IBM T.J. Watson Research Center

Defining and Monitoring Service Level Agreements for dynamic e-Business

Alexander Keller, *alexk@us.ibm.com*
Heiko Ludwig, *hludwig@us.ibm.com*

LISA 02 | 11/07/2002 | Philadelphia, PA, USA

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

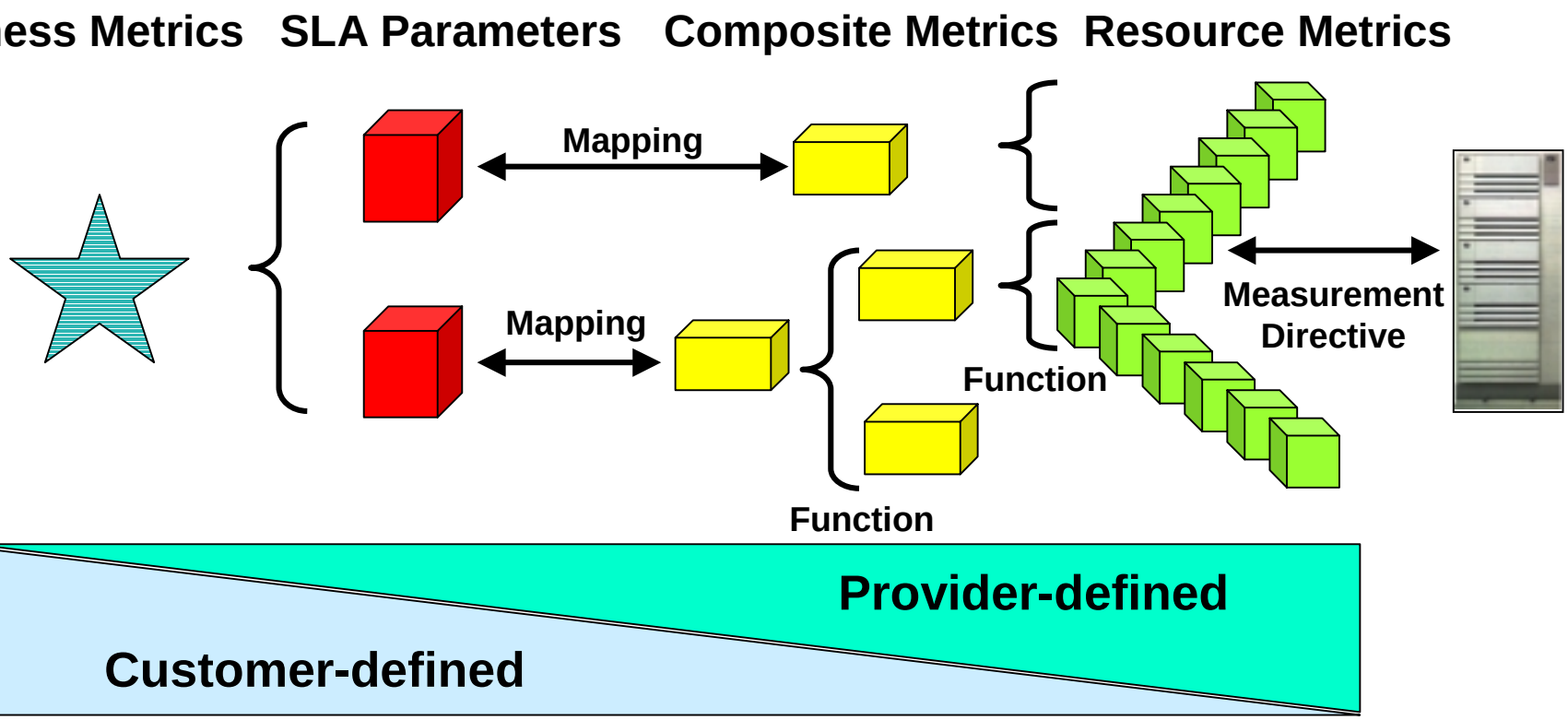
- Today: Confined to Availability
  - "Availability% := (n - #hours_Svc_down)* 100 / n"
  - "… Users being able to establish a TCP Connection to the Server…"
  - "…Customer's ability to access the Software Application on the Server…"
  - "… if the Server is responding to HTTP Requests issued by monitoring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

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

What is needed?

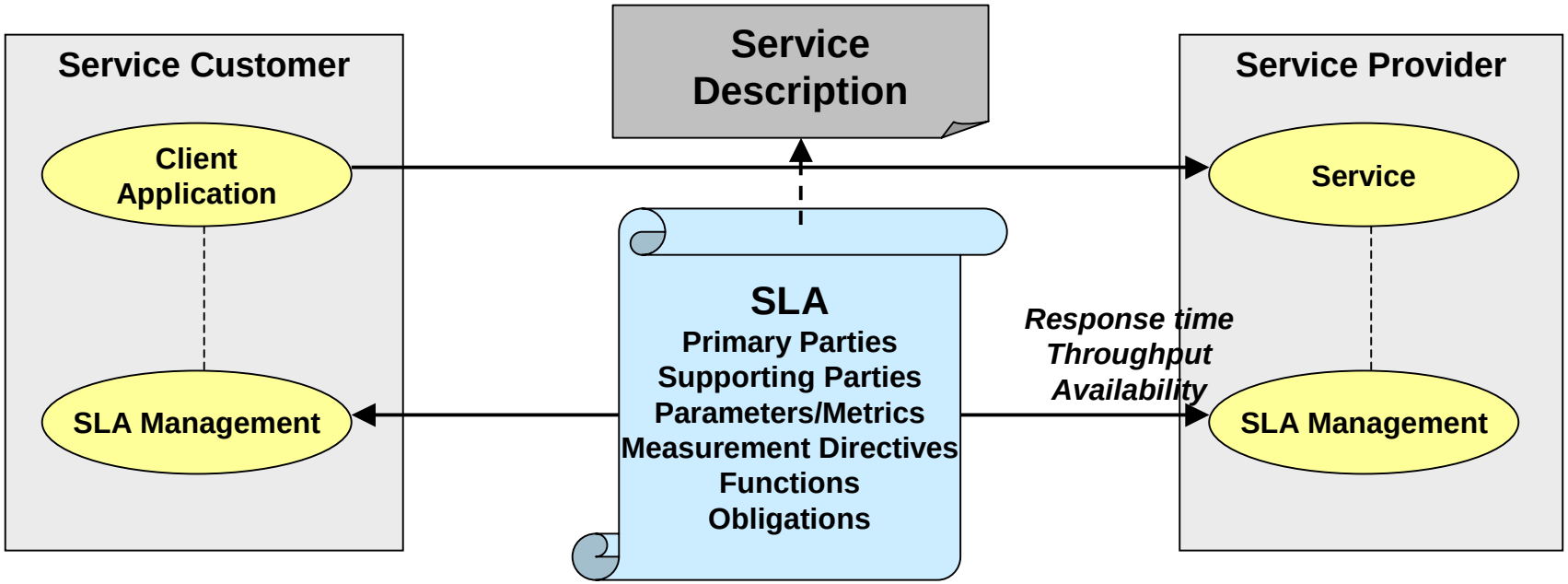
- Define new SLAs "on demand" (e.g., Grid, Virtual Enterprises, Web Services)
- Accommodate ANY QoS Parameter Definition and Service Level
- Go beyond "Availability": Response Time, Throughput, Bandwidth…
- Connect to existing Application and Resource Instrumentation
- Support Customer/Provider Relationships 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



- 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)

Web Service Level Agreement (WSLA) Framework

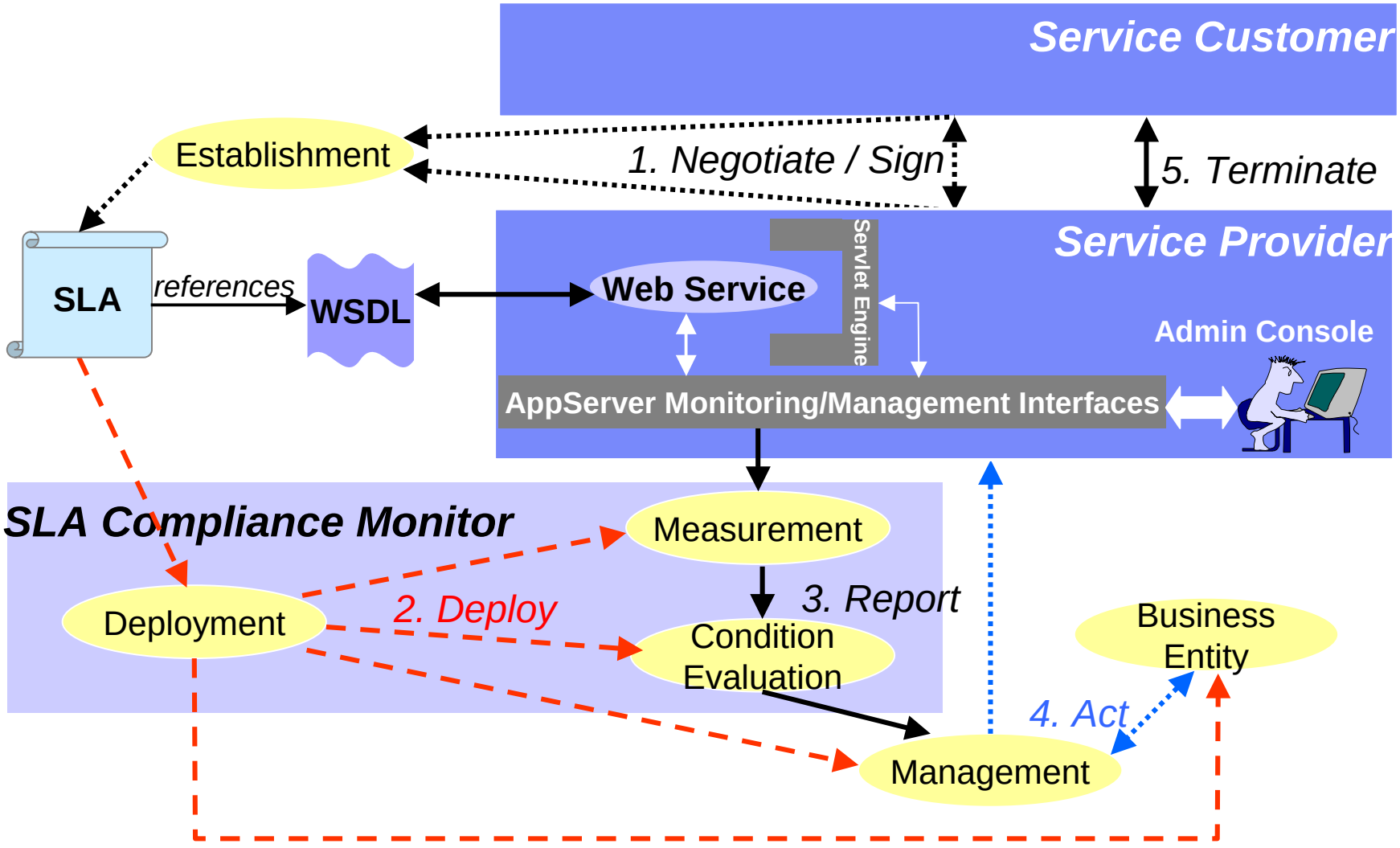


- SLA annotates an existing Service Specification:
  - References Service Description (e.g., Web Services: WSDL)
  - Other Service Descriptions possible, e.g., for Business Processes, Messaging, IT Resources
- XML Schema based Language for SLAs,
- Runtime Architecture comprising several SLA Monitoring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

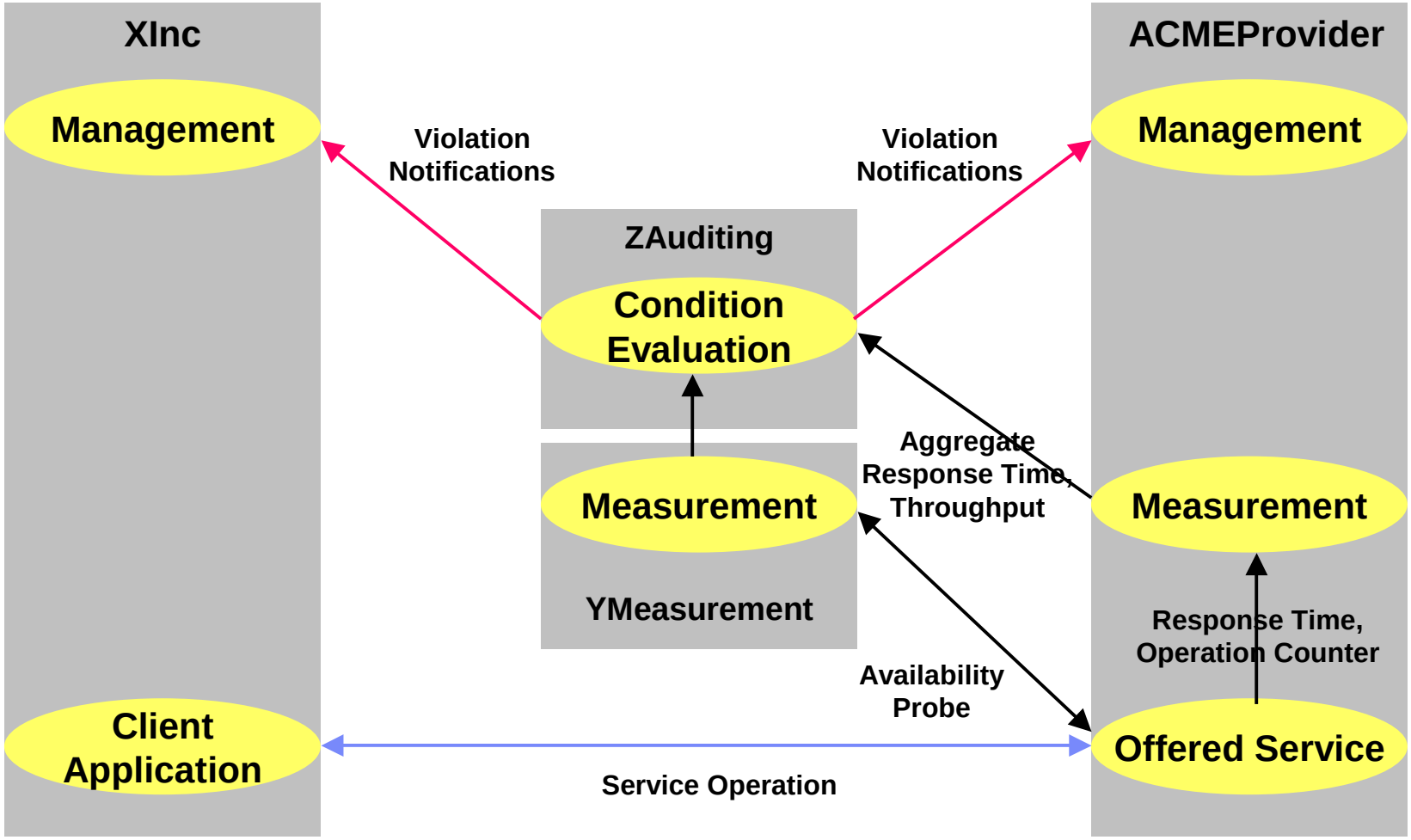
The WSLA Services: Atomic Building Blocks

- 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



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

Parties:
Signatory Parties
Supporting Parties
Service Description:
Service Operations
Bindings
SLA Parameters
Metrics
Measurement Directives
Functions
Schedule
Obligations:
Validity Period
Predicate
Actions

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

Parties:
Signatory Parties
Supporting Parties
Service Description:
Service Operations
Bindings
SLA Parameters
Metrics
Measurement Directives
Functions
Schedule
Obligations:
Validity Period
Predicate
Actions

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"

Example: Defining SLOs with Constraints

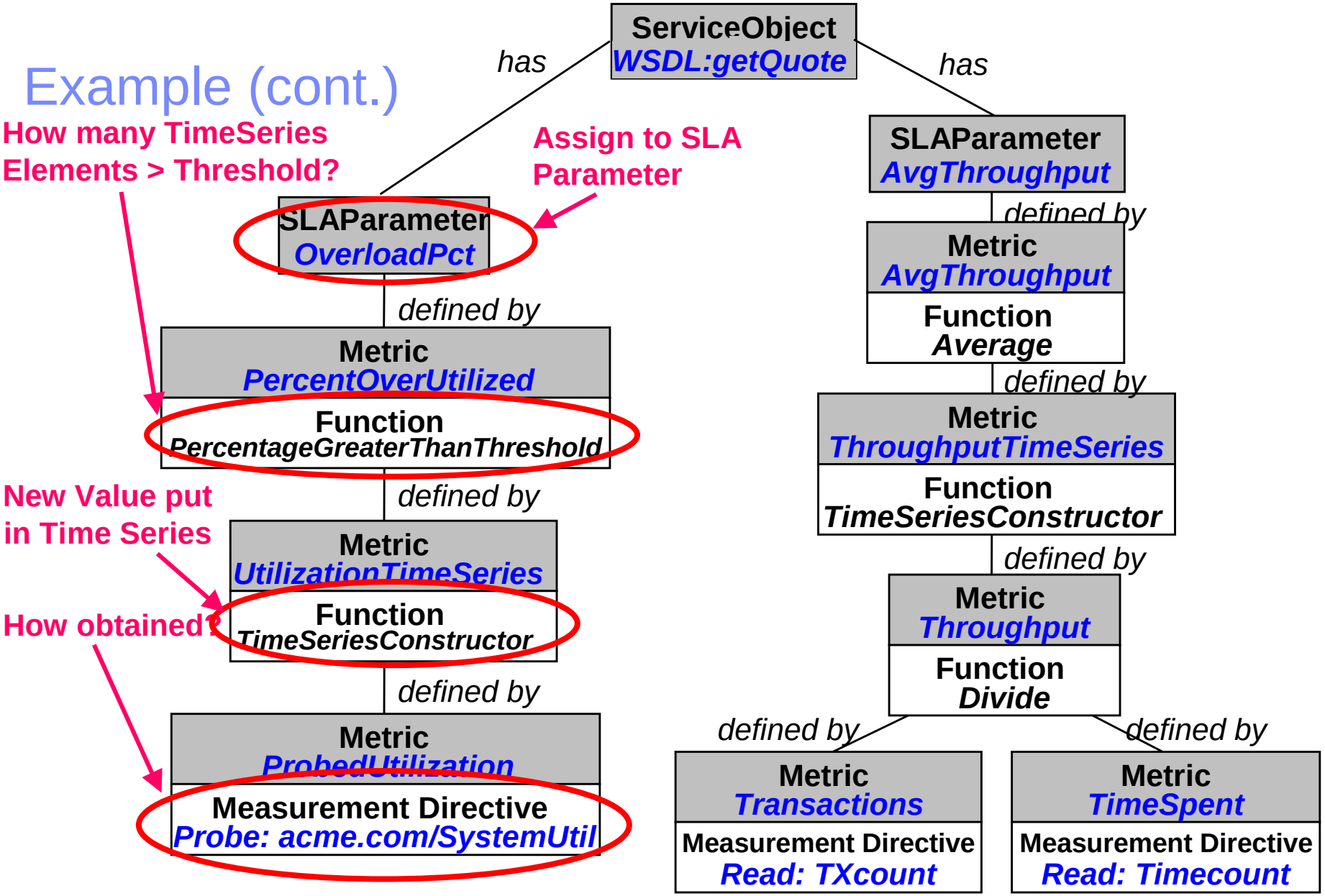
- Why define Constraints for Service Level Objectives?
  - If your hosted Site becomes too popular and creates excessive Load, your Throughput SLO may be impossible to fulfill
  - Service Provider needs to protect himself against this Situation
- SLOs are defined for regular Workloads

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.)



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**

```
<SLAParameter name="OverloadPct" type="float" unit="Percentage">
 <Metric>OverLoadPct</Metric>
 <Communication>
  <Source>YMeasurement</Source>
  <Pull>ZAuditing</Pull>
  <Push>ZAuditing</Push>
 </Communication>
</SLAParameter>

<Metric name="OverloadPct" type="float" unit="Percentage">
  <Source>YMeasurement</Source>
  <Function xsi:type="PctGTThreshold" resultType="float">
    <Schedule>BusinessDay</Schedule>
    <Metric>UtilizationTimeSeries</Metric>
    <Value>
      <LongScalar>0.8</LongScalar>
    </Value>
  </Function>
</Metric>

<Metric name="UtilizationTimeSeries" type="TS" unit="">
  <Source>YMeasurement</Source>
  <Function xsi:type="TSConstructor" resultType="float">
    <Schedule>Every5Minutes</Schedule>
    <Metric>ProbedUtilization</Metric>
    <Window>12</Window>
  </Function>
</Metric>
```

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

```
<ServiceLevelObjective name="SLO_for_AvgThroughput">
 <Obliged>ACMEProvider</Obliged>
 <Validity>
   <Start>2001-11-30T14:00:00.000-05:00</Start>
   <End>2001-12-31T14:00:00.000-05:00</End>
 </Validity>
 <Expression>
  <Implies>
    <Expression>
     <Predicate xsi:type="Less">
      <SLAParameter>OverloadPct</SLAParameter>
      <Value>0.3</Value>
     </Predicate>
    </Expression>
    <Expression>
     <Predicate xsi:type="Greater">
      <SLAParameter>AvgThroughput</SLAParameter>
      <Value>1000</Value>
     </Predicate>
    </Expression>
   </Implies>
  </Expression>
 <EvaluationEvent>NewValue</EvaluationEvent>
...

</ServiceLevelObjective>
```

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