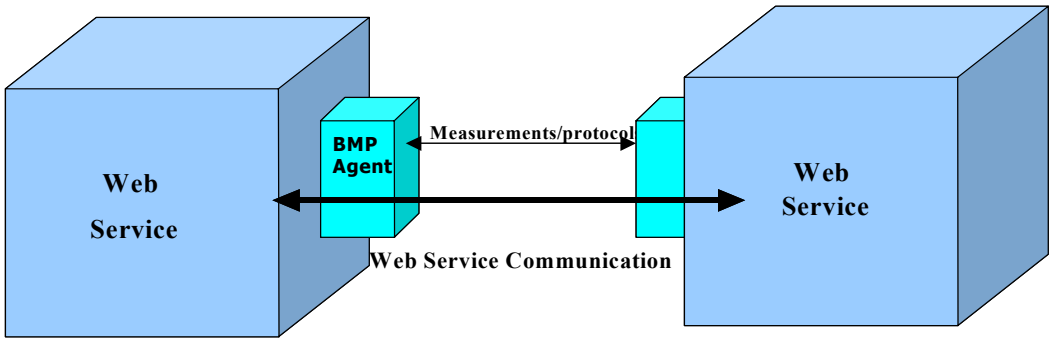


Fig2: BMP Agents at two ends exchange measurements and protocols for management

However, before service level agreements can be specified for web services, it is necessary to understand what all comprises a typical infrastructure and what all can be guaranteed on them.

2 Web Service Infrastructure

A web service infrastructure would comprise of large number of business processes. These business processes will usually comprise of set of activities. Each activity will be handled by either humans (as is the case in work-flow management systems), automated systems (based on legacy systems or state of the art application servers) or some times will be outsourced to external e-businesses. In Figure 3 a simple example of a web service infrastructure is shown. This particular business is set up by PCMaker.com that receives orders from companies/humans interested in buying PCs. It has internal business processes like user authentication, PC manufacturing, preparation of invoices etc. These business processes are defined in terms of WSFL/XLANG.

For some of the PC order parts, it needs to contact it supplier and similarly uses a shipping company to ship the PCs it makes. The PCMaker.com web service has operations, namely login, order_request, Send_invoice, and Send_shipment. It also has other operations, namely Order_parts and Ship_order. These descriptions are captured in Web Service Description Language (WSDL).

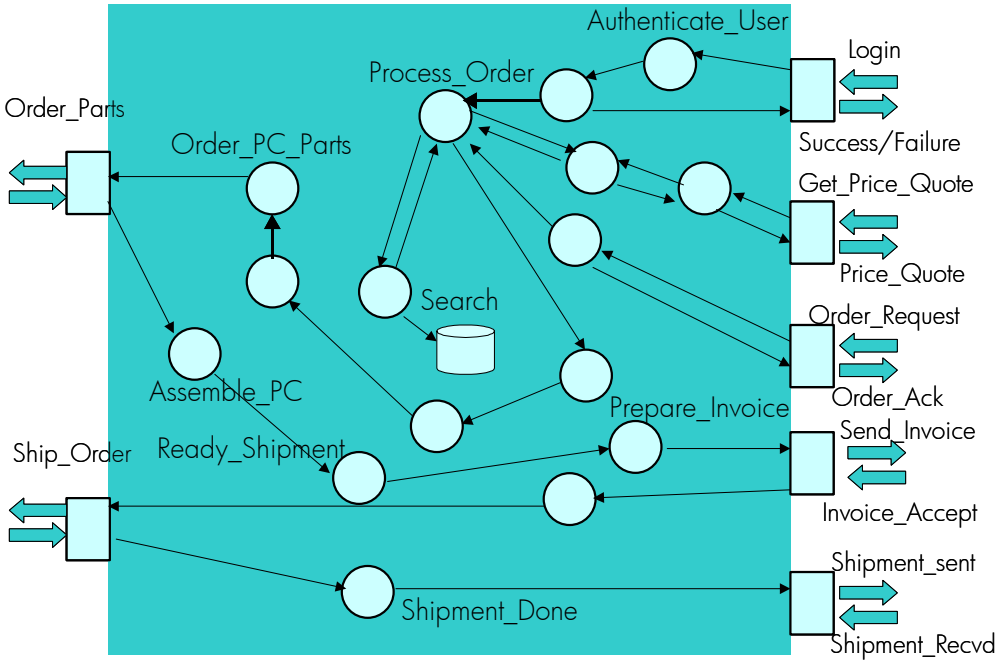


Fig3 A typical web service and business process infrastructure

Behind the logical business processes, web services and operation they support are software that support the web service infrastructure, namely web sites, web server farms, applications servers (Legacy software) and the business processes that are executed either on Process Manager, MQSeries, Web Methods platforms or are plain physical business processes which involve interaction between a disparate set of systems, humans and software.

3 Web Service SLA

Protocols like BTP and ebXML enable web service to web service interactions to be captured through a set of well defined processes. These processes are distinct from the internal business processes as mentioned above. Parts of these processes could be sub-processes defined by standards such as say RosettaNet PIPs. This enables the fact that web services can undertake business by executing an orchestration of business transactions amongst themselves. This involves definition of a combined process between the two partners which in turn is bi-sected according to the roles undertaken by the partners (namely customer, provider). Each party executes the process belonging to their role. These processes involve a particular sequence of invocation of each other's operations through message exchanges between them. The operation and message exchange interfaces are already captured in WSDL descriptions as explained. These processes also interface with internal business processes that are defined in process definition languages like WSFL or XLANG.

While, WSDL introduces concepts such as messages, operations, ports, and end points – which are useful for describing the operations of any web service, WSFL introduces the notion

of activities and process flows. So, one way to create a flexible SLA formalization is to build upon these concepts. In other words, one can create a flexible SLA formalization by associating "quality metrics" to the formalizations that are already defined in WSDL, WSFL or BTP/ebXML. Here are some examples that show how such association can be done.

- Response time of a web service operation.
- Average response time between two set of messages
- Response time of a process flow.
- Average response time of a set of process flows of a particular type
- Security of an operation.
- Number of times an activity is executed in a flow.
- Cost of executing an operation.
- Availability of an end point.
- Recoverability of an end-point

The concept of service level agreements and guarantees is missing as yet in the world of web services and business transactions. We introduce the concept of SLAs/contracts amongst web services in this article. An SLA has a set of Service Level Objectives (SLOs) as specified.

A typical contract between a company manufacturing PCs (say PCMaker.com) and a company buying PCs (PCBuyer.com) for a period of 6 months will be as follows:

SLO1: PCMaker's e-procurement system *will be available to Ford, Monday to Friday from 9AM-5PM, 99.9 % of the time*

SLO2: PCMaker shall deliver the ordered goods *on an **average** within 10 days of the receipt of a purchase order*

SLO3: PCMaker shall invoice PCBuyer for any goods ordered *within 6 hours*

SLO4: Payment of goods by PCBuyer shall be done ***always** within forty-five days of the receipt of invoice from PCMaker.*

Each SLO has a functional part (that refers to a system, endpoint, a process, or a set of processes…) and a guarantee part (italicized) applied on the functional part. The guarantee is on a system, a particular instance of a construct (process/operation/message..) or on a set of such constructs. SLA monitoring involves monitoring whether these guarantees on the functional parts are being met.

In order to automate SLA monitoring, we propose a specification language that enables definition of precise and flexible SLAs, and is described in detail in section 3.1. Section 3.2 describes the instrumentation aspects that enables correlation of web service and business process data. The Business Management Platform Agent (BMP Agent) that automates and distributes the SLA monitoring process is described in section 3.3. In section 3.4, the implementation details of the BMP Agent are described.

3.1 SLA specification

The first enabler for automated SLA management is a **flexible** but **precise** formalization of what an SLA is. The flexibility is needed since we neither completely understand nor can anticipate all possible SLAs for all the different types of web service providers. This will also help create a generic SLA management system for managing a range of different SLAs. The precision is essential so that an SLA management system can unambiguously interpret, monitor, enforce, and optimize SLAs.

Examples of the lack of flexibility and precision in existing SLA formalizations are discussed in [1]. Detailed explanation of how we have addressed flexibility and precision in coming up with SLA formalization are also presented in [1]. Below is a summary of the formalization. A point to note is that the SLA specification is quite generic and is independent of the domain it is applied to (in this case that of web services).

An SLA is specified over a set of data that is measurable. An SLA typically has a date constraint (start date, end date, nextevaldate) and a set of Service Level Objectives (SLOs). An SLO in turn has typically a day–time (Mo-We, 6:00PM-8:00 PM) constraint and a set of clauses that make up the SLO. A clause is based on measured data. This is referred to as a *measuredItem*. A measuredItem can contain one or more *items*. A measuredAt element determines where the measurements are taken (provider, consumer side). A clause evaluation is triggered either when an event happens, e.g. say a message arrives, an operation completes or at a fixed time, say at 6PM. We call this an *evalWhen* component of an SLO. Once the evalWhen trigger arrives, a set of samples of measuredItem are obtained applying a sampling function. The *evalOn* component determines how this sample is computed. The sample set is a constrained set of measured data that is constrained by the evalOn component. Examples of evalOn components may be a number or a time period, e.g. the 5 longest running transactions, or all the samples for last 24 hours. A function (*evalFunc*) is thereafter applied on the sample set so obtained. An example of evalFunc would be average response time function < 5 ms. The evalFunc[2] must be a mathematical function that is expressible in terms of its inputs and logic. The following grammar shows a portion of this formalization.

```
SLA  = dateconstraint SLO*

Dateconstraint = startdate enddate nextevaldate

SLO = daytimeconstraint clause*

Daytimeconstraint = Day* time

Clause = measuredItem evalWhen evalOn evalFunc evalAction

MeasuredItem = Item*

Item  = measuredAt constructType constructRef
```

[2] The evalFunc could be expressed in MathML or SQL or any other functionally complete language

As an example, a clause like *At 6 PM the Average response time for the 5 longest running bookbuy transactions measured on the client side should be < 5 ms* can be broken up into a, measuredItem (Item:bookbuy transaction, measuredAt:Consumer), evalWhen (at 6PM), evalOn function (set of 5 longest running transactions), the evalFunc (average response time < 5 ms) and evalAction (Notify administrator). The complete set of examples of how complex SLAs can be represented in it are presented in [1].

3.2 Instrumentation

In order to ensure that guaranteed SLAs can be evaluated and their compliance measured, it is necessary that raw measurement data be collected about the managed system. This managed data is obtained through instrumentation of processes, activities that are executed, and messages that go in and out of the e-business infrastructure.

3.2.1 Instrumenting the web service

It is necessary to interfere with message exchanges among web services in order to collect information about the interactions with business partners. An acceptable solution should not impose any modifications or limitations on existing web services. Since SOAP is rapidly becoming the preferred standard for web service interactions, we assume SOAP messages are used among web services in order to submit request and response messages. We have implemented a small proxy component tries to capture incoming and outgoing messages, and records data about the message exchanges, then forwards the captured messages to the actual recipients. We have considered various alternatives for easily attaching a proxy component to existing web services in order to listen to incoming and outgoing messages: port sniffing, server-side filters (Microsoft's ISAPI, or Netscape's NSAPI), API provided by web services themselves, and modification of SOAP toolkit. Since SOAP is widely accepted for message exchange, port sniffing and server-side filters are not suitable, because the message contents are encrypted by SOAP toolkit. Most web services do not provide an API for controlling or querying about their activities due to security issues or simply because the web service developers did not feel any need for such interfaces. Consequently, we have chosen to keep track of message exchanges among web services by modifying SOAP toolkit.

The most popular implementations of SOAP toolkit share common components, called routers. SOAP routers receive the messages from SOAP clients and submit them to the receivers. SOAP toolkit encrypts the message at the sender site, and decrypts it only when it reaches the receiver's site. A proxy can be easily attached to SOAP toolkit routers with minor modifications to the toolkit. This is the most appropriate way to automatically attach a proxy in order to capture SOAP messages and collect information from those messages. It does not require any modifications to existing web services, and does not require re-compilation of existing SOAP toolkit installation. We used this approach for collecting data from SOAP message exchanges among web services.

In order to correlate individual message exchanges with each other, we use the notion of Global Flow (GF) as described within our assumptions above. The GUID is used for keeping track of a GF. Every time our proxy component catches a message that is exchanged between

web services, it first checks whether a GUID exists. If a GUID does not exist in the message, the proxy inserts a GUID into SOAP header of the message. All web services and other software components propagate the GUID in their communications. Consequently, our proxy components that are attached to SOAP toolkits at business partner sites can easily figure out which SOAP message is sent in the context of which previous messages.

3.2.2 Instrumentation of business process

Since activities of web services are automated by business processes at the back-end, it is necessary to collect data from those software components in order to gather detailed information about internal activities of a business, and correlate those internal activities with external message exchanges. As we indicated among our assumptions, most business process management systems log data about internal business process executions into a raw log file or database. For example, HP Process Manager (HPPM) logs execution data into a raw file, which is then uploaded into database tables by a dedicated process. A proxy component can be configured in order to read logged data from proper database tables. This component can also correlate the message exchanges with internal process executions using the GUID that is passed through all web services and their back-end software components.

3.3 SLA Monitoring

As minimal human intervention is desirable in web services it is necessary to create monitoring engine that can take care of a variety of specifications and monitor the necessary management data. We believe that the SLA formalizations described above are precise enough to be able to create or customize an SLA monitoring engine on the fly. To simplify the discussion, we will describe the details of the engine as if it manages a single SLA between two services. Such an engine has then two components – one on the service provider side and one on the service consumer side. Extending our notion to a large number of SLAs requires that the engine keep track of the state of multiple SLAs simultaneously, and be able to relate each measurement to one or more affected SLAs. Extending our notion of two services to a large number of interacting services requires the engine's components to take the dual role of acting as both "service providers" in some SLAs and as "service consumers" in some SLAs.

The instance data so collected has to be modeled in the high performance database and a data warehouse so that service level agreements can be monitored on top of the modeled data. The high performance database is updated for every transaction instance data that is received. The data warehouse is updated at regular intervals of time for keeping the data for a longer period of time.

3.3.1 SLM Engine

The SLM Process Controller executes the management processes for the SLM engine. These management processes are distinct from the business processes that are internally executed in the web services infrastructure as discussed in section 1. These management process flows are created and managed for a variety of purpose. These flows are defined in WSFL and are exposed to other BMP agents through WSDL specification of their own. These BMP agents thus can initiate management related conversation with each other. The BMP agent process

controller executes the *SLA monitoring process flow* for undertaking SLA evaluation and reporting.

As the specification typically has startdate, enddate, daytimeconstraint, evalWhen, evalOn and evalFunc components to it, each of these constitutes a generic component that can be used by our SLA Management engine. In addition, we have also identified the most common variants of these generic components, which can be readily parameterized by the engine for a large number of possible combinations of SLAs. Using a new, evalWhen, evalOn, or evalFunc component in an SLA requires an administrator to first develop such a component within the framework of our engine and then to add it to the engine.

The *model generator* receives the WSDL/WSFL specifications and creates a model of the web service in the *model repository*. All the measurements collected from the web service (e.g., ongoing conversations, performance measurements, etc) are attached to this model. The instrumentation in the web service is responsible for collecting these measurements and passing them on to the *management handler* to be stored in the model repository. If the measurements are collected on the client side (as determined by the measuredAt components of the items in SLA clauses), then the *communicator* is responsible for receiving the measurements and storing them into the repository. SLM Engine process controller receives the SLA executes a monitoring process flow (as shown in figure 8, and explained in next section) and accordingly informs the SLA customizer which in turn customizes the alarms at the Alarm Manager (depending on the evalWhen and dateconstraint components). The Alarm Manager comprises of the SLO Validity Period Monitor, and triggers (time based and event based). The SLA customizer also creates an SLO object in the SLA/contract repository and registers it as the call back handler of the alarms. The SLO object maintains the state of the SLO (valid, active, invalid). If a registered alarm for start-date of an SLO arrives the state of the SLO is changed from init to valid. The SLO is invalidated when the end-date trigger arrives. In between as the evalWhen alarms are triggered (because of a time or an event happening) the SLO evaluator evaluates the SLO. The *SLO evaluator* obtains the required management information (based on evalOn, daytime Constraint and the evalFunc constituent of the specification) from the high performance database in memory. The SLO evaluator determines compliance/violations. The *SLA violation engine* maintains the record for violations, their timestamps, the levels of violation, and the clauses that are violated (both in memory and in log files). The business cockpit can be used for looking and visual analysis of the current SLAs, SLOs, their violation records. The violation records will also be used for triggering *contract assurance* processes and actions as specified by evalAction constituent of the SLO.

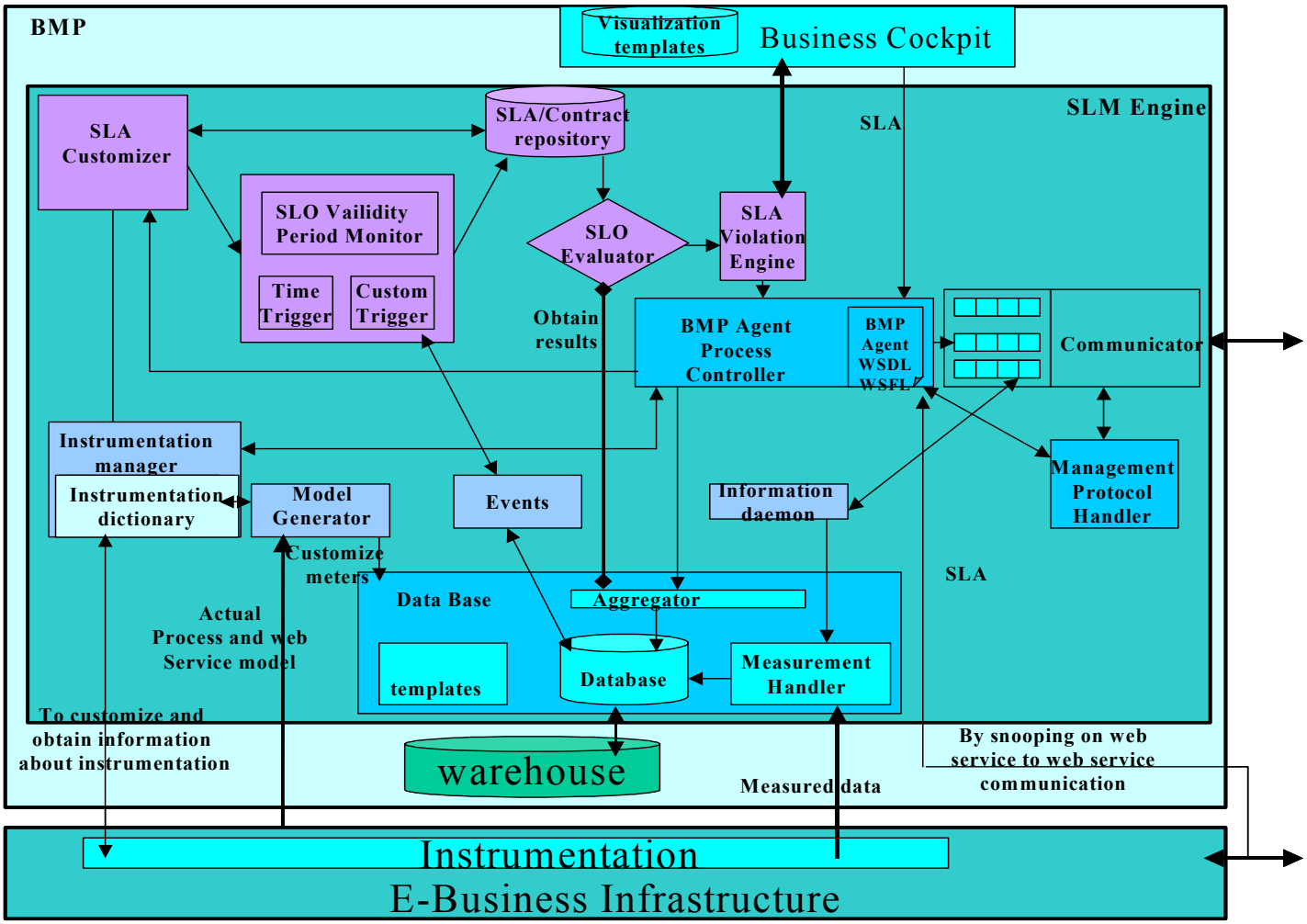


Fig4: The BMP Agent

3.3.1.1 Management Information Modeling

The *model generator* component receives the WSDL/WSFL specifications and creates a model of the web service in the *model repository*. The instrumentation dictionary contains information about the instrumentation and thereby the metrics that are available for various components of the web service. It can then combine the service model with the metrics available at each of the web service model component. This combined model is created in the repository. Subsequently when the actual measured data are stored by the measurement handler, the management data is stored according to the combined model.

All the measurements collected from the web service (e.g., ongoing interactions, performance measurements, etc) are attached to this combined model. The instrumentation in the web service is responsible for collecting these measurements and passing them on to the *management information handler* to be stored in the model repository. If the measurements are collected on

the client side (since the measuredAt component says so in an SLA), then the *communicator* is responsible for receiving the measurements and storing them into the repository.

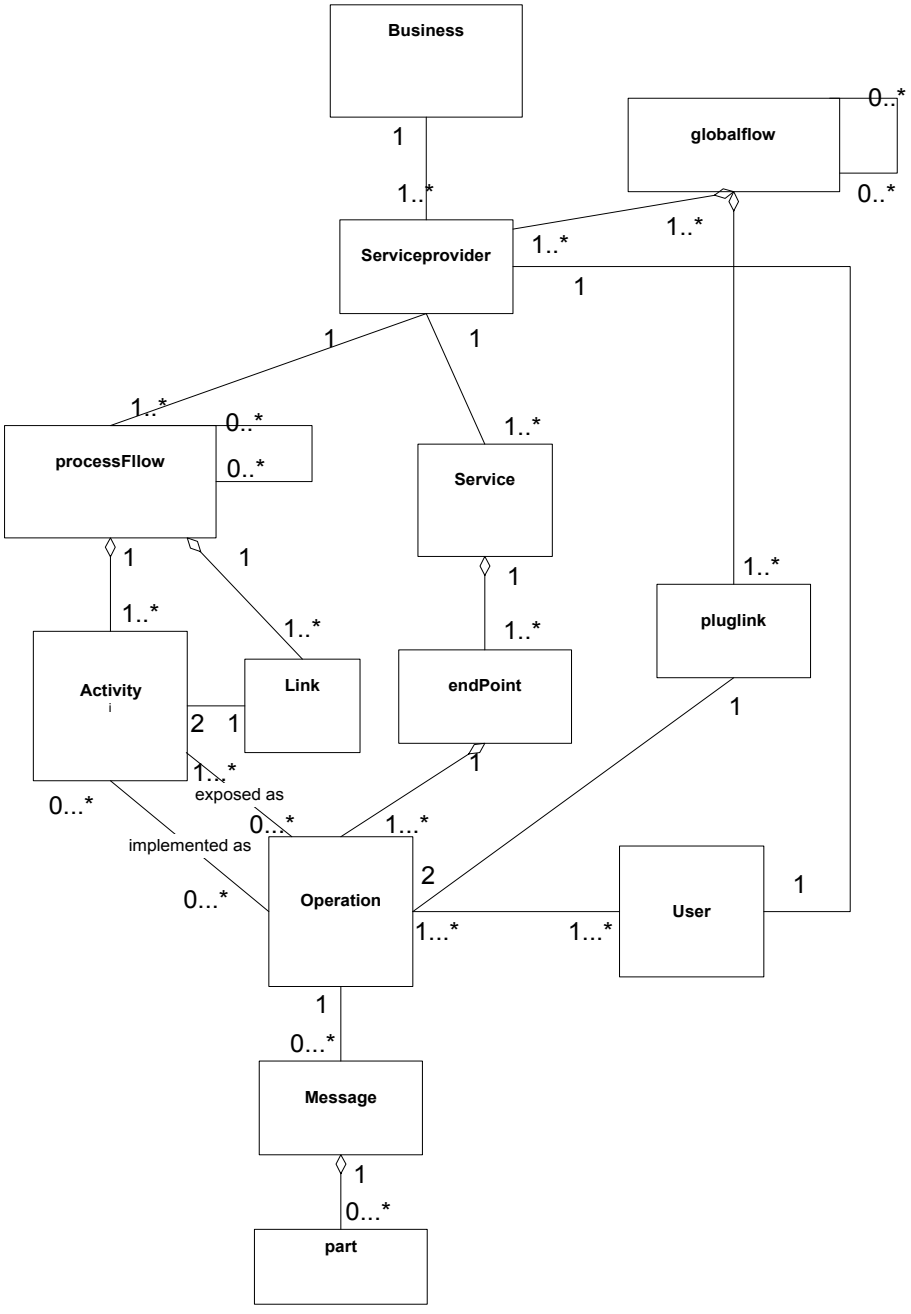


Fig5: General Web service + Business Process Model

Business	An organization that executes business processes. The business marks the boundaries of an administrator's *domain of responsibility* . A business can put out one or more service providers. A service provider controls its Business Process Flows.
ProcessFlow	A sequence of one or more workflow *activities* that achieve some intended purpose on behalf of the business.
Activity	Logical entities that form a workflow. Is realized by one or more applications and exposed as one or more operations
Application	Implements an activity.
Operation	Exposed part of the activities in a WSDL description
Message	An Operation is made up of one or more messages
User	A specific business, which invokes operations. A user could be a service provider too in a B2B scenario.
Service provider	A service provider provides services and Business Process Flows.
Conversation	Logical grouping of messages that can be done using context attribute
SLA	An agreement that web services decide upon
SLO	Service Level Objectives that form part of the SLO, usually based on the constructs defined in the model

In the managed object model used by the SLM engine, the basic web service and business process constructs are viewed as derived from a base class. We term the base class as the *managed object*. Every managed object has a set of *attributes*. An attribute is defined in the *attribute definition*. The attribute definition comprises of the *identifier*, *name*, *datatype*, *calculable*, *units* of the attribute. The identifier uniquely refers to an attribute definition while the name provides a label for it. The permissible data types are namely,

- **counter,**
  - **counter-threshold**
- **gauge,**
  - **gauge-threshold**
- **opaque**
  - **boolean**
  - **integer**
    - **uInt16**
    - **uInt32**
  - **string**
  - **timeTicks**
  - **uri,**
    - **objectId**
    - **url**
    - **physAddress**
      - **ipAddress**
      - **netAddress**
      - **nsapAddress**

Calculable determines whether an attribute conforming to the definition will be summable. There are three different values possible for calculable, namely non-calculable, summable and

non-summable. Non-calculable attributes are those that cannot be calculated (e.g. strings). Summable attributes are those that can be summed over multiple instance values. Units is a string that defines the specific units of the attribute (Bytes, ms..). New attributes can be defined by creating new attribute definitions and attaching them to the managed objects. This enables extensibility of the managed object model.

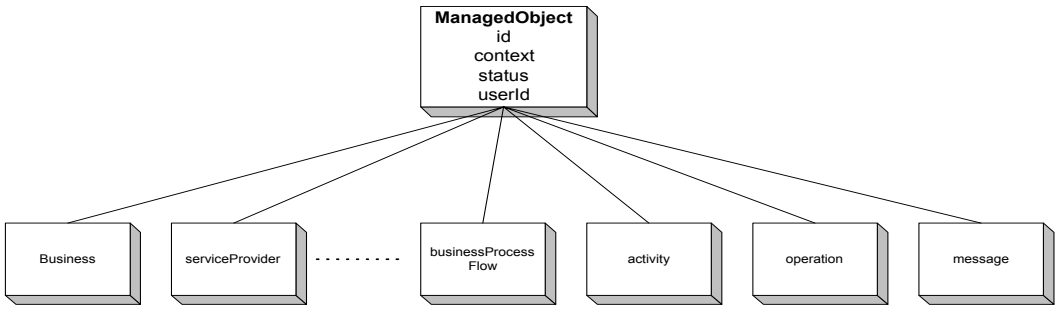


Fig 6: Hierarchy of managed object class and other web service constructs

The managed object has the *base attributes* of id, context, status, userId. All the other constructs, like operations, activity, processFlow, globalFlow, .. etc extend managed object. All the constructs thus have id, attribute, context, status, userId and other attributes that are specific to them. The additional attributes that would need to be measured at the different web service constructs (in addition to the base attributes) are shown in figure7.

The basic managed object model is extensible. At each of the constructs new attributes conforming to the data types mentioned above can be defined through new attribute definitions. This will allow for management systems that are capable of collecting additional information about the constructs. Also derived attributes can be defined that manipulate the base attributes.

In addition, metrics can be defined on top of the managed object model as defined in the previous section. A management system may create a metric object for modeling a (set of) managed object(s). The ITU-T model is quite applicable in our case of managed systems modeled through web service and business process abstractions [6]. The ITU-T metric object model for example provides for definition of mean monitor, moving average mean monitor. Mean and variance monitor, mean and percentile, mean and min max monitor.

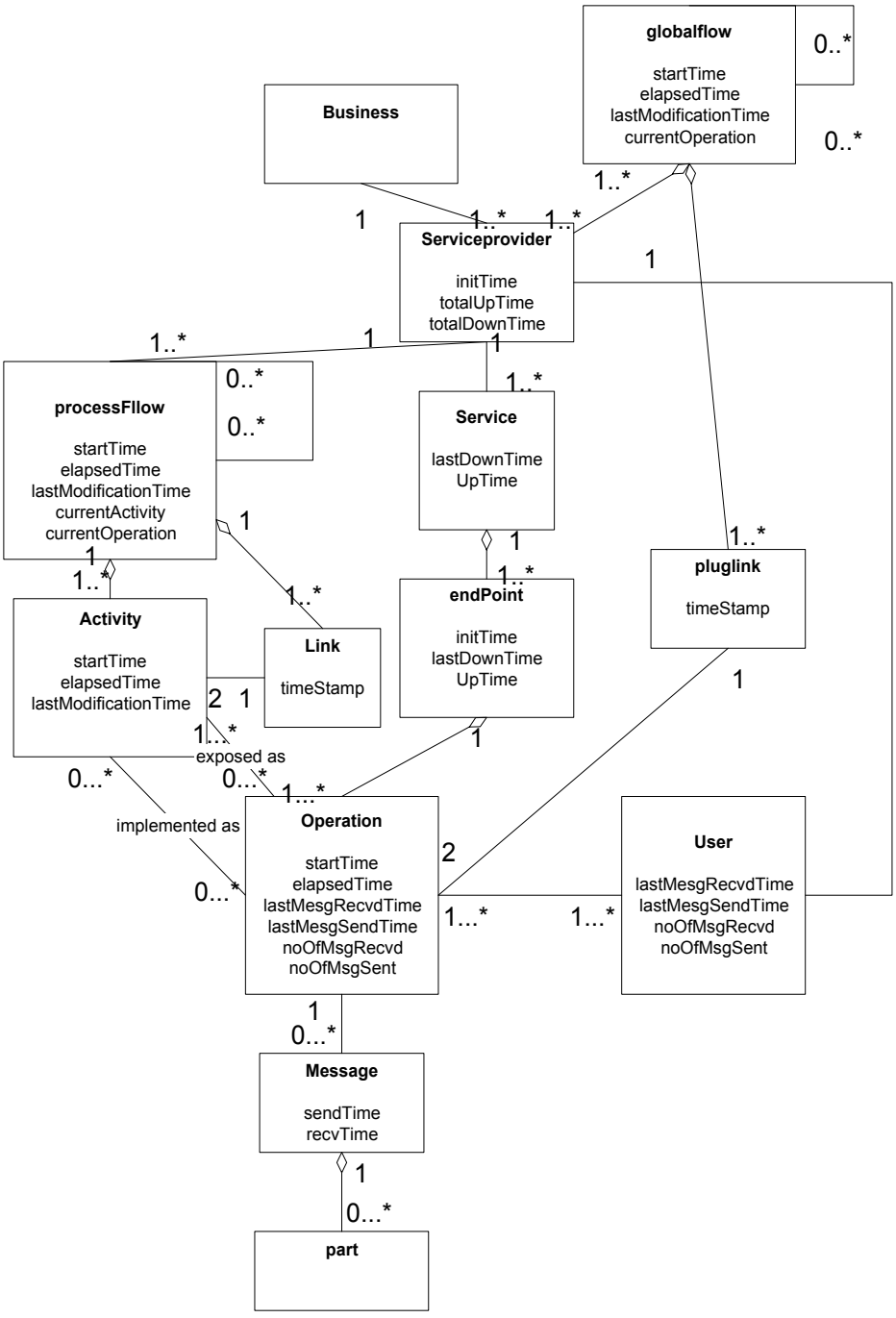


Figure 7: Managed Object model

The management data is thus collected and modeled in the databases in the SLM engines on both sides. If all the measuredItems are local then the SLA can be evaluated on the local data. However, if the measuredItems refer to attributes on web services on either side the data so collected needs to be exchanged between the SLM engines.

The data is continuously measured, modeled and stored in the database (and consequently in the data warehouse at regular intervals of time) as shown in Figure 4. The BMP Agent Process Controller receives the SLA specification either by snooping on the web service to web service communication or directly through the business cockpit. Once the SLA is received the Service Level Monitoring process flow is executed on both the provider-side and the customer-side.

3.3.1.2 Service Level Monitoring Process Flow

The process consists of the following steps:

(a) The process (SLM process) is initiated as soon as an SLA is received as input.
(b) Decide where the measurements are to be carried out. This is marked on every measured item in the SLA using *measuredAt*.
(c) Decide where the evaluation of the SLA is to be done. The SLA evaluation is carried out at the customer side, if the SLA has items that are all measured at the customer side. Similarly, if all the measured items are measured at the provider side, the SLA evaluation is carried out at the provider side. At the end of evaluation the SLM engines exchange violation report through SLA Violation Report Exchange protocol.
(d) If however, some of the items are measured at the customer side, and some of them are measured at the provider side, then the evaluation is carried out at the provider side. This last case, however requires that the customer-side measurements are transferred to the provider-side.
(e) If some of the measurements have to be transferred from customer side to provider side, initiate *measurement exchange protocol*. The measurement exchange protocol takes care of transferring measurements at the right frequency and right level of aggregation. This is described in detail in the next section.
(f) If the engine is responsible for the SLA evaluation, it sends the SLA to its SLA customizer that in turn creates the SLO, stores it in the SLA repository, customizes the alarms in the Alarm Manager and registers the SLO object as the call back handler for them. Once configured, the components of the SLA monitoring engine described above automatically trigger the evaluation of the SLA.

3.3.1.2.1 Measurement Exchange Protocol

When the evaluation of an SLA depends on measurements from both the customer-side and provider-side, a measurement protocol is needed for transferring the measurements from the former to the latter. Such a protocol should be designed with the following objectives in mind: (a) minimize the amount of data that is transmitted between the two sides, and (b) transfer the data in time for the evaluation of SLA to take place when triggered.

To fulfill these two objectives, the SLA monitoring engines on both sides should agree on (a) what measurements need to be transferred and at what level of aggregation, and (b) how frequently they should be transferred. The type and level of aggregation of the measurements depends on both evalFunc and measuredAt. To specify the level of aggregation, we use typical sampling functions such as count (t), totaled, averaged, movingAvg(lastN), minN, maxN, threshold. In the case when the sampling function cannot be determined from the evalFunc, we ship all the measurements from the customer-side to the provider-side. The reporting frequency depends on evalWhen.

The measurement protocol handles both the agreement on level of aggregation and frequency, as well as the transfer of agreed measurements from customer-side to provider-side. There are in essence 5 different types of messages that form the protocol.

- Init: sent by the consumer to the provider for clauses whose measurement data need to be exchanged. The init message carries possible choices of sampling function, interval, duration and reporting interval details that the consumer supports as shown below.

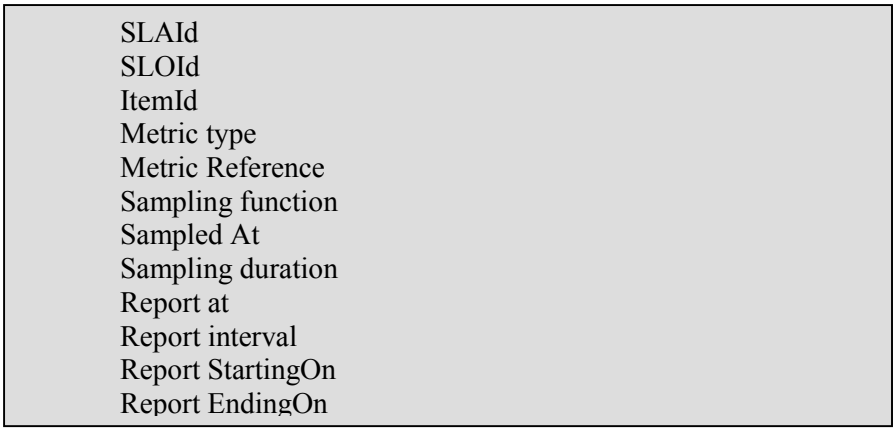
- Request: The provider decides the exact measurement specification (sampling function, sampling params and reporting params) that it chooses and specifies it in its request message.

- Agreement: The consumer sends this message if it agrees to the request

- Start: message from provider to commence the reporting.

- Report: actual measurement report messages

- Close: message to terminate the reporting.



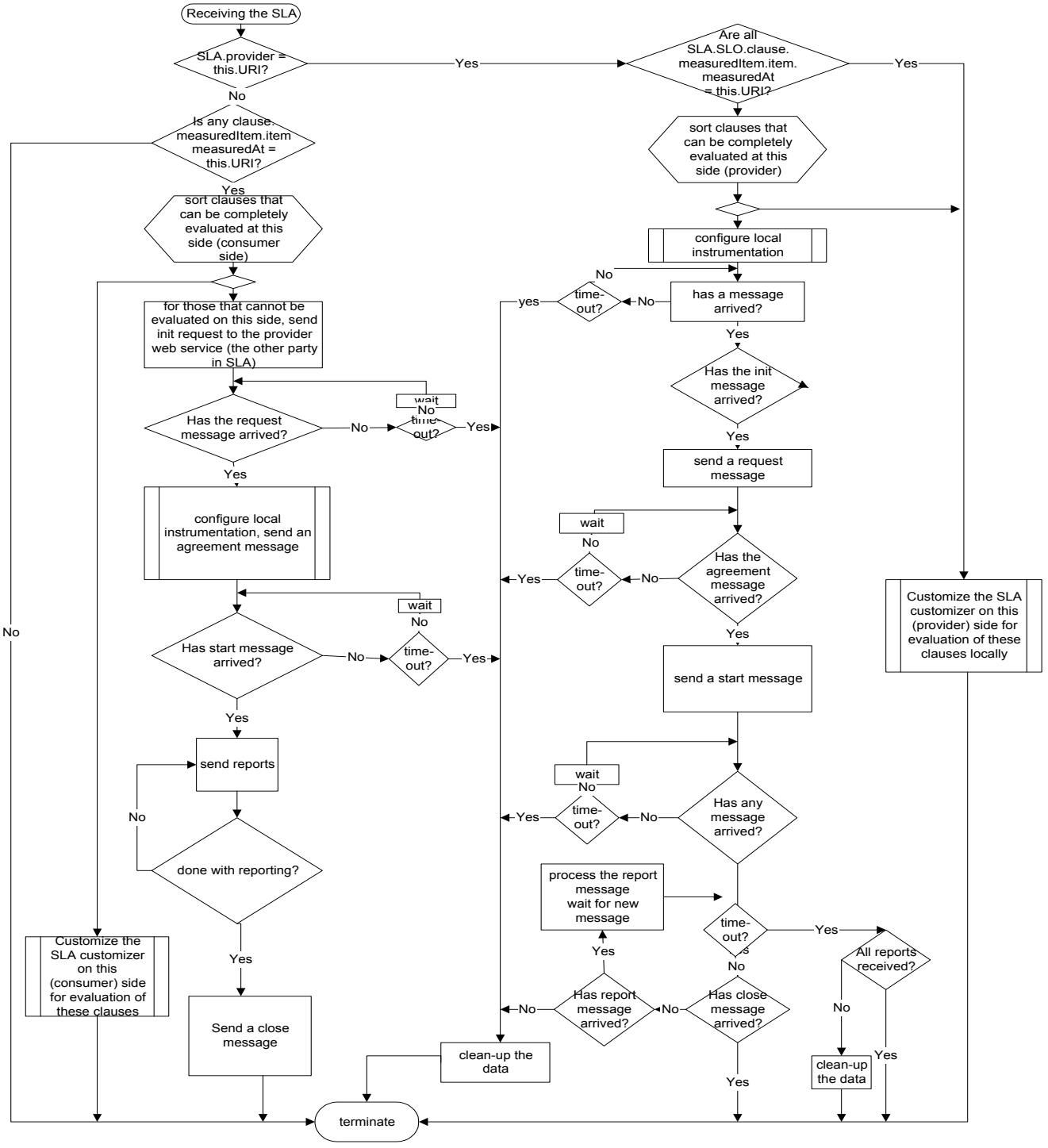


Fig8: The measurement process at the two BMP Agents

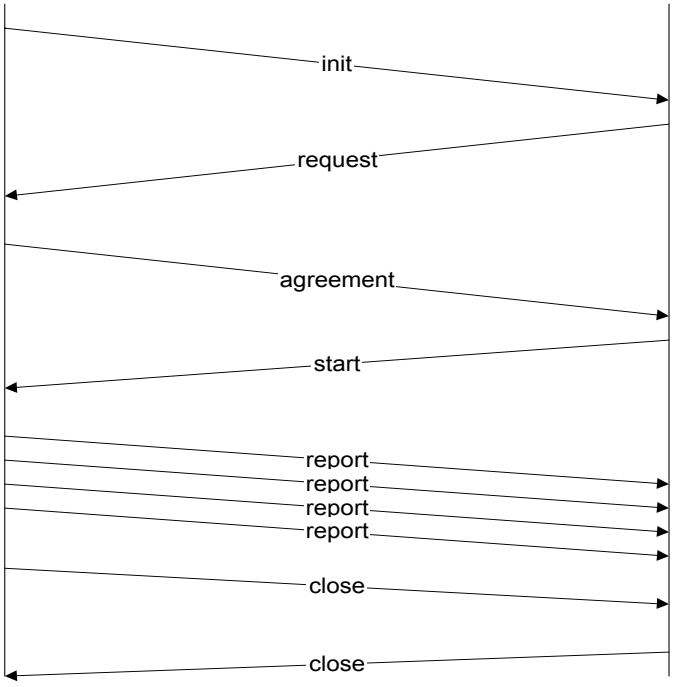


Fig9: The measurement exchange protocol for exchanging measurements collected between the BMP Agents

3.3.2 Violation Engine

Once the SLOs are invoked by the Alarm Manager, the SLO evaluator evaluates the function (evalFunc) of the SLO. The query that is created uses daytime constraint, evalOn and evalFunc components of the SLA specification. The results of these evaluations are compared against thresholds and the details of the evaluation are maintained as a Violation Record in the violation engine. It is also appended to the log File. The violation records can be used for controlling the web service and business process infrastructure for contract assurance purpose and for visual analysis by business managers.

3.4 Implementation

A Business Management Platform Agent was implemented (in Java). The BMP uses Apache SOAP toolkit to exchange messages with each other. They execute management processes on HP Process Manager. A sample web services scenario as described earlier and shown in Figure 10 was implemented and the messages, business processes involved were instrumented. For the web service scenario the actual business processes were also created on HPPM. HPPM provides a Java API to control process executions by other software components. A proxy component uses this Java API to feed in the GUID into HPPM process instances and retrieve it when necessary. The web services also use Apache SOAP toolkit for exchanging messages with each other. The SOAP toolkit was modified to collect the message correlation and instrumentation data. The measured data was stored and modeled in mySql database and Oracl9i data warehouse.

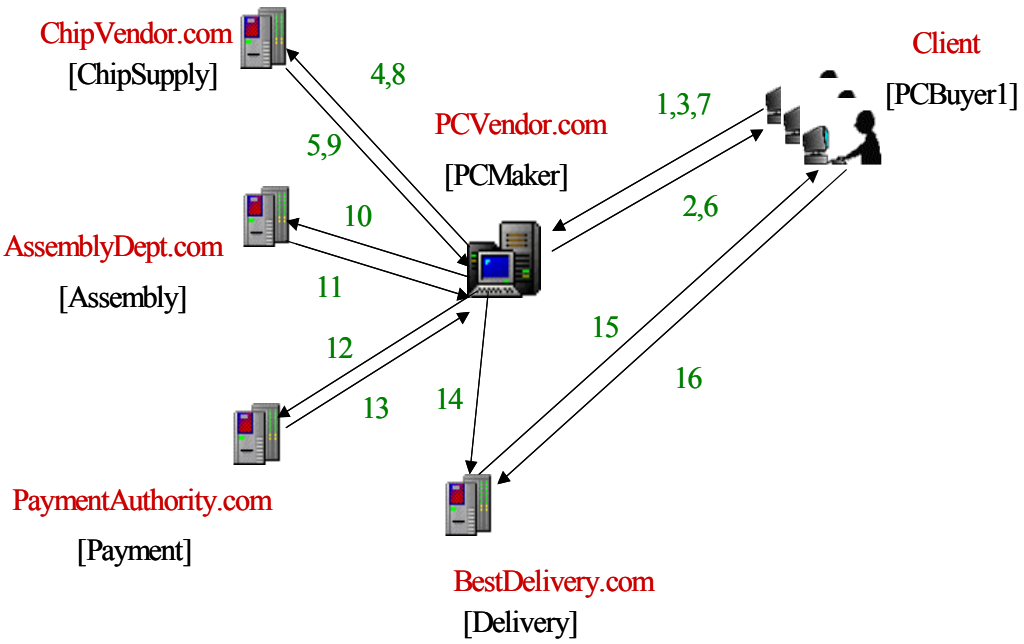


Fig10: The web services scenario that was implemented

The message exchanges in the current scenario:

#	MSG_TYPE	SENDER	RECEIVER
1	submitLoginmsg	PCBuyer1	PCMaker
2	ConfirmLoginmsg	PCMaker	PCBuyer1
3	SubmitQuoteRequestmsg	PCBuyer1	PCMaker
4	RequestChipQuotemsg	PCMaker	ChipSupply
5	SendChipQuotemsg	ChipSupply	PCMaker
6	SendQuotemsg	PCMaker	PCBuyer1
7	SubmitPORequestmsg	PCBuyer1	PCMaker
8	SendChipPOmsg	PCMaker	ChipSupply
9	RespondChipPOmsg	ChipSupply	PCMaker
10	SendAssemblyPOmsg	PCMaker	Assembly
11	RespondAssemblyPOmsg	Assembly	PCMaker
12	SendPaymentPOmsg	PCMaker	Payment
13	RespondPaymentPOmsg	Payment	PCMaker
14	SendDeliveryPOmsg	PCMaker	Delivery
15	SendDeliveryNotificationmsg	Delivery	PCBuyer1
16	sendReceiptNotificationmsg	PCBuyer1	Delivery

The example scenario as discussed earlier was implemented. The implemented scenario has two SLAs between PCMaker.com and its customers namely PCBuyer1.com, PCBuyer2.com. The two SLAs were namely SLA1 and SLA2. Each SLA has a single Service Level Objective. The first SLA is with PCBuyer1.com. It guarantees that between the dates of 0.2/15/02 and 07/15/02 all the invoice processes from 9-5 and on weekdays will be undertaken in 6 hours. The evaluation will be done every day at 6 PM.

<sla>
<slaId>2</slaId>
<partnerName>PcBuyer1.com</partnerName>
<startDate>Fri Feb 15 00:00:00 PST 2002</startDate>
<endDate>Mon Jul 15 00:00:00 PDT 2002</endDate>
<slo>
<sloId>1</sloId>
<dayTimeConstraint>Mon-Fri: 9-17</dayTimeConstraint>
<measuredItem>
<item>
<constructType>process</constructType>
<constructRef>PcMaker.com/Invoice</constructRef>
<measuredAt>PcMaker.com</measuredAt>
</item>
</measuredItem>
<evalWhen>6PM</evalWhen>
<evalOn>all</evalOn>
<evalFunc name =”averageResponseTime” operatior =”LT” Threshold =”6” unit =”hours”></evalFunc>
</slo>
</sla>

This SLA is signed between PCMaker.com and PCBuyer2.com. It guarantees that between the dates of 0.2/15/02 and 07/15/02 all the PC Delivery processes from 9-5 and on weekdays will be done on an average within 6 hours. The evaluation of the SLAs will be done every day at 6 PM.

<sla>
<slaId>1</slaId>
<partnerName>PCBuyer1.com</partnerName>
<startDate>Fri Feb 15 00:00:00 PST 2002</startDate>
<endDate>Mon Jul 15 00:00:00 PDT 2002</endDate>
<slo><sloId>1</sloId >
<dayTimeConstraint>Wed-Thu: 12-17</dayTimeConstraint>
<measuredItem>
<item>
<constructType>process</constructType>

<constructRef>PCMaker.com/PCDelivery</constructRef>
<measuredAt>PCMaker.com</measuredAt>
</item>
</measuredItem>
<evalWhen>6PM</evalWhen>
<evalOn>all</evalOn>
<evalFunc name=”averageResponseTime”operator=”LT” threshold = ”6”
unit=”hours”></evalFunc>
</slo>
</sla>

Also in order to demonstrate an SLA based on measurements from two different sites we created the following SLA based on two messages from two different end-points. This SLA is between PCMaker.com and PCBuyer1.com, but is based on two measuredItems. The BMP Agent at PCBuyer1.com sends the measurements to PCMaker.com for evaluation of the SLA everyday just before 6 PM and keep sending the reports from startDate to endDate.

<sla>
<slaId>3</slaId>
<partnerName>PcBuyer1.com</partnerName>
<startDate>Fri Feb 15 00:00:00 PST 2002</startDate>
<endDate>Mon Jul 15 00:00:00 PDT 2002</endDate>
<slo>
<sloId>1</sloId>
<dayTimeConstraint>Mon-Fri: 9-17</dayTimeConstraint>
<measuredItem>
<item>
<constructType>message</constructType>
<constructRef>PcMaker.com/submitPORequestmsg</constructRef>
<measuredAt>PcMaker.com</measuredAt>
</item>
<item>
<constructType>message</constructType>
<constructRef>PcBuyer1.com/sendReceiptNotificationmsg</constructRef>
<measuredAt>PcBuyer1.com</measuredAt>
</item>
</measuredItem>
<evalWhen>6PM</evalWhen>
<evalOn>all</evalOn>
<evalFunc name =”averageResponseTime” operatior =”LT” Threshold =”2” unit
=”days”></evalFunc>
</slo>
</sla>

The BMP Agent corresponding to PCMaker.com is loaded with the SLAs as mentioned above. These SLAs are passed as input to the Bmp Agent Process controller that in turn determine that these SLAs are all locally measured and are then passed to the SLA customizer. The SLA customizer creates the SLO objects and customizes the Alarm Managers. The evaluations are done as these alarms arrive. The snapshots of BMP Agent console are shown in Figure 11,12.

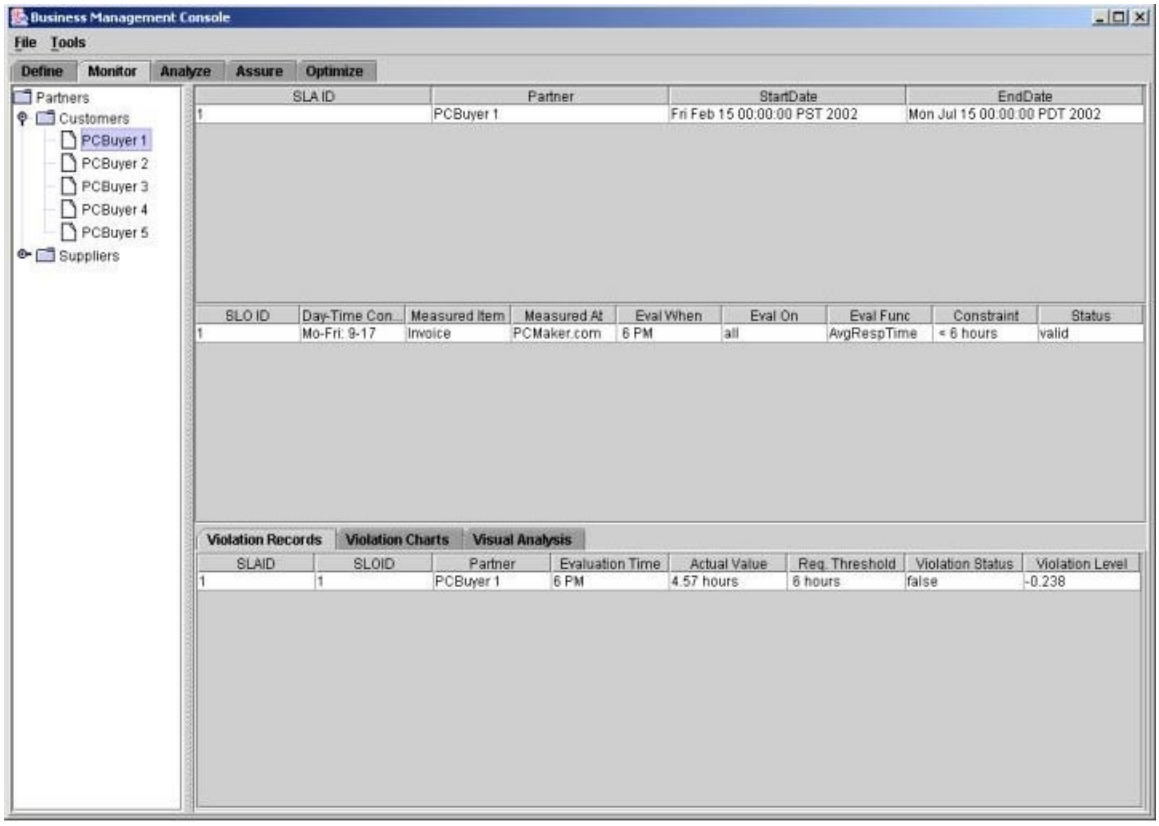


Fig 11: The Console of the BMP Agent

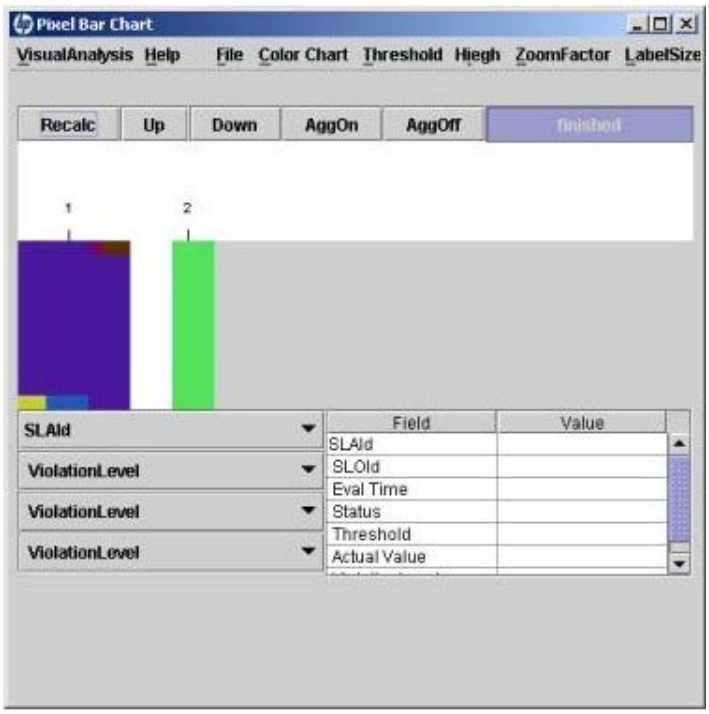


Figure 12: The Visual Analysis of the SLA Violation Logs

3.5 Related Work

One of the earlier important works that researched SLA management in a federated environment is presented in [3]. The SLA management engine requires a service model that determines the services offered in the domain as well as dependencies between the service components. It also needs the measurements available from them at each level to be specified in the model. A systems dictionary is required that specifies which plugins to use to gather which information. As the contracts defined in contract definition language are mapped to measurements from the systems dictionary and the process in not totally automated the specification can lead to ambiguities. A contract is defined by a triple (P,M,A), where P is a set of properties, A is the set of assertions and M is the set of methods available on the contract. An *assertion* is an atomic group of statements agreed upon between the parties agreeing to the contract. Statements in an assertion are made up of logical predicates whose values can be uniquely determined. The logical predicates are composed using variables as well as logical operators, quantifiers, set operations and constraints on these variables. An example assertion may be *response time < 25 ms*. For automation, it is necessary that the assertions be unambiguously specified. Also the web services will sign numerous SLAs with multiple parties over time and the SLA management process should be automated as much as possible. An assertion as mentioned above could lead to ambiguities. This could mean an instance response time or average response time. Again if it is average response time that is being referred to, is it averaged over every 5 minutes, an hour or 24 hours. It is also necessary to indicate when the averages are calculated. Is it at 6PM everyday? Or at any time? Also SLAs based on measurements from multiple sites have not been addressed in the work.

Inter-domain communication has been handled in telecommunication networks [6][7]. However, unlike the Internet their networks are regulated and typically designed to offer a single type of service. Also they have not looked at SLA management in a federated environment.

Most published work refers to managing network services and end to end mapping of network QoS [8][9]. However, they have not focused on protocols for sharing management information, and have not provided mechanisms to guarantee SLAs through unambiguous specification and auto-customization of federated SLA management engines and federated protocols to enable compliance.

C3DS [10] project exploits distributed object technology to create a framework for complex service provisioning. It uses MOM/Agent, Transactional workflow and Architecture description Language technologies to provide control to administrators to dynamically reconfigure agents and workflows deployed by it through a specification. Our approach is to automate the process of federated management and does not need administrators to perform SLA management. The C3DS approach has also not looked into SLA management problem.

Conclusion and future work

Service Level Agreements are difficult to specify in a clear and unambiguous manner. It is equally difficult to automate the monitoring of these SLAs. In addition, most of the SLAs deal with provider side guarantees and neglect client side measurements. In this article, we have proposed an automated and distributed SLA monitoring engine that monitors an SLA specified in our language.

There is often a sequencing involved among the SLOs of an SLA. Only if an SLO is fulfilled can the next SLO be evaluated. A sequencing logic on top of the specification is easy to describe. To execute the SLO sequencing an engine is required. This engine hands over the SLOs to be executed to the monitoring engine, and may initiate actions that are part of the functional part of the SLO on an execution engine. For example, an SLO 2 specifying a delivery timeliness guarantee may depend on an SLO 1 specifying a payment process timeliness guarantee. In this case the sequencing engine, evaluates SLO1 first, may initiate a payment PIP (say as specified in RosettaNet specification) and pass the SLO1 to the monitoring engine. The monitoring engine evaluates the timeliness guarantees and informs the sequencing engine about the outcome, which in turn can move to SLO2 according to the specification. In future we intend to create an over-all architecture for SLA life-cycle management which will include the sequencing engine, execution engine and the SLA monitoring engine. We also intend to undertake SLA conflict-detection and automatic contract/SLA assurance.

Acknowledgement

We would like to thank Aad van Moorsel for providing insights, ideas and help in developing the ideas presented in the paper. We would also like to thank Ming Hao for helping us in visual analysis of SLA and SLO logs through pixel bar charts.

References

1. Sahai A, Durante A, Machiraju V. Towards Automated SLA Management. HPL-2001-310
2. Jerome Daniel, Bruno Traverson, and Sylvie Vignes. A QoS Meta Model to Define a Generic Environment for QoS Management. Third International IFIP/GI Working Conference, USM 2000. Munich, Germany, September 12-14, 2000. In Proceedings Lecture Notes in Computer Science 1890 titled "Trends in Distributed Systems: Towards a Universal Service Market". Springer Verlag.
3. Bhoj P, Singhal S, Chutani S. SLA Management in a federated Environment. HPL-98-203.
4. Lewis D, Bjerring L. An inter-domain Virtual Private Network Management System. In the proceedings of NOMS 96
5. Lewis et al. Experiences in Integrated Multi-Domain Management. IFIP/IEEE International Conference on Management of Multi-Media Networks and Services, Montreal, Canada, 1997.
6. Hall J (editor). Management of Telecommunication systems and Services: Modelling and Implementing TMN based Multi-Domain Management, Lecture Notes in Computer Science 1116, Springer-Verlag, ISBN 3-540-61578-4, 1996
7. Telecommunication Management Network (TMN) at ITU-T. Formerly CCITT. http://www.itu.int
8. Aurrecoechea, C., Lazar, A.A. and Stadler, R., Open Network Services for Management, IEEE Conference on Open Architectures and Network Programming, San Francisco, CA, April 3-4, 1998.
9. Huard, J.-F. and Lazar, A.A., On QOS Mapping in Multimedia Networks, *21th IEEE Annual International Computer Software and Application Conference (COMPSAC '97)*, Aug. 13-15, 1997, Washington, D.C.
10. *Shrivastava S.* C3DS Platform for Service Provisioning. C3DS Technical Report number 44. , 20 pages, 2001 http://www.newcastle.research.ec.org/c3ds/trs/abstracts/44.html
11. Dirk Thiβen and Helmut Neukirchen. Internet Trading and Load Balancing for Efficient Management of Services in Distributed Systems. Third International IFIP/GI Working Conference, USM 2000. Munich, Germany, September 12-14, 2000. In Proceedings Lecture Notes in Computer Science 1890 titled "Trends in Distributed Systems: Towards a Universal Service Market".
12. Long T P, Jong W B, Woon HJ. Management of service level agreements for multimedia Internet service using a utility model. IEEE communications Managezine Vol 39, no.5, May 2001
13. Forbath T. Why and how of SLAs [service level agreements]. Business Communications Review, Vol 28. No. 2, Feb 1998
14. Chatterjee BS, Sydir M, Lawrence T. Taxonomy for QoS specifications. In the proceedings of WORDS'97, February, 1997
15. Lewis L, Ray P. Service Level Management: Definition, Architecture, and Research Challenges. In the proceedings of IEEE GlobeCom'99.
16. Katcgabaw M, Lutfiyya H, and Bauer M. Driving Resource Management with Application-Level Quality of Service Specifications. In the proceedings of ICE 98, USA.
17. Tierney B, Crowley B, Gunter D et al. A Monitoring Sensor Management System for Grid Environments. http://www-didc.lbl.gov/papers/JAMM.HPDC00.pdf
18. Mesnasce D, Almeida V, Fonesca R, Mendes M. Resource Management Policies for E-Commerce Servers.
19. Campbell A, Aurrecoechea C., Hauw L .QoS review Architectures, Proceedings of the 4th International Workshop on Quality of Service (IWQoS)
20. Wolter K, Van Moorsel A. The Relationship between Quality of Service and Business Metrics: Monitoring, Notification and optimization – HPL-2001-96.

21. Langer M, Nerb M. Defining a Trouble Report Format for the Seamless Integration of Problem Management into Customer Service Management HP OpenView University Association (HP-OVUA) Plenary worhsop, Bologna, Italy, 1999.

22. Hauck R, Reiser H. Monitoring of Service Level Agreements with Flexible and Extensible Agents. HP OpenView University Association (HP-OVUA) Plenary worhsop, Bologna, Italy, 1999.

23. Fonesca M, Agoulmine N, Cherkaoui O. Active Networks as a flexible approach to deploy QoS Policy based Management. HP OpenView University Association (HP-OVUA) worhsop, Berlin, 2001

24. Nakamura Y et al. ENMA: The WWW Server Performance Measurement System via Packet Monitoring. In the proceedings of INET 99, San Jose, CA, USA, 1999.

25. Web Services Description Language (WSDL) http://www.w3.org/TR/wsdl

26. Web Services Flow Language (WSFL) . http://www.ibm.com/software/solutions/webservices/

27. Tele Management Forum SLA Management Handbook, GB917, public evaluation version 1.5 , June 2001. http://www.tmfcentral.com/kc/repository/documents/GB917v1.5.pdf.

28. Samani M, Sloman M. Monitoring of Distributed Systems (A Survey). Imperial College Research Report DOC 92/93. Sept, 1992.