Defining and Monitoring Service Level Agreements for dynamic e-Business

Alexander Keller, alexk@us.ibm.com
Heiko Ludwig, hludwig@us.ibm.com
Why should SysAdmins care about SLAs?

- How much does it cost you to guarantee a Response Time less than 1 sec.?
- How much do you bill a Customer for a Throughput of 1000 TAs/sec?
- How much Revenue is lost per Hour of Downtime of Server X?
  - Express System Resources in Financial Terms
- What are realistic Thresholds for Response Time/Throughput/Bandwidth?
- Can you accommodate additional Workload and accept another Customer?
- How much Workload do SLA Measurements put on Server X?
- How does this impact your SLAs with other Customers?
  - SysAdmins will become involved in SLA Negotiation (today: Lawyers)
- If your Systems become overloaded, which Customer will be starved out?
  - Classify Customers according to Revenue
  - SLA Violation may not be due to technical Failure, but Result of Business Decision
- What’s more expensive?
- A Disk-Crash on a Server or an overloaded Ethernet Segment?
- Depends on how much the Customer pays whose Data is hosted there!
  - Fix Outages according to Customer Classification (today: Severity of Outage)
Real-world SLAs – and their Requirements

- Today: Confined to Availability
  - “Availability\% := (n - \#hours\_Svc\_down) * 100 / n”
  - “… Users being able to establish a TCP Connection to the Server…”
  - “…Customer’s ability to access the Software Application on the Server…”
  - “… if the Server is responding to HTTP Requests issued by monitoring SW…”

BUT: There is no agreed-upon Definition of “Availability”!

What is needed?
- Define new SLAs “on demand” (e.g., Grid, Virtual Enterprises, Web Services)
- Accommodate ANY QoS Parameter Definition and Service Level
- Go beyond “Availability”: Response Time, Throughput, Bandwidth…
- Connect to existing Application and Resource Instrumentation
- Support Customer/Provider Relationships of arbitrary Depth
- Delegate SLA Monitoring Tasks to Third Parties
- Address Confidentiality Requirements of the Parties (“Need to know”)
- Automated Setup of Monitoring Environment based on SLA Definition
Terminology: SLA Parameters, Metrics, Functions

Business Metrics  SLA Parameters  Composite Metrics  Resource Metrics

- The analyzed SLAs share a common Structure:
  - Involved Parties, SLA Parameters
  - Metrics used as Input to compute SLA Parameters
  - The Functions that define how Metrics are aggregated
  - How Metrics are retrieved from Managed Resources (Measurement Directive)

Customer-defined  Provider-defined
Web Service Level Agreement (WSLA) Framework

- SLA annotates an existing Service Specification:
  - References Service Description (e.g., Web Services: WSDL)
  - Other Service Descriptions possible, e.g., for Business Processes, Messaging, IT Resources
- XML Schema based Language for SLAs,
- Runtime Architecture comprising several SLA Monitoring Services
The WSLA Services: Atomic Building Blocks

- **Establishment & Deployment Services**
  - Supports negotiation and authoring of SLAs
  - Deploys the relevant (!) Parts of the SLA to the different Parties
  - E.g., multiple Measurement Services may not “see” each other

- **Measurement Service**
  - Probes and measures Resource Metrics according to SLA Specification and aggregates them into SLA Parameters

- **Condition Evaluation Service**
  - Compares SLA Parameters obtained from Measurement Service against specified Service Levels
  - Notifies the involved Parties that a Violation has occurred during a valid Time Period

- **Management Service & Business Entity (not yet supported)**
  - Carries out corrective Actions, provided they do not violate Business Policies
  - Access to - proprietary - Tuning Controls and Configuration Parameters of managed Resources often not available,
  - Must be checked against Business Policies embodied by Business Entity
SLA Lifecycle in the WSLA Architecture

1. Negotiate / Sign
2. Deploy
3. Report
4. Act
5. Terminate

Establishment

SLA

WSDL

Web Service

AppServer Monitoring/Management Interfaces

Service Customer

Service Provider

Admin Console

SLA Compliance Monitor

Deployment

Measurement

Condition Evaluation

Management

Business Entity
Delegating SLA Monitoring Tasks to Third Parties

Measurement Service Providers guarantee Accuracy and Objectivity (e.g., Keynote Systems)
## Typical Structure and Elements of an SLA

<table>
<thead>
<tr>
<th>Parties:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signatory Parties</td>
</tr>
<tr>
<td>Supporting Parties</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Service Description:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service Operations</td>
</tr>
<tr>
<td>Bindings</td>
</tr>
<tr>
<td>SLA Parameters</td>
</tr>
<tr>
<td>Metrics</td>
</tr>
<tr>
<td>Measurement Directives</td>
</tr>
<tr>
<td>Functions</td>
</tr>
<tr>
<td>Schedule</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obligations:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity Period</td>
</tr>
<tr>
<td>Predicate</td>
</tr>
<tr>
<td>Actions</td>
</tr>
</tbody>
</table>

### Involved Parties:
- IDs and Interfaces of Signatory Parties
- IDs and Interfaces of Supporting Parties

### Service Characteristics & Parameters:
- Operations offered by Service
- Transport encoding for Messages
- Agreed-upon SLA Parameters (Output)
- Metrics used as Input
- How/where to access Input Metrics
- Measurement Algorithm
- Measurement Duration, Sampling Rate

### Guarantees & Constraints:
- When is SLA Parameter guaranteed?
- How to detect Violation (Formula)
- Corrective Actions to be carried out
### SLA Structure Example: Service Throughput

#### Involved Parties:
- “customer.com”, “provider.com”
- “msp.com, keynote.com, …”

#### Service Characteristics & Parameters:
- “StockQuoteService:GetQuote()”
- “SOAPGetQuote”
- “average throughput of service”
- “#Requests(svc)”
- “www.msp.com/getMetric?Requests(svc)”
- “AVG(#Requests(svc))”
- “over 24 hours, every 60 minutes”

#### Guarantees & Constraints:
- “weekdays, 9am-5pm”
- “> 1.000 TA/second”
- “open TT”, “pay penalty/premium”

<table>
<thead>
<tr>
<th><strong>Parties:</strong></th>
<th><strong>Involved Parties:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Signatory Parties</td>
<td>“customer.com”, “provider.com”</td>
</tr>
<tr>
<td>Supporting Parties</td>
<td>“msp.com, keynote.com, …”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Service Description:</strong></th>
<th><strong>Service Characteristics &amp; Parameters:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Service Operations</td>
<td>“StockQuoteService:GetQuote()”</td>
</tr>
<tr>
<td>Bindings</td>
<td>“SOAPGetQuote”</td>
</tr>
<tr>
<td>SLA Parameters</td>
<td>“average throughput of service”</td>
</tr>
<tr>
<td>Metrics</td>
<td>“#Requests(svc)”</td>
</tr>
<tr>
<td>Measurement Directives</td>
<td>“www.msp.com/getMetric?Requests(svc)”</td>
</tr>
<tr>
<td>Functions</td>
<td>“AVG(#Requests(svc))”</td>
</tr>
<tr>
<td>Schedule</td>
<td>“over 24 hours, every 60 minutes”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Obligations:</strong></th>
<th><strong>Guarantees &amp; Constraints:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity Period</td>
<td>“weekdays, 9am-5pm”</td>
</tr>
<tr>
<td>Predicate</td>
<td>“&gt; 1.000 TA/second”</td>
</tr>
<tr>
<td>Actions</td>
<td>“open TT”, “pay penalty/premium”</td>
</tr>
</tbody>
</table>
Example: Defining SLOs with Constraints

- Why define Constraints for Service Level Objectives?
  - If your hosted Site becomes too popular and creates excessive Load, your Throughput SLO may be impossible to fulfill
  - Service Provider needs to protect himself against this Situation
- SLOs are defined for regular Workloads
  BUT: What is a “regular Workload”? Needs to be defined within the SLA!

Example:
- “If the System Load is over 80% for more than 30% of the Time, the Obligation of a Service Provider to guarantee 1000 TAs/sec is waived”
- 2 SLA Parameters:
  - AvgThroughput: Average Throughput for TAs; must be > 1000 TA/sec
  - OverloadPct: The % amount of time System Utilization is => 80%
- Both Parameters are measured every 5 Minutes on an hourly Basis
- In an SLO, OverloadPct is used as Precondition for AvgThroughput
Example (cont.)

ServiceObject

has

WSDL:getQuote

has

SLAParameter

AvgThroughput
defined by

Metric

UtilizationTimeSeries

Function

TimeSeriesConstructor
defined by

Metric

Transactions

Measurement Directive

Read: TXcount

defined by

Metric

TimeSpent

Measurement Directive

Read: Timecount

defined by

Metric

AvgThroughput

Function

Average
defined by

Metric

ThroughputTimeSeries

Function

TimeSeriesConstructor
defined by

Metric

Throughput

Function

Divide
defined by

Metric

PercentOverUtilized

Function

PercentageGreaterThanThreshold
defined by

Metric

ProbedUtilization

Measurement Directive

Probe: acme.com/SystemUtil

defined by

How obtained?

New Value put in Time Series

Assign to SLA Parameter

How many TimeSeries Elements > Threshold?
Defining SLA Parameters and Metrics:

Assignment of Metric to SLA Parameter

Who Communicates with whom? And how?

Define the Metric: How many Values (in %) of a “Utilization” Time Series are over a Threshold of 80%?

Create the Time Series: - probe every 5 Minutes - keep the last 12 Values
SLOs in the WSLA Language:

ACMEProvider guarantees the SLO

The SLO is valid for 1 Day

Time Format: RFC 3060

Precondition:
OverloadPercentage < 30%

Guarantee:
Average Throughput > 1000

Send NewValue Event to registered Parties whenever Guarantee is broken

© 2002 IBM Corporation
Conclusions and Outlook

WSLA supports:
- Flexible Specification of inter- and intra-organizational SLA Parameters
- Highly customizable Service and IT Resource-Level SLOs
- Nested Customer/Provider Relationships
- Definition of third ("supporting") Parties in SLA Management
- Formal, XML-Schema based Description Language
- Applicable to various Kinds of Services (Web Services, Storage, eUtilities etc.)

SLA Compliance Monitor Implementation Part of IBM Web Services Toolkit

Current Work:
- Comprehensive SLA Framework, comprising:
  - Business Metrics and Pricing,
  - Business Processes, Workflow and Service Composition,
  - SLA Editing and Reuse of common SLA Artifacts,
  - Integration with existing Management Frameworks (WBEM / CIM)
Let us know what you think!

Download WSTK 3.2 with SLA Compliance Monitor from:

http://www.alphaworks.ibm.com/tech/webservicestoolkit