2012-12

An investigation into specifying service level agreements for provisioning cloud computing services

Kelley, Nancy J.

Monterey, California. Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

AN INVESTIGATION INTO SPECIFYING SERVICE LEVEL AGREEMENTS FOR PROVISIONING CLOUD COMPUTING SERVICES

by

Nancy J. Kelley

December 2012

Thesis Co-Advisors: Man-Tak Shing
James Bret Michael

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704–0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE December 2012	3. REPORT TYPE AND DATES COVERED Master's Thesis

4. TITLE AND SUBTITLE An Investigation into Specifying Service Level Agreements for Provisioning Cloud Computing Services	5. FUNDING NUMBERS
6. AUTHOR(S) Nancy J. Kelley	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943–5000	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943–5000	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number _____N/A_____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited	12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

Within the U.S. Department of Defense (DoD), service level agreements are a widely used tool for acquiring enterprise-level information technology (IT) resources. In order to contain, if not reduce, the total cost of ownership of IT resources to the enterprise, the DoD has undertaken outsourcing its IT needs to Cloud service providers. In this thesis, we explore how service level agreements are specified for non-Cloud-based services, followed by determining how to tailor those practices to specifying service level agreements for Cloud-based service provision, with a focus on end-to-end management of the service-provisioning.

14. SUBJECT TERMS Service Level Agreement (SLA), Cloud Service Level Agreement (CSLA)	15. NUMBER OF PAGES 113
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540–01–280–5500 Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

AN INVESTIGATION INTO SPECIFYING SERVICE LEVEL AGREEMENTS FOR PROVISIONING CLOUD COMPUTING SERVICES

Nancy J. Kelley
Civilian, United States Navy SPAWAR SSC Pacific
B.S., Texas A&M University, 1994

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN SOFTWARE ENGINEERING

from the

NAVAL POSTGRADUATE SCHOOL
December 2012

Author: Nancy J. Kelley

Approved by: Man-Tak Shing
Thesis Co-Advisor

James Bret Michael
Thesis Co-Advisor

Peter J. Denning
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Within the U.S. Department of Defense (DoD), service level agreements are a widely used tool for acquiring enterprise-level information technology (IT) resources. In order to contain, if not reduce, the total cost of ownership of IT resources to the enterprise, the DoD has undertaken outsourcing its IT needs to Cloud service providers. In this thesis, we explore how service level agreements are specified for non-Cloud-based services, followed by determining how to tailor those practices to specifying service level agreements for Cloud-based service provision, with a focus on end-to-end management of the service-provisioning.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ACD ..... Automated Call Distribution
ASA ..... Average Speed to Answer
AZ ..... Availability Zone
B1 ..... Boundary One
CA ..... Certificate Authority
CADF ..... Cloud Auditing Data Federation Working Group
CCE ..... Cloud Computing Environment
CDRL ..... Contract Data Requirements List
CIA ..... Confidentiality, Integrity, Availability
CIM ..... Common Information Model
CMWG ..... DMFT Cloud Management Working Group
COI ..... Community of Interest
CONOPS ..... Concept of Operations
CSLA ..... Cloud Service Level Agreement
CSP ..... Cloud Services Provider
DAN ..... Dial Access Network
DAR ..... Data at Rest
DHS ..... Department of Homeland Security
DMFT ..... Distributed Management Task Force
DMZ ..... Demilitarized Zone
DNS ..... Domain Name Services
DoD ..... Department of Defense
DoS ..... Denial-of-Service
DSL ..... Domain Specific Language
EBS ..... Elastic Block Store
EC2 ..... Elastic Compute Cloud
ELB ..... Elastic Load Balancing
EVA ..... Enterprise Validation Authority
FedRAMP ..... Federal Risk and Authorization Management Program
FISMA ..... Federal Information Security Management Act
GPO ..... Group Policy Object
GSA ..... General Services Administration
IaaS ..... Infrastructure as a Service
IT ..... Information Technology
KPI ..... Key Performance Indicator
LAN ..... Local Area Network
NIST ..... National Institute of Standards and Technology
NOC ..... Network Operations Center
OCCI ..... Open Cloud Computing Interface
OMB ..... Office of Management and Budget
PaaS ..... Platform as a Service
PDT ..... Pacific Daylight Time

PII ..... Personally Identifiable Information
PIN ..... Personal Identification Number
PKI ..... Public Key Infrastructure
PMO ..... Program Management Office
QoE ..... Quality of Experience
QoS ..... Quality of Service
Qu4DS ..... Quality Assurance for Distributed Services
RAS ..... Remote Access Server
RBSLA ..... Rule-Based Service Level Agreements
RDS ..... Relational Database Service
RuleML ..... XML-based Rule Markup Language
SaaS ..... Software as a Service
SLA ..... Service Level Agreement
SLALOM ..... Language for SLA specification and monitoring
SLAng ..... XML language for creating SLAs
SLD ..... Service Level Description
SLM ..... Service Level Management
SLO ..... Service Level Objective
SOD ..... Segregation of Duties
SPAWAR ..... Space and Naval Warfare
SSC ..... System Center
TBD ..... To Be Determined
TEC ..... Tivoli Enterprise Console
UPS ..... Uninterruptable Power Supply
US ..... United States
VM ..... Virtual Machine
VPN ..... Virtual Private Network
WAN ..... Wide Area Network
WS ..... Web Service
WS-Agreement ..... Web Services Agreement Specification
WSLA ..... Web Service Level Agreements
WSML ..... HP Web Service Management Language
XML ..... Extensible Markup Language

I. INTRODUCTION

A. EXECUTIVE OVERVIEW

A Service Level Agreement (SLA) is a formal written agreement between the service provider and the service customer or user. It defines the parameters of service the user expects and the provider guarantees to deliver. An SLA can be contractual to deal with providers either outside the organization or in-house. SLAs have been used since the 1960s, mainly by fixed line telecom operators as part of their contracts with their corporate customers. Now they are commonly used in a wide range of service contracts in almost all industries, especially in Information Technology (IT) organizations. The following list is an example of some of those businesses using SLAs: e-Businesses, Internet Service Providers (ISPs), Telecom Companies, IT Service Providers, Sales, Data Centers, Facilities Management, Recruitment, Document Storage, and Business Continuity Services. SLAs are defined at several different levels:

- Customer-based SLA–covers all the services used by one specific customer group.
- Service-based SLA–encompasses one service for all customers.
- Multilevel SLA–covers two or more types of SLAs.
- Corporate-level SLA–covers all generic service level issues for all customers.
- Customer-level SLA–covers all services used by a customer group.
- Service-level SLA–covers a specific service used by a customer group.

This thesis will discuss a portion of a Customer-level SLA covering a DoD software-intensive system.

B. HYPOTHESIS

Effective SLA deployment is a key issue for maintaining stability of services and protecting assets for traditional client-server systems software as well as for Cloud-based systems. SLAs are a means for managing and maintaining network and software services. This thesis analyzes a subset of services provided and consumed by an existing DoD software-intensive system to determine how SLAs can be effectively deployed to define,

manage, and maintain the desired functionality and quality of these services. In addition the thesis explores how the SLAs need to be modified and strengthened for them to be effective for similar services provided by Cloud-based systems. The thesis also touches on the use of machine-readable SLA management of Cloud services and infrastructure.

C. METHODOLOGY

This thesis presents a comparison of the SLA deployment for two different types of software-intensive systems offering the same services. By software-intensive, we mean a system whose requirements are implemented for the most part in software. The first system is a real-world system, which for proprietary reasons will remain anonymous. This system is a typical client-server system. The second system is a hypothetical Cloud-based system providing a subset of the services provided by the first system. A Cloud-based system is a type of distributed system that runs on top of a client-server system and differs from the on-premise distributed system in the following ways:

- Services are delivered over the Internet.
- It is sold on demand.
- The service is managed completely by the provider.
- The provider is fully responsible for the performance, reliability and scalability of the computing environment.

An SLA for a traditional network system covers network support services, application performance, client-side services and server-side services. A Cloud system SLA can only cover server-side services, it cannot directly cover application performance. This thesis will focus on seven server-side services:

1. Help Desk
2. Email Services
3. Web and Portal Services
4. File Share Services
5. Print Services
6. Network PKI Logon Services
7. RAS Services

D. THESIS ORGANIZATION

Following the introduction and background chapters, creation of an SLA is introduced in Chapter 3. In that chapter an example of a human-readable SLA is given for a typical client-server environment. Chapter 4 follows with a case study of the verification and enforcement of an actual SLA. Cloud SLAs (CSLA) are discussed in Chapter 5. A comparison is made between SLAs for traditional client-server systems and Cloud-based systems followed by a discussion of the requirements of a Cloud SLA. The final chapter summarizes the major challenges as well as the benefits that are associated with the use of SLAs.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

Cloud computing is an evolving technology that is becoming more popular. Many businesses and organizations are turning to Cloud computing in order to cut back on costs and increase scalability. This chapter provides a brief summary of Cloud-based systems and Service Level Agreements.

A. WHAT IS A CLOUD-BASED SYSTEM?

Cloud computing is rapidly becoming the new buzz word in information technology. According to the official National Institute of Standards and Technology (NIST) definition, "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [1].

The NIST definition lists five essential characteristics of Cloud computing, three "service models" and four "deployment models" that, in combination, describe ways to deliver Cloud-based services.

1. Essential Characteristics

On-demand self-service. A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider [1].

Broad network access. Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin and thick client platforms (e.g., mobile phones, tablets, laptops, and workstations) [1].

Resource pooling. The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different

physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth [1].

Rapid elasticity. Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the amount of provisionable resources appears to be unlimited and available at any time [1].

Measured service. Cloud-based systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service [1].

2. Service Models

Software as a Service (SaaS). The capability provided to the consumer is to use the provider's applications running on a Cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying Cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings [1].

Platform as a Service (PaaS). The capability provided to the consumer is to deploy onto the Cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying Cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment [1].

Infrastructure as a Service (IaaS). The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying Cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls) [1].

3. Deployment Models:

Private Cloud. The Cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises [1].

Community Cloud. The Cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises [1].

Public Cloud. The Cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the Cloud provider [1].

Hybrid Cloud. The Cloud infrastructure is a composition of two or more distinct Cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., Cloud bursting for load balancing between Clouds) [1].

As the adoption of Cloud computing has grown, individuals and enterprises are finding that Cloud use can help lower IT costs and increase business agility. However, even though Cloud services offer economic benefits, they also come with significant potential risks in safeguarding information and unavailability of data stored in the Cloud, along with the challenges of migrating the non-Cloud IT infrastructure and services to the Cloud.

B. SLA–AN OVERVIEW

An SLA is a formal written agreement that defines the parameters of service the user expects and the provider has guaranteed to deliver. This formal written agreement or legal contract specifies the minimum expectations and obligations existing between the service provider and the customer. In order for an SLA to be fully respected it must be agreed upon by both parties. Coming to an agreement can be difficult as each stakeholder may have a different need or view of what is required.

Service-oriented applications are commonly deployed across a network to provide quick, reliable access to software to users who depend on them to complete job-related tasks. These applications need to be managed and maintained to continuously provide reliable tools for the end users. A SLA can be used to determine whether these services

are being maintained and delivered at the quality level agreed upon. If implemented and maintained correctly throughout its life cycle, an SLA can improve availability and security of the network.

The life cycle of an SLA consists of five phases shown in Figure 1. The first phase of the SLA life cycle is *SLA Template Development*. This is the phase when the SLA templates are developed. Next *Negotiation* is done and the contracts are completed. *Implementation* is the phase involving SLA generation. After the SLA is completed, it is *Executed*, monitored and maintained. The SLA must undergo frequent *Assessment* to determine if any changes are needed. Then the cycle starts over at the beginning and will continue to do so until the SLA is terminated.

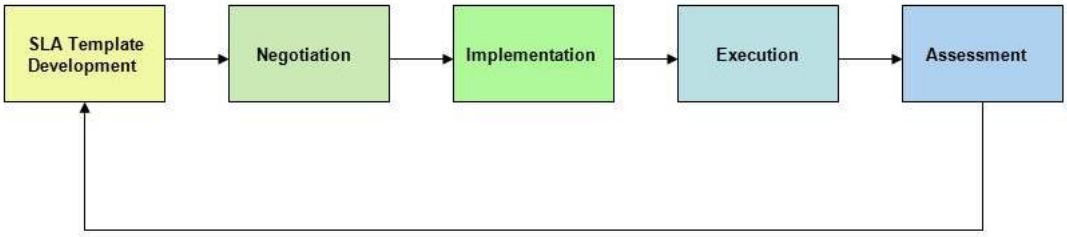

Figure 1. SLA Life Cycle. After [2].

An SLA can be used to address basic performance and availability standards or it can be much more complex. Implementation can start during the development of new systems, maintenance of existing legacy systems, or during post-production support. They can also be used for outsourcing services that were previously performed in-house. To be effective an SLA must be a living document that changes throughout a system's life cycle as updates in software or hardware are needed. This trait will add to the complexity involved in managing and maintaining the SLA.

Communication is a critical component in SLA creation and over the entire lifetime of the SLA. Discussions with the consumers and providers must be held on a

regular basis, especially when updates to the system are required. Another useful form of communication that can be used is customer feedback, especially through the use of surveys. Keeping the SLA updated using information from the meetings and surveys will benefit all concerned. One of the benefits is avoidance of misunderstanding between the stakeholders as a carefully maintained SLA would clarify contract-specific terms and objectives. Another benefit is useful information can be derived from the data collected and reported as required by the SLA. This information is obtained while monitoring parameters such as security, backup procedures, and uptime. The customer can analyze this data to make better decisions while the service provider can use it to determine ways to improve their service.

Effectual communication is only one of the necessary elements required to ensure the development and management of a beneficial SLA. Determining the most effective metrics as well as the responsible parties for maintaining the service levels required are also needed and will aid in successful deployment and operation of the system. The metrics and roles have to be agreed on by all stakeholders, as well as provisions for review and revision, reports required and monitoring tools to be used.

Other requirements are definition of the scope or boundary of the SLA, its duration, what will be considered breaches or violations, and appropriate remedies and penalties. This is a challenging process but once an agreement is reached the specified parameters can be monitored in order to detect agreement-breaches. Then, when an SLA breach is detected, the appropriate mitigation or remedy (predefined in the contract) can be implemented, including penalties and incentives. An example of an incentive is compensating the service provider for compliance. Penalties or fines could be used for noncompliance.

III. CREATION OF MEANINGFUL SLAS

SLAs come in several different varieties. For example, an *Enterprise* SLA is an agreement between the service provider and all customers of an entire organization. A *Customer* SLA is an agreement between the service provider and a specific customer group of the organization. And a *Service* SLA is an agreement between the service provider and the customers of a particular service; this is the type of SLA this thesis addresses.

No matter which type of SLA is used, it must be written so that it encourages the appropriate behavior of the End User as well as the Service Provider. An effective SLA is based on the requirements of the user and the technology offered by the provider and is understandable, measurable, and realistic. It must also successfully manage customer expectations, add measurable value to the services it covers and lead to changes that improve services. To accomplish all this, the key components are metrics–an SLA is only as good as its metrics. The metrics must be easily collected and measure the right performance characteristics required to ensure the agreed upon service levels are met.

A. THE PROCESS

Before a meaningful SLA can be developed, it is necessary that all stakeholders have an understanding of the services that are currently being used within the system and the service levels required by the users. These services will need to be described in enough detail to ensure everyone understands the requirements. It is also important for everyone to understand the goals of the organization and the outputs of the processes or services required by the end users. Interviews or surveys can be used to help establish the end users' needs and requirements.

Once an understanding of the customers or users' needs is reached a baseline of the current service levels must be obtained. The baseline is used to determine performance thresholds for the services that will be monitored. Realistic maximum and minimum thresholds must be found for each service. An effective method to produce the

"baseline" information on the current levels of service being provided is to do benchmark testing. This type of testing aides in quantifying the level of a service and verifying that defined service levels are being met.

The next step in the process is determining what Quality of Service (QoS) parameters or metrics will need to be measured in order to adequately verify compliance. QoS refers to the ability of an application or network to deliver predictable results. Metrics (user metrics, process metrics and performance metrics) are standard measures used to quantitatively evaluate an organization or service to determine if the service provider is meeting its commitments. Determining relevant metrics that can accurately measure service availability and performance is a difficult task, but must be done as accurately as possible to develop a meaningful SLA.

Taking the time to accurately choose and enforced SLA metrics ensure:

- Correct performance attributes are measured to verify that the customer is receiving their required level of service and the provider is accomplishing an adequate level of profitability
- Metrics are collected without difficulty and costly overhead
- Metrics contain relevant, useful data
- Service provider is given an adequate chance to satisfy the customer
- Expectation is for reasonable and attainable performance levels are expected [3]

Developing well-defined metrics is crucial to SLA success. To do so, we need to establish critical processes/customer requirements, develop measures, and establish targets which the results can be scored against. Performance metrics are used in the example SLA discussed in this thesis. By gathering performance metrics, workload characteristics, usability heuristics and quality metrics, and matching them to real-world criteria, a SLA will be developed. It is also necessary to determine what monitoring tools, logs, software agents, and monitoring software are available for use to gather the metrics.

It would be ideal to have the metrics be monitored using existing tools. However, it may be necessary to develop custom tools. When a large number of services are monitore,d it can be more efficient to monitor only a portion or sample of the services. A sample size for generating the mean scores will be used in the example. There are many different categories of performance metrics including:

- Configuration Management
- End-User Problem Resolution
- Network Problem Resolution
- Move, Add, Change
- Information Assurance Services
- Help Desk
- Email Services
- Web and Portal Services
- File Share Services
- Print Services
- Network PKI Logon Services
- RAS Services
- Network Problem Resolution
- End User Problem Resolution
- MAC Request Resolution - Move, Add, Change
- Availability
- Latency/Packet Loss
- Voice and Video Quality of Service
- Information Assurance Services:
- Configuration Management

The performance of each of the service categories may be defined by several different attributes. Common performance attributes include those shown in Table 1.

Table 1. Performance Attributes

Availability	Indicates whether a function or mechanism can continue to maintain operation so that services may be accessed or used as needed. Often expressed as a percentage.
Performance	Expected performance characteristics needed to establish resource commitments. A measure of what is achieved or delivered by a system, person, team, process, or IT service.
Accessibility	Ensures that a technology or content-type is usable or accessible by users when needed.
Integrity	Assures accuracy and completeness as well as adequate performance to specifications.
Reliability	The ability of a system or component to perform its required functions under specific conditions for a specified period of time.
Transparency	Openness and accountability
Confidentiality	Ensures information is not accessed by unauthorized individuals.
Security/Safety	Protection, access control, and authenticity
Functionality	The operations, capabilities and usefulness of an application.
Efficiency	Measurement of response time, interoperability, user accessibility
Recoverability	Refers to being able to restore to a normal condition if a system or an application suffers an unexpected failure of function.
Scalability	Ability to handle a growing amount of work

Each service category can have several tests associated with it. For example, the subset of Service Categories discussed in this thesis could require the following test cases:

1. Help Desk–Average Speed of Answer (Telephone Calls)
   Average Speed of Response (Voice Mail/Email)
   Call Abandonment Rate
   First Call Resolution

2. Email Services–User Email Availability
   Email Server Service Availability
   Email Client Responsiveness
3. Web and Portal Services–End-to-End Performance
4. File Sharing Services–Server Availability
   Client Responsiveness
   Backup/Restore
5. Print Services–Availability
6. Network PKI Logon Services–Client Responsiveness
   Server Responsiveness
7. RAS Services–Service Availability
   Client Responsiveness

B. FORMAT OF AN SLA

SLAs are typically created in plain-text format (human-readable). However, in the case of automated SLA management a machine-readable SLA in addition to a text format should be used. Using human-readable format may introduce less overhead, but because machines cannot read it interoperability is restricted. If only a machine-readable format is used, it is more difficult for humans to interpret. This is why it is helpful to use both formats. Machine readability is also important when supporting large numbers of inquiries when human interpretation, negotiation, and enforcement can take longer and increase the opportunity for errors to arise. Machine-readable SLAs support automated service discovery and the selection of services based on qualities specified in the SLA by interpreting SLA parameters at runtime. This enables automated support for achieving dynamic (i.e., runtime) negotiation and the selection of different quality levels at runtime. To accomplish this requires:

- A sufficiently expressive language for encoding quality-attribute specifications and describing the different levels of qualities that a single-service implementation can provide

- A language for allowing consumers to query providers in order to find services that meet their desired level of quality
- A mechanism that facilitates the selection of quality levels based on runtime conditions such as the number of concurrent users
- An ability to assign priority to requests according to defined rules, such as lowering the priority of requests from a consumer that exceeds a pre-specified number of transactions in a given time period, and
- An ability to support logical expressions for dynamic negotiation of quality-attribute tradeoffs [4]

There are many advantages realized when using SLAs in machine-readable format.

- A machine-readable format supports automatic negotiation between service users and providers [4].
- Sometimes the SLA specifies measures that should be taken by the service user and/or service provider when a deviation from the SLA or a failure to meet the asserted service qualities occurs. In addition, machine-readable SLAs enable measures that can be triggered automatically (e.g., an email notification) [4].
- A billing system can parse the SLA in order to obtain the rules to automatically calculate charges to the service user [4].
- An automated SLA management system that measures and monitors the quality parameters uses the SLA as input [4].

Currently, there are only a few standard formats for machine-readable SLA specifications available, such as IBM's Web Service Level Agreement (WSLA)

framework and Web Services Agreement (WS-Agreement) specification which appears to be more widely accepted.

- **WSLA**–Extensible Markup Language (XML) Schema based language for specifying and monitoring SLAs for Web Services [4].
- **WS-Agreement**–Web Services protocol for establishing agreement between a service provider and a customer using an extensible the XML language [4].

An important difference between WSLA and WS-Agreement is that WS-Agreement XML is highly extensible. This enables the creation of customizable templates necessary in a dynamic SLA.

A WS-Agreement SLA is an XML document that contains the following parts:

- A mandatory unique ID for the agreement followed by an optional name context for the agreement, metadata about the agreement, and user-defined attributes
- The terms of the agreement
- Service terms that identify/describe the services covered by the SLA
- Guarantee terms that specify the levels of agreed upon service quality [5]

The following is a snippet showing two service description terms from the same agreement—one specifying that 32 CPUs will be used to execute a job and the other specifying that 8 CPUs will be used:

```
<wsag:ServiceDescriptionTerm
wsag:Name="numberOfCPUsHigh"
wsag:ServiceName="ComputeJob1">
```

<job:numberOfCPUs>32</job:numberOfCPUs>
</wsag:ServiceDescriptionTerm>
<wsag:ServiceDescriptionTerm
wsag:Name="numberOfCPUsLow"
wsag:ServiceName="ComputeJob1">
<job:numberOfCPUs>8</job:numberOfCPUs>
</wsag:ServiceDescriptionTerm> [5]

This example of machine-readable code is a good illustration of why machine-readable SLAs may not be readable by most humans. It is possible to develop a tool that can convert machine-readable specifications to human-readable specifications and vice versa. Even the CASE tools from late 1980s to early 1990s had this capability. The example SLAs used in the rest this paper will be in human-readable format. Other ongoing research projects for SLA specification include:

- **SLAng**–XML language for creating SLAs. It was developed as part of the Trusted and Quality of Service Aware Provision of Application Services (TAPAS) project at University College London (UCL).
- **Rule-based Service Level Agreements** (RBSLA)–a declarative rule based SLA language based on XML-based Rule Markup Language (RuleML) and interoperable with other languages. It focuses on knowledge representation concepts for Service Level Management (SLM) of IT services.
- **HP Web Service Management Language** (WSML)
- **Apache Neethi–**a framework enables Apache Web services stack to use WS Policy as a way of expressing requirements and capabilities.
- **SLALOM**–a language for SLA specification and monitoring

C. COMPONENTS OF AN SLA

The components used for an SLA vary depending on the type and needs of the organization. All SLAs should at least cover:

- Introduction and Purpose / Descriptions of Service
- Services to be Delivered / Service Standards
- Performance, Tracking and Reporting / Duration
- Problem Management / Roles and Responsibilities
- Fees and Expenses / Evaluation Criteria
- Customer Duties and Responsibilities

SLAs have two distinct major parts: a technical section and a contractual or legal section. Service expectations are defined in the technical section. Test procedures are also described there. The contractual section defines such things as fees, non-performance penalties, and schedules. A typical SLA consists of several sections or subcomponents.

Table 2. Components of an SLA From [6]

(continued on next page)

Section	Description
Executive Summary	This is a summary section describing the general purpose of the Document – to meet or exceed the service-level measurements that are mutually agreed on. This should include the purpose of the document and the duration of the agreement. It should define the stakeholders or ownership for the service levels agreed on within the enterprise and the scope of the areas that are included.
Description of the Services	Within this section is a detailed description of each of the services and the committed performance levels associated with them.
Service-Level Definitions	For each functional area (e.g., email services or print services), a minimum number of key SLAs should be included. A sample of the description of the data points that should be prepared for each SLA are: · Definition — The key business service (function/process/procedure) that is being measured, reported and continuously improved. · Measurement time frame — The days, dates and times when the defined SLA is to be measured, usually indicating the inclusion or exclusion of recognized national holidays. · Assumptions/responsibilities — Statement of specific requirements that must be met by the IS organization and business units to remain in compliance with the SLA.

Section	Description
	· Service-level metric — Relevant measurement of required work performed by the IS organization. Although these service levels are commonly measured in percentage terms, IS organizations need to design pertinent measurements that can be expressed in terms of business performance. · Measurement formula — Description of mathematical formula and example. · Reporting measurement interval period — Reporting period for measurement that determines exceeding, meeting or not meeting target SLAs. · Data sources — Location(s) from where data is collected, including a description of what is collected, where it is collected, how it is stored, and who is responsible for it. · Escalation activity — Describes who is notified and under what conditions as out-of-compliance situations occur, including day-to-day and measurement period out-of-compliance situations. · Escalation management — Identifies to whom the out-of compliance activities are forwarded on recognition. · Contractual/exceptions/penalties/rewards — Describes, and refers to, any contractual exceptions, penalties and rewards that are included in the contract. · Reward/penalty formula — Description of mathematical formula and example. · If the enterprise employs severity or priority codes, generally they would be described within this section.
Service-Level Management	Numerous processes need to be documented regarding the management of service levels, including: measurement tracking and reporting, business continuity, problem escalation guidelines, service/change requests, new services implementation, approval process and the service-level review process.
Roles and Responsibilities	This section outlines the roles and responsibilities of all the parties to ensure that the service objectives are met. This includes the IS organization, the various business units, and any external services providers that may be used. It should also identify governance committees or key stakeholders managing this contract.
Appendixes	Appendixes are used to include additional information that might be relevant to the agreement, such as the hardware and software supported.

Table 2 (continued)

You may also need to include other requirements or subsections such as:

1. Delivery of Services: Describes how the service provider will deliver the service or services.
2. Metrics: How will the service delivery be measured?
3. Key Performance Indicators (KPIs): Describe the KPIs and the responsible party for producing the KPIs.
4. Schedules: Timelines for testing, reporting, and mitigation.
5. SLA Changes: Will the SLA remain fixed or will it be allowed to change when necessary? If changes are allowed, what are the procedures used?
6. Support Hours: Describes the normal support hours and the after-hours support times. Also describes any additional charges that may be incurred for support outside of normal hours.
7. Record Retention: How long will records be kept? How will they be disposed of when no longer needed? How will confidential information be protected?

The number and type of requirements needed to make the SLA effective is determined by the type of services being provided. Some SLAs might only require a portion of these items to be sufficient. But to be effective the SLA must include how service-level thresholds will be measured or monitored along with the required reports, data sources, and contract exceptions. It must also include penalties/rewards for noncompliance/compliance, as well as termination guidelines. There are at least three situations in which an SLA may be terminated.

1. The service defined in the SLA has completed.
2. The time period over which the SLA has been agreed upon has expired.
3. The provider is no longer available [7].

D. SLD TEMPLATE EXAMPLE

The Service Level Descriptions (SLDs) are critical components of an SLA. They contain the compliance objectives of each service to be monitored and the processes to be

used to assess them. The SLDs are basically the test cases. An SLA may contain many SLDs. Therefore it is desirable to use a template when creating them. SLD templates can be found on the Internet and in many books. The templates can be tailored to the needs of the organization.

An example of creating an SLD or test case will be given in this chapter. A SLD includes how service-level thresholds will be measured or monitored along with the required reports, data sources, and contract exceptions. It can also include penalties/rewards for noncompliance/compliance. The SLDs need to be understandable, measurable, and realistic. The following diagram is an example of an SLD template. This is the template used in this thesis.

<table>
<tr><td colspan="2">SERVICE NAME:</td><td>SLA:</td></tr>
<tr><td colspan="3">Service Description:</td></tr>
<tr><td colspan="2">Performance Category:</td><td>Increment</td></tr>
<tr><td colspan="3">Performance Category Description:</td></tr>
<tr><td colspan="3">Measurement CONOPS:</td></tr>
<tr><td>Who:</td><td colspan="2">Frequency:</td></tr>
<tr><td>Where:
User Population:
Sample Size:
Sample Unit:
Where Measured:</td><td colspan="2">How Measured (i.e., captured):
Measurement Formula:
Frequency of Measure:
Weighting (as applicable):</td></tr>
<tr><td>Aggregation of Data:</td><td colspan="2"></td></tr>
<tr><td>SLA Success Criteria</td><td colspan="2"></td></tr>
<tr><td>SLA Target</td><td>Server Availability</td><td></td></tr>
</table>

Figure 2. Example SLD Template

Most of the information required in this SLD template is self-explanatory. However, a brief description of each entry is given.

1. **SERVICE NAME:** Service type or name

2. **SLA:** SLA document version number

3. **Service Description:** Specific description of service

4. **Performance Category:** Service type or category

5. **Increment:** Increment of test case

6. **Performance Category Description:** Describes the service and/or sub-services that must be measured in order to determine SLA-compliancy of the service.

7. **Measurement CONOPS:** Specific information defining the goals and objectives of the SLA measurement:
   - Strategies, policies, constraints
   - Specific processes used
   - Statement of responsibilities
   - Exclusions/Inclusions

8. **Who:** This states "who" will perform the measurement or testing.

9. **Frequency:** How often will measurements be gathered to validate SLA-compliancy?

10. **Where:** Where the test will be conducted?

11. **User Population:** Who are the users of this service?

12. **Sample Size:** The number of observations or measurements needed to represent the system architecture.

13. **Sample Unit:** The element or set of elements considered for selection in the sample.

14. **Where Measured:** Where will the measurement be made?

15. **How Measured (i.e., captured):** Testing or measuring procedure used to obtain SLA-compliancy data. What types of testing tools will be used?

16. **Measurement Formula:** Formula used to calculate a score or percentage

17. **Frequency of Measure:** Period of time over which measurements are to be taken.

18. **Weighting (as applicable):** Describes score weighting if used.

19. **Aggregation of Data:** Process in which information is expressed in a summary form for the purposes of analysis or reporting. The value is derived from the aggregation of data occurrences within the same data subject.

20. **SLA Success Criteria:** States what requirements have to be met to pass the SLA.

21. **SLA Target:** Score or percentage needed to meet compliancy-specifications.

Using the SLD or test case for Network PKI Logon Services, an example is given to show what information would be needed to fill out the required SLD.

Assumptions used in this example:

- The network service provider is a Contractor.

- A customer survey was used to gain the required information to determine the best metrics to use.
- It was agreed upon that the services would be tested and reported on monthly.
- The organization is located at several geographically separated locations.
- Not all of the sites are fully operational to full performance requirements.
- A smartcard is used by the organization for user authentication.
- The organization employs a typical wide area network (WAN).
- A baseline of the current service levels has been obtained.
- The service provider is referred to as the Contractor.
- The customer is referred to as the End-User or User.

SERVICE NAME: NETWORK PKI LOGON SERVICES	**SLA: 100**
Service Description: Public Key Infrastructure (PKI) is a system of digital certificates, Certificate Authorities, and other registration authorities that verify and authenticate the validity of each party involved in an Internet transaction.	
Performance Category: Network PKI Logon Services	**Increment 1 SLA: 100.1**
Performance Category Description: Network PKI Logon Services is the Contractor-provided service for end-user access to the Enterprise Validation Authority (EVA) Server and Active Directory in support of end-user logon to the network. This SLA excludes logon to the network through Remote Access Server (RAS0 and web-based activities. The performance measure for Network Logon Services is Client Responsiveness. **Client Responsiveness:** Percentage of transactions that fall within the response time to successfully complete cryptographic network logon from a network-attached workstation. The measurement is the time required for the supporting infrastructure to process the end-user cryptographic log on request. Client Responsiveness shall be measured by sampling. Initial sample size shall consist of 50 measures that provide a representative sampling of the deployed architecture (e.g., local and distant user-to-server farm connections).	

<table>
<tr><td colspan="4">Measurement CONOPS: The measurements are obtained manually by using a stopwatch. After inserting the smart card into the smart card reader and entering the appropriate Personal Identification Number (PIN) number when prompted, the start point for the measurement is the time from when the “Return” key is depressed. The stopping point will be when the screen depicting “Loading your person settings” is shown on the monitor. Every time measurement is recorded and forwarded to a collection point – to the SLA Collector. All time trials are used to compute the percentage above or below the threshold value. All measurements for this SLA will be taken by an administrator and periodically observed by the end user. Test will occur during the morning hours, 0800 – 1000 local.
The test will utilize a smartcard with minimum 32 KB chip.</td></tr>
<tr><td>Who: Contractor</td><td colspan="3">Frequency: Monthly</td></tr>
<tr><td>Where: Client Site

User Population:
All Users.

Sample Size: 50

Sample Unit:
Test Account at sample site

Where Measured:
Client workstation to EVA Server and Active Directory</td><td colspan="3">How Measured (i.e., captured):
Stopwatch test at sample sites

Measurement Formula:
Number of attempts successful within the required Time Interval during the test period / Total number of attempts during the test period

Frequency of Measure:
0800–1000 Local time

Weighting (as applicable):
Equal weighting</td></tr>
<tr><td>Aggregation of Data:</td><td colspan="3">Performance data for sites that have not yet achieved Full Performance will be aggregated at the site level and the SLA targets will apply at the site level.
Performance data for sites that have achieved Full Performance will be aggregated at the user population level and the SLA targets will apply at the user population level.</td></tr>
<tr><td>SLA Success Criteria</td><td colspan="3">All targets must be met to pass the SLA.</td></tr>
<tr><td>SLA Target</td><td>Client Responsiveness</td><td>Time Interval</td><td>Percentage Complete</td></tr>
<tr><td></td><td></td><td>≤ 30.0 sec</td><td>≥ 90.00%</td></tr>
</table>

Figure 3. SLD Example (SLA 100.1)

The organization in this case study uses PKI extensively and therefore it is a critical issue to include in their SLA. PKI supports the secure exchanging of information. In this case, PKI is used with smart card to provide extra security by implementing two-

factor authentication (something the user has and something the user knows). As with all security measures, PKI Network Login application must be cost effective and usable. When using PKI, time and cost savings will be gained by reducing the amount of time that users spend logging in to multiple applications each day. Proper protection of information also helps to reduce financial losses that result from the theft of unprotected information. Smart cards provide many advantages such as security, ease of use, and portability. One problem of using PKI Network Login is it does add to the network load resulting in degradation of bandwidth and increased response time for users. To ensure the system architecture is effectively supporting PKI Network Logon implementation, especially during peak periods of use, Client Responsiveness should be tested as described in the example SLD.

A sample size of fifty measures was chosen for this SLD after testing determined this would be enough to adequately represent the deployed architecture. Due to the size of the organization's network, testing every single device on the network would prove to be inefficient as it would cause severe overload and congestion of the network. By using a controlled amount of data we take advantage of the process known as data aggregation to convey a close approximation of what we would obtain by testing all devices. The testing will occur during the morning hours, 0800 – 1000 local, when a large number of users will be logging in. Equal weighting was used as all users should experience the same access time. The SLA will be passed only if greater than or equal to 90.00% of the clients tested are successful with a logon time of less than or equal to 30.0 seconds. Logon time should never be more than 30.0 seconds, this was determined from the customer survey to be longest logon time acceptable by the user. No automatic monitoring software was available for testing; therefore this test will be conducted manually using a stopwatch with results recorded in a spreadsheet for evaluation.

Data aggregation was also based on the operational status of the site being tested. This was done to prevent penalizing sites based on the analysis of other sites that are not up to the same level of development. Sites that are not fully operational are aggregated at the site level. Sites that are fully operational are aggregated at the user population level. The remaining services the organization was concerned with are as follows:

1. Help Desk
2. Email
3. Web and Portal Services
4. File Share Services
5. Print Services
6. RAS Services

To complete the SLA for this organization, the same SLD template would be used to create the test cases for these services. Readers can refer to the Appendix for details of the SLDs.

E. SLA SIGNOFF

The initial draft SLA detailing responsibilities, penalties, incentives, deliverables, documentation, methodology for verification, escalation procedures, and management of the SLA must be mutually agreed upon by all stakeholders. It is also essential to include contracting officials in SLA negotiations or at least allow them to review the draft before the negotiation process begins. A meeting between the Customer and the Service Provider should be held to discuss the draft until an agreement is reached. At this time all stakeholders will have the opportunity to review the draft, bring up any questions they have, and present suggestions. Monitoring tools or products should be discussed as well to determine which would be the most cost effective. Another topic for this meeting would be reports, their format, their periodicity, and their distribution. Reports are extremely important in that they provide the mechanism by which management can determine whether actual performance meets service thresholds. The reports and other deliverables are usually outlined in the Contract Data Requirements List (CDRL).

The meetings need to continue until an agreement is reached. Feedback from all meetings would then be used to finalize the document. Meetings of this type will be required periodically throughout the duration of the SLA; an SLA is not a static document, it can be changed as needed. Once all changes are made, the stakeholders must sign the SLA indicating their approval.

V. VERIFYING AND ENFORCING SLA COMPLIANCE

After an SLA has been finalized and accepted by all stakeholders it can be executed. This is when the SLA must be verified or monitored as well as enforced using the agreed upon documented methods. The purpose of this stage is to measure and quantify the service quality expected by the customer as defined in the SLA. Proactive steps will ensure a high degree of consistency in service is delivered. All stages of the SLM life cycle are affected by the data gathered during this stage.

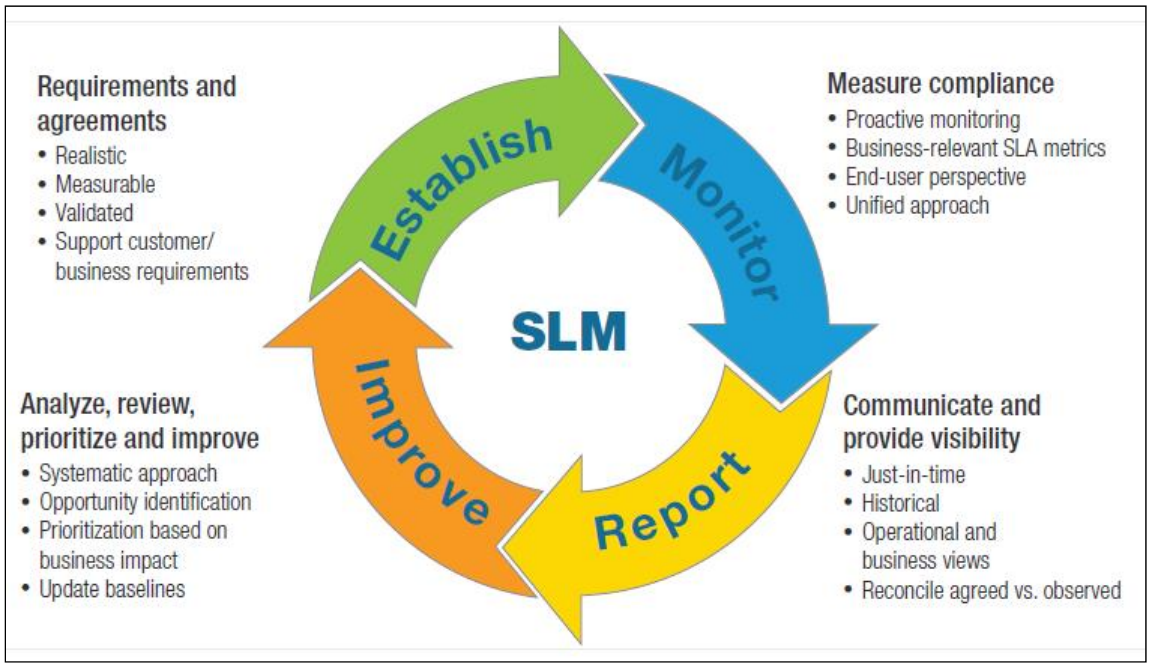

Figure 4. The SLM Life Cycle. From [8]

A. VERIFICATION

Verification is critical to the success of any SLA. It is impossible to tell if the terms of an SLA are being met without monitoring and verification. Verifying service level compliance involves the process of quantifying service quality using metrics that measure service performance and end-user satisfaction. To do this effectively, the actual achieved service levels (QoS parameters) are measured; these measurements will also

serve as inputs for the SLA report. Ideally, a fully automated method will be used for monitoring and verification of SLA compliance. However, it is not always possible to fully automate the process as the SLDs in this thesis illustrated. Also automated tools may not be affordable, some combination of manual and automated processes can be used instead.

The product of testing and verification is a SLA results report. The Service Level Manager and stakeholders review these reports. The report consists of the actual scores each service obtained during the most recent test-and-verification cycle. The Service Level Manager uses this information to determine the required time to mitigate any non-compliance and what penalties apply. The aim of a Service Improvement Program (SIP) is to mitigate service violations and to gradually improve the level of service, akin to the software reliability growth processes used in software engineering.

The SLA report indicates areas for improvements. Improvements could necessitate changes in requirements of the SLA itself. The key idea here is continuous improvement of the level of service being provided.

B. ENFORCEMENT

SLA enforcement follows the verification and monitoring step in the SLA management life cycle. Service level guarantees must be enforced with contractual provisions for failure to comply with the SLA. These provisions include penalties for violations and compensation or incentives for meeting or exceeding the service-level guarantees in the contract. The challenge here is to strike the right balance between penalty and reward to motivate the service provider to fulfill or exceed their obligations. In addition to financial penalties the SLA can include escalation policies requiring that any non-compliance issues receive immediate attention not only from the individuals directly responsible for maintaining the service but also higher level management in the provider organization. Examples of types of penalties for SLA non-compliance include:

- A decrease in the agreed payment for using the service, that is, a direct financial sanction [9]

- A reduction in price to the consumer, along with additional compensation for any subsequent interaction [9]
- A reduction in the future usage of the provider's service by the consumer [9]
- A decrease in the reputation of the provider – and subsequent propagation of this value to other clients [9]

For instance, one penalty scheme is to assign penalties based on the weighting of the unmet Service Level Objective (SLO). SLA breaches can be defined by a broad category of violations applying a weight strategy such as the following list describes.

- 'All-or-nothing' provisioning: provisioning of a service meets all the SLOs – that is all of the SLO constraints must be satisfied for a successful delivery of a service
- 'Partial' provisioning: provisioning of a service meets some of the SLOs – that is some of the SLO constraints must be satisfied for a successful delivery of a service
- 'Weighted Partial' provisioning: provision of a service meets SLOs that have a weighting greater than a threshold (identified by the client) [7]

In general, penalties should be accessed soon after a SLA to focus the attention of the provider on mitigating the non-compliance issues.

C. SLA CASE STUDY

We describe here the SLA results for the example traditional client-server system. The actual SLA/contract covered may more services than the seven discussed in this thesis. Network services were provided by a contractor and both the contractor and the user participated in the monitoring of the SLA requirements. Most of the monitoring done by the end-user had to be done remotely for economic reasons because the sites were

geographically dispersed. The SLA/contract was in effect for ten years. Three months of data for each of the service categories listed in section 1-C was evaluated.

Metrics were collected using a combination of automatic scanning tools, custom scripts, and manual processes where no automated tools were available. A machine-readable SLA was not used and no actual data was gathered at runtime. The scan data was reviewed to determine if any inconsistencies existed that may have been generated by false findings. The data came from several different sources, but no automated tool for data fusion and decision making was available. Instead, scripts were used to combine and filter the data and generate measures and metrics.

The SLA required assessment of several different sites per month with a report due by the fifth day of the month following the month in which the monitoring was completed. The report was automatically generated using custom scripts to format the data. The report contained an executive summary listing the scores for each site as well as the overall score for the entire network. Also included in the report was a separate section for each site monitored that month listing the specific findings for the site. All raw scan data was also provided as required in the SLA. The report was electronically disseminated to stakeholders in the provider's organization and in the customer's organization to review and use as needed. The end-user used the report to ensure services were being provided as required. The contractor used the reports in order to determine where they needed to mitigate problems and improve services to better end-user satisfaction. Upper level management from both entities used the reports to discuss future changes and budgeting issues.

During the execution of this SLA no awards were ever earned by the Network Services provider. In fact, no awards were included in the SLA contract. The contractors were scored either a "Pass" or a "Fail" for each month and they were paid in full for passing months. In months where there were failures, their payment was decreased by an amount based on the cause and impact of the failure. Table 3 shows the results of the SLA for three months.

Table 3. SLA Results

Service	Results	Recommendation
Help Desk	Failed 1 month	Better training for Help Desk operators.
User/Server E-mail	Failed all 3 months	Ensure the probes are online while testing occurs to ensure data presented is accurate
Web and Portal Services	Failed all 3 months	Increase virtual throughput of network segments using newer data stream compression techniques.
File Share	Passed all 3 months	None
Print	Passed all 3 months	None
Network PKI Logon Services	Passed all 3 months	None
RAS	Passed all 3 months	None

Other types of failures also occurred within the three months studied such as:

- Antivirus not running or out-of-date
- Critical patching not done
- Firewalls configured incorrectly
- Unauthorized shares found on the network
- Incorrect Group Policy Objects (GPOs) applied to devices

Failures resulted from a variety of situations such as:

- Age of the infrastructure: end-of-service, end-of-life devices and applications
- Bugs in patches that prevented the patches from being installed correctly
- No actual patches or remediation techniques available
- Funding unavailability
- Inadequate testing done prior to operational deployment
- Inability of scanning tools to access devices remotely
- Bugs in scanning tools causing them to verify requirements incorrectly

A review of the monthly findings by both parties led to both recommendations for improvements and some disagreements where requirements were not clearly defined. In some instances the requirements were interpreted very differently by the customer and provider. Even slight misinterpretations can be an issue. Situations such as this required SLA updates in order to define the requirements more clearly so that they were interpreted the same by all stakeholders. Improving the precise meaning of the requirements is task for which formal methods could be applied, but formal methods were not used in this real-world example.

V. USING SLAS TO MAINTAIN CLOUD OPERABILITY

The use of Cloud computing services is increasing. Organizations are finding it necessary to develop SLAs that can be applied effectively to Cloud-based services delivered over a local area network (LAN) or wide area network (WAN). NIST refers to SLAs for a cloud service as a Cloud Service Level Agreement (CSLA). The CSLA is a portion of the service contract where the Cloud service level expected by the customer is defined. The same SLA standards should be applied in a Cloud computing environment (CCE) as would be when outsourcing network services. However, when employing a CCE the customer must be more careful when evaluating their needs, what is negotiable, and how much guarantees and assurances are worth to them.

A. CLOUD STANDARDS

It is important to note that there are currently gaps in the standardization of many aspects of Cloud computing. This makes it difficult to monitor across multiple Clouds as no common set of metrics exist which can be used across services from different Cloud providers. Work is currently being done to develop Cloud standards for monitoring, best practices, basic metrics, as well as standardized machine-readable languages for creating SLAs.

Using a standardized machine-readable language to create an SLA aides in monitoring more complex configurations such as Cloud-based systems. It also enables the use of automated verification and monitoring tools which is helpful in cases where Cloud-based systems have been adopted or IT service providers are managing large numbers of SLAs for different customers and different types of services. There are currently a few commercial SLM tools available such as HP OpenView, BMC Patrol, IBM Tivoli, Microsoft Application Center, and CA Unicenter. However, these tools can only handle simple static rules with a limited set of parameters. Several groups are working to come up with more robust solutions. The Distributed Management Task Force (DMTF), formerly “Desktop Management Task Force,” is one industry organization that is currently developing standards for SLA management and compliance in the Cloud.

Their continuing work has led to the development of several standards for Cloud SLA development, monitoring, and verification (Table 4).

Table 4. DMFT Cloud Standards Development

Common Information Model (CIM)	describes the management aspects of services and resources at various levels of abstraction and decomposition.
Cloud Auditing Data Federation Working Group (CADF)	a group developing open standards for federating Cloud audit information. "The CADF is also working closely with the DMTF Cloud Management Working Group (CMWG) to reference their resource model and interface protocol work" [56].
Open Cloud Computing Interface (OCCI)	a set of open community led specifications delivered through the Open Grid Forum. It is a vendor independent, platform neutral, general purpose set of specifications for Cloud-based interactions with resources. "OCCI provides a protocol and API design components, including a fully-realized ANTLR grammar, for all kinds of Cloud management tasks. The work was originally initiated to create a remote management API for IaaS model based services, allowing for the development of interoperable tools for common tasks including deployment, autonomic scaling and monitoring. It has since evolved into a flexible API with a strong focus on integration, portability, interoperability and innovation while still offering a high degree of extensibility" [10].

The U.S. General Services Administrations (GSA) established the Cloud Computing Program Management Office (PMO) in April 2009. Their work includes developing a standardized approach to security assessment, authorization, and continuous monitoring for Cloud-based systems with the development of the government-wide Federal Risk and Authorization Management Program (FedRAMP). An Office of Management and Budget (OMB) policy currently requires federal agencies to use FedRAMP when authorizing Cloud services.

The purpose of FedRAMP is to:

- Ensure that Cloud-based services used government-wide have adequate information security;
- Eliminate duplication of effort and reduce risk-management costs; and
- Enable rapid and cost-effective procurement of information systems/services to Federal agencies [11].

FedRAMP is the result of collaboration with GSA, NIST, the Department of Defense (DOD), and the Department of Homeland Security (DHS). It provides standard contract clauses and general guidelines on SLAs in Cloud computing environments. The FedRAMP approach uses a " 'do once, use many times' framework to save cost, time, and staff requirements to conduct redundant agency security assessments" [12]. FedRamp is also Federal Information Security Management Act (FISMA) compliant. FISMA requires that Federal agencies adequately safeguard their information systems and assets.

B. CLOUD-BASED SYSTEM SLAS VS. TRADITIONAL CLIENT-SERVER SYSTEM SLAS

When developing SLAs for Cloud-based systems, different measures for service performance may be required than what are used for traditional client-server system SLAs. The Cloud is comprised of both the applications delivered as services over the Internet and the hardware and systems software that provide those services. End users access Cloud services through the networks they are connected to. Cloud SLAs are then written with the assumption that the client-server system is up and running and the performance measures are defined from the time a service request is received by the service provider to the time that a service is performed by the provider. In other words, the response time for Cloud services is measured at the server ends and excludes the network communication time while the response time for services in a traditional client-server system is measured at the client end and includes the communication time through the network. The time taken for the request to reach the provider and the time taken for the result of the service to reach the requester would be excluded from the Cloud SLA

performance calculations. These durations would be part of the SLA for the network service. Cloud SLAs must also emphasize service reliability rather than component reliability because end-users will expect services to be reliable and to meet a performance or quality of experience (QoE) standard. QoE in Cloud-based systems can be measured in terms of response time.

A Cloud-based system presents a dynamic environment in which different routes for requests for network bandwidth and computing services are used. The technology is not quite as stable as with an on premise-dedicated server provided by a traditional client-server system. While a Cloud-based system can manage larger load peaks it must also be taken into consideration that storage capacity or a requested service may be in use at the same time by others on the same platform. This can cause a variation in response time as the system may have to re-provision space or services. Additionally, Cloud services may be subject to load fluctuations as they are delivered over the Internet which is also subject to load fluctuations and can lead to outages. Outages in the Cloud are slightly different than what is seen in a client-server system. When a traditional server experiences an outage it causes a direct outage of the service or application. In a Cloud the outage of a single server might only cause a temporary reduction in performance.

Cloud SLA violations are more apt to occur during load fluctuations. These fluctuations usually cannot be predicted therefore a static testing schedule may not be acceptable. A Cloud service provider could use automatic negotiation and dynamic SLM processes to deal with the load fluctuations and provide better service for their customers and at the same time lower the rate of SLA violations caused by the fluctuations. Developing processes using standards and best practices for managing SLAs can make runtime interpretation possible as well as the dynamic discovery and selection of resources and services. Resources such as storage, process, memory, and network bandwidth could be reassigned based on consumer need and network load to allow the network to run more efficiently and improve the customer's experience. This type of SLA management could only be used by the service provider and possibly large enterprises such as the United States (US) DoD who are able to obtain more in-depth information about how the service provider's services are managed and how the users' data is stored.

Service providers typically consider technical issues related to operations of distributed systems to be proprietary and at a minimum the internal workings should be made transparent to the customers. The U.S. government has leverage to obtain the technical information, whereas most smaller enterprises cannot and are only concerned with the levels of a relatively small list of QoS parameters like number of servers of a particular type and type of encryption.

The risks and methods used to monitor the Cloud are somewhat different from traditional client-server systems. For example, the customer may have an agreement with one provider, while the service is actually delivered by various subcontractors of the Cloud provider. In this case the consumer has no explicit contractual relationship with any of these additional providers. Adding to this problem, the customer might not have any knowledge of the subcontractors unless the provider chooses or is required to disclose them. This would make it difficult if not impossible to create a CSLA to adequately monitor service levels. Situations like this are not good for all parties as even the "unknown" parties could incur risks that the consumer could be responsible for.

Other risks may not be unique to the Cloud but they are magnified by its use. Some of the risks both computing paradigms share are:

- Loss of business focus
- Solution does not meet business and/or user requirements; does not perform as expected.
- Not all requirements identified or wrong solution selected.
- Gaps between business expectations and service provider capabilities, contractual errors
- Compromised system security and confidentiality
- Invalid transactions or transactions processed incorrectly
- Expensive compensating controls
- Reduced system availability.
- Questionable integrity of information/data
- Poor software quality
- Testing not adequate, high number of failures

- Resources allocated insufficiently
- Responsibilities and accountabilities are unclear
- Inaccurate billings
- Reputation
- Potential for fraud

In addition, Cloud computing also has some unique risks:

- **Greater dependency on third parties:**
  - Increased vulnerabilities in external interfaces
  - Increased risk in aggregated data centers (security, privacy, and economic risk)
  - Immaturity of the service providers with the potential for service provider going out of business
  - Increased reliance on independent assurance processes
- **Increased complexity of compliance with laws and regulations:**
  - Greater magnitude of privacy risk
  - Transborder flow of personally identifiable information
  - Affecting contractual compliance
- **Reliance on the Internet as the primary conduit to the organization's data introduces:**
  - Security issues with a public environment availability issues of Internet connectivity
- **Due to the dynamic nature of Cloud computing:**
  - The location of the processing facility may change according to load balancing
  - The processing facility may be located across international boundaries
  - Operating facilities may be shared with competitors
  - Legal issues (liability, ownership, etc.) [13]

Confidence and assurance in the Cloud is quite different too. With traditional client-server systems assurance is better understood as the boundaries and frameworks are well defined as illustrated in Figure 5. The SLA boundaries for this network are very clear; they are the End-User Site and the Internet Provider. Assurance can be provided by reviewing historical data from traditional client-server systems while little or no historical data may be available for the Cloud.

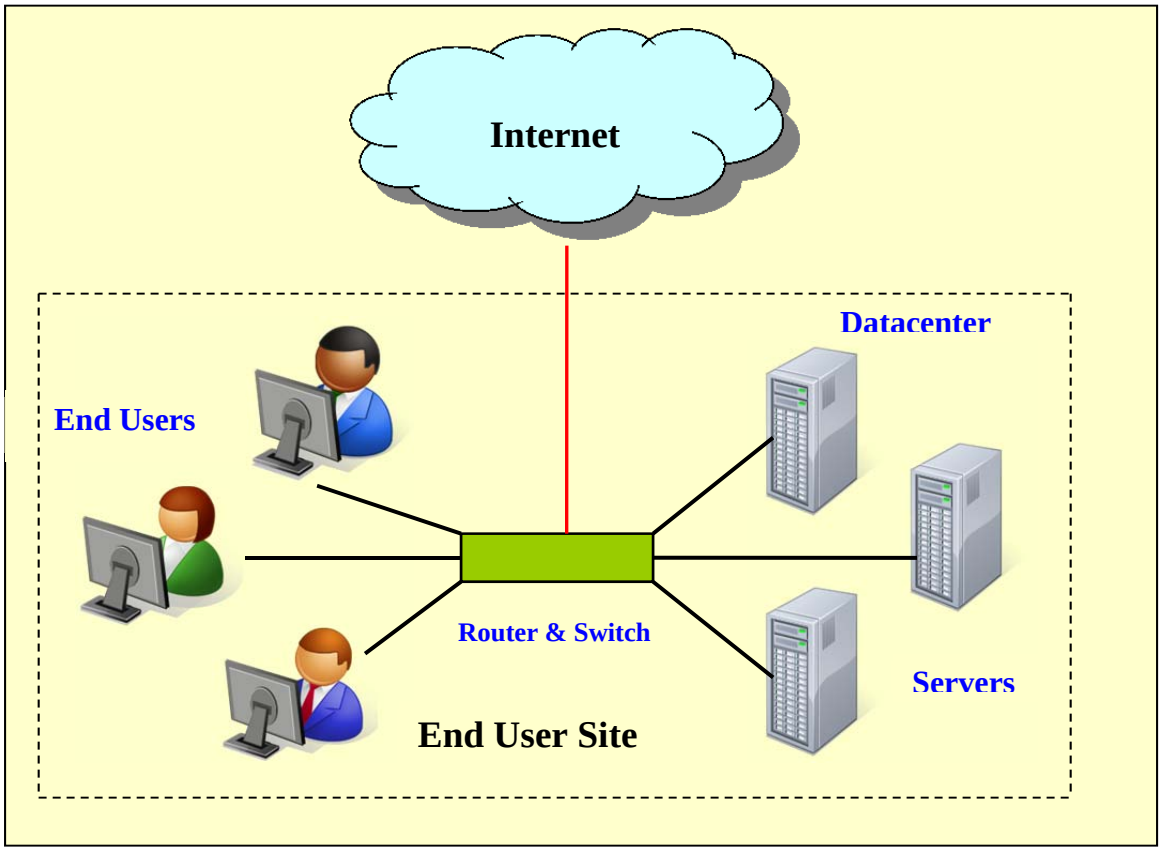

Figure 5. Client-server System Example

Figure 6 is a representation of a simple Cloud-based system and shows the more complex boundaries associated with it. In this example, the boundaries include the End Users' site, the ISP providing the Internet provider, and the Cloud Service Provider's site or sites. A Cloud-based system can be much more complex, comprised of multiple individually managed domains and contain a mixture of Telco and IT equipment and services increasing the challenge to provide end-to-end SLA monitoring. Providing assurance in the Cloud will require new methods. Cloud Systems also utilize shared resources which are sometimes located in different geographic locations making it a major challenge to define boundaries and isolate client-specific transactional information. Methods that focus on transactional data can also be less effective in the Cloud. For these reasons continuous, real-time, process-aligned methods are needed to provide assurance in the Cloud.

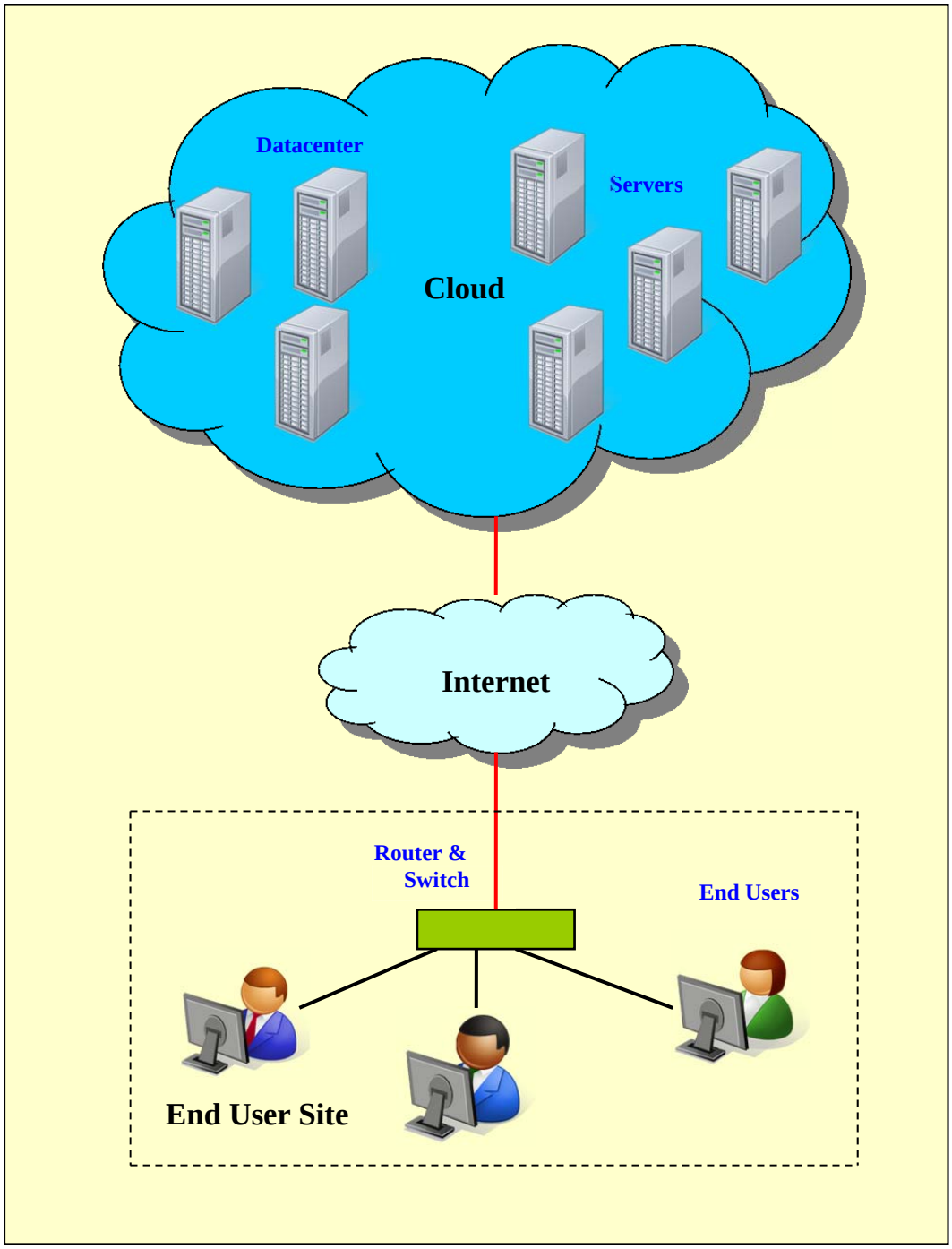

Figure 6. Cloud-based System Example

C. REQUIREMENTS FOR CLOUD SLAS

Leveraging Cloud computing can provide enterprises with significant cost savings and increased efficiency. Some of the key business benefits gained are cost containment, immediacy, scalability, efficiency, resiliency, and availability. However, along with the benefits come risks and security concerns to be considered. When developing an SLA for a Cloud-based system we need to keep in mind the new issues being brought into play. Clear expectations must be stated regarding the handling and usage of data. Many changes in governance issues must also be addressed. The following table describes some of these issues.

Table 5. Five Key Governance issues around Cloud Computing [14]

(continued on next page)

Issue		Considerations
Transparency	Effective and robust security controls must exist to assure information is secured against unauthorized access, change or destruction.	‣ How much transparency is required? ‣ What needs to be transparent? ‣ Will transparency aid malefactors? ‣ Who will have access to customer information? ‣ Does the provider maintain Segregation of Duties (SoD) between its employees? ‣ How are different customers' information segregated? ‣ What controls are in place to prevent, detect and react to security breaches?
Compliance	Data may not be stored in one place and might not be easy to retrieve.	‣ If data is requested by authorities, ensure it can be provided without compromising other information. ‣ Audits are done by legal, standard and regulatory authorities that demonstrate there can be plenty of overreach in such seizures.

Issue		Considerations
Trans-border Information Flow	The actual physical location of the information may be an issue.	‣ Physical location determines jurisdiction and legal obligation. ‣ Personally identifiable information (PII) laws vary greatly in different countries. ‣ Something may be legal in one country but illegal in another.
Privacy	Providers must prove that privacy controls are in place and can prevent, detect and react to security breaches effectively.	‣ Before service provisioning commences, information and reporting lines of communication need to be established. ‣ Test communication channels periodically during operations.
Certification	Customers must be assured that provider is doing the right things.	‣ Third-party audits and/or service auditor reports should be used to provide assurance.

Table 5 (continued)

When using Cloud services data and applications are controlled by a third-party, "The cloud services delivery model will create clouds of virtual perimeters as well as a security model with responsibilities shared between the customer and the cloud service provider. This shared responsibility model will bring new security management challenges to the organization's IT operations staff" [15].

- Adequate transparency from Cloud services to manage the governance (shared responsibilities) and implementation of security management processes such as detection and prevention solutions to assure the customers that the data in the Cloud is appropriately protected.
- What security controls must the customer provide over and above the controls inherent in the Cloud-based platform.
- How must an enterprise's security management tools and processes adapt to manage security in the Cloud [15].

Data at Rest (DAR) is a major concern when utilizing a Cloud service. The customer's data must be stored in an agreed upon location in order to define jurisdictional boundaries. Jurisdictional boundaries determine data access legalities and usage rights. Compliance to regulations and laws in different geographic regions can become challenging to deal with. Proper legal advice is critical to ensure that the contract specifies the responsibilities and liabilities of both the CSP and the Cloud customer. While there is little legal precedent regarding liability in the Cloud at this time, NIST has developed three definitions to improve the understanding of the complex legal concept of liability. There are two types of liability, direct and indirect, both of which have limitations associated with them.

Liability (Direct): Liability for damage caused to the customer by the provider. In this context, "direct liability" is taken to mean liability for losses to the customer relating to the loss or compromise of data hosted on the Cloud service.

Liability (Indirect): A legal obligation resulting from damages awarded to an injured party because of the negligent act of a third party. In the context of Cloud computing, typically used to cover indirect, consequential or economic losses arising from a breach by the provider.

Liability Limits: Terms seeking to limit the extent of any damages that the provider is held liable for [12].

To cover the unique challenges of implementing a Cloud-based system, a CSLA should contain a set of requirements that will be established by the Cloud consumer that must be followed by the Cloud provider. Examples include logging, auditing, licensing, and information management requirements. Information management requirements can be complex, the location where data can be stored, privacy issues, data preservation, and what will be done if data is seized all have to be considered.

D. CLOUD SLA METRICS

There are two major categories of Cloud metrics: business metrics and technical metrics that enable the business SLA to be met. When considering metrics in a Cloud SLA, it is recommended by NIST that the consumers and providers:

- Understand the business objectives for the Cloud opportunity.
- Understand the context and where the stakeholders fit into the Cloud ecosystem.
- Understand potential cascading SLAs and associated metrics.
- Understand enabling "technical metrics" vs. more visible "business metrics".
- Identify the set of metrics that align with prioritized objectives.
- Understand the usage cost models that are applied.
- Clarify how the metrics will be used and what decisions will be made.
- Ensure these metrics are defined at the right level of granularity and can be monitored on a continuous basis.
- Determine available standards that help provide a consistent measurement method. (some will evolve as Cloud computing matures)
- Understand the value and limitations of the metrics collected.
- Analyze and leverage the metrics on an ongoing basis as a tool for influencing business decisions [12].

The metrics used are dependent on the type of service model being used such as IaaS, PaaS, and SaaS, as well as the types of services provided. Using the correct metrics when monitoring Cloud computing services is a crucial ingredient for successful monitoring and enforcement of Cloud SLAs.

E. EXAMPLE SLD FOR A CLOUD-BASED SYSTEM

An example SLD (Figure 3) that applied to a traditional client-server system was discussed in this thesis. A subset of seven performance categories were selected from a

group of performance categories typically used in traditional client-server systems. Not all of these service-performance categories would apply to a Cloud-based service provider. Some of the test cases could be similar; however, there are some concerns with the Cloud that do not apply to the traditional client-server system and vice versa. For example, some performance measures may not apply to a Cloud-based system, while others must be measured using different techniques and metrics. Table 6 is a short list depicting some of the differences in measuring performance.

Table 6. Comparison of Traditional Client-Server System and Cloud-based System Performance Category Concerns

(continued on following page)

Performance Category	Traditional Client-Server System Concerns	Cloud-based System Concerns
Help Desk	- Average speed to answer phone calls - Average speed of response – Voice Mail - Average speed of response – E-mail - Call abandonment rate - First call resolution	This service does not apply to a Cloud-based application.
Email Services	- User E-mail availability (access and storage of email messages) - E-mail server end-to-end performance - E-mail server availability - E-mail client responsiveness	- Applications must be complete and available on demand to the customer - Must provide secure E-mail archiving * Encrypt and protect integrity of data in transit * Privacy
Web and Portal Services	End-to-end performance - PKI Services - Domain Name Services (DNS) support	Applications must be complete and available on demand to the customer. Traditional licensing and asset management may be different. - Reliability and redundancy of Internet connectivity used by the customer and the CSP - Security issues with a public environment - Availability issues of Internet connectivity

Performance Category	Traditional Client-Server System Concerns	Cloud-based System Concerns
File Share Services	- File Share Server availability - Client responsiveness	- Data may not be immediately located in the event of a disaster. Recovery time objectives should be stated in the contract. - The shift from in-house processing/storage of data to a system where data travels over the Internet to and from one or more externally located and managed data centers raises significant issues concerning: * Ownership of data * Disposition of data * Data breaches * Location of data * Legal/government requests for access to data * Data leakage * Data privacy
Print Services	- Print Server availability - When a Cloud-based service provider is being used, the client-server system is responsible for sending the print job to the appropriate printer, with the particular options the user selected, and providing job status to the application.	- Print Server availability - Printer must be Cloud-ready - Data in transit or at rest must be protected/encrypted
Network PKI Logon Services	Client responsiveness	PKI use for single sign-on and file/message encryption: - PKI deployments face challenges due to VM snap-shotting and insufficient entropy - PKI duplication issues may exist - Certificate Authority separation (separation between CAs and customers) – a customer should only be able to see and use its own CAs
RAS Service	- RAS Service availability - Client responsiveness	- Shared responsibility in security management - RAS Service availability

Table 6 (continued)

The SLD (Figure 3) in Section III.B would not be applicable to a cloud-based service provider. This SLD measures Network PKI ***Logon*** Services which would be the responsibility of the customer's network, this should be in the user's site SLA. PKI is just as important in Cloud-based environments as it is in traditional client-server systems to ensure secure communication. However, the Cloud-based service provider has a different role in PKI management. The Cloud-based service provider is responsible for providing end-to-end encryption and credential management allowing users to send and receive encrypted data as well as working with encrypted data to defend against denial-of-service (DoS) attacks, data theft, and unauthorized use. Testing PKI services in the Cloud would require a different approach as shown in Figure 7 and 8. Figures 7 and 8 show the SLDs for measuring the availability and notification portion of the Cloud-based PKI service.

<table>
<tr><td colspan="2">SERVICE NAME: CLOUD-BASED System - PKI SERVICES</td><td>SLA: 100</td></tr>
<tr><td colspan="3">Service Description: Public Key Infrastructure (PKI) is a system of digital certificates, Certificate Authorities, and other registration authorities that verify and authenticate the validity of each party involved in an Internet transaction. PKI is important in Cloud-based environments to ensure secure machine to machine communication. PKI provides authentication, non-repudiation, integrity, and confidentiality of the messages exchanged. It also, provides secure Web Services across domain boundaries and secure data storage.</td></tr>
<tr><td colspan="2">Performance Category: Cloud-based System PKI Services - Availability</td><td>Increment 1
SLA: 100.2</td></tr>
<tr><td colspan="3">Performance Category Description: PKI Services are available from the Cloud provider for end-user access to the Enterprise Validation Authority (EVA) Server and Active Directory in support of end-user Public Key Infrastructure (PKI) for file/message encryption and access to secure websites.

This SLA excludes PKI use for single sign-on/logon to the network, this is the responsibility of the customer's client-server system.</td></tr>
<tr><td colspan="3">Measurement CONOPS: The PKI Service will be available 99.95% of the time during the month being measured.</td></tr>
<tr><td>Who: Cloud Service Provider</td><td colspan="2">Frequency: Monthly</td></tr>
</table>

<table>
<tr><td>User Population:
All Users.

Sample Size: N/A

Sample Unit:
N/A

Where Measured:
Measured at all sites based on actual availability.</td><td colspan="3">Penalty for SLA Violation:
A credit equal to one day's charge for the monthly recurring fee will be issued to the End User.

Frequency of Measure:
Ongoing

Weighting (as applicable):
Equal weighting</td></tr>
<tr><td>Limitations:</td><td colspan="3">Unavailability during scheduled maintenance will not be a violation as long as the maintenance time does not exceed industry standards for the type of maintenance being performed.</td></tr>
<tr><td>SLA Success Criteria</td><td colspan="3">All end user sites must have access to PKI Services at least 99.99% of the time.</td></tr>
<tr><td rowspan="2">SLA Target</td><td rowspan="2">Availability</td><td>Time Interval</td><td>Percentage Complete</td></tr>
<tr><td>One Month</td><td>(goal >= 99.99%)</td></tr>
</table>

Figure 7. SLD Example (Availability SLA-100.2)

<table>
<tr><td>SERVICE NAME: CLOUD-BASED System - PKI SERVICES</td><td>SLA: 100</td></tr>
<tr><td colspan="2">Service Description: Public Key Infrastructure (PKI) is a system of digital certificates, Certificate Authorities, and other registration authorities that verify and authenticate the validity of each party involved in an Internet transaction. PKI is important in Cloud-based environments to ensure secure machine to machine communication. PKI provides authentication, non-repudiation, integrity, and confidentiality of the messages exchanged. It also, provides secure Web Services across domain boundaries and secure data storage.</td></tr>
<tr><td>Performance Category: Cloud-based System PKI Services - Notification</td><td>Increment 1 SLA: 100.3</td></tr>
<tr><td colspan="2">Performance Category Description: Cloud Service Provider will contact end users within 15 minutes of determining PKI Service is or will not be available.
This SLA excludes PKI use for single sign-on/logon to the network, this is the responsibility of the customer's client-server system.</td></tr>
<tr><td colspan="2">Measurement CONOPS: The End User will be contacted within 15 minutes of a PKI Service outage.</td></tr>
</table>

<table>
<tr><td>Who: Cloud Service Provider Security Operations Center</td><td colspan="3">Frequency: Monthly</td></tr>
<tr><td>Where: Cloud Provider Site

User Population:
All Users.

Sample Size: N/A

Sample Unit:
N/A

Where Measured:
Measured at all sites based on actual availability.</td><td colspan="3">Penalty for SLA Violation:
A credit equal to one day’s charge for the monthly recurring fee will be issued to the End User.

Frequency of Measure:
Throughout the entire month.

Weighting (as applicable):
Equal weighting</td></tr>
<tr><td>Limitations:</td><td colspan="3">End User contact information should be kept updated by the End User. If contact information isn’t kept updated, this SLA does not apply.
Only one credit will be given for any one single violation of this SLA..</td></tr>
<tr><td>SLA Success Criteria</td><td colspan="3">All targets must be met to pass the SLA.</td></tr>
<tr><td rowspan="2">SLA Target</td><td rowspan="2">Notification</td><td>Time Interval</td><td>Percentage Complete</td></tr>
<tr><td>(goal <= 15 min)</td><td>N/A</td></tr>
</table>

Figure 8. SLD Example (Notification SLA-100.3)

In Increment 1 SLA 100–2 and 100–3 we measure only the portion of PKI transactions that are actually performed in the Cloud. The user would have to first use their PKI certificate to log onto the client network. Then to be able to access and use encryption services provided in the Cloud the client would need to interact with the Cloud’s PKI management services. The Cloud could provide access to secure websites requiring PKI use as well as store/create/access encrypted data files using applications available through the Cloud. To correctly create an SLA for a Cloud-based server, only the transactions actually performed in the Cloud can be measured. The Cloud service provider cannot be scored on issues the client’s network is responsible for. Activity on

the customer-side (client network) still needs to be monitored to be able to trace the location of the problem, which actually could be some combination of provider service and customer-side issues.

F. CLOUD-BASED SYSTEM SLA–ENFORCING AND MONITORING

A Cloud SLA (CSLA) is a service-based agreement enforced and monitored by gathering data measuring the end-user experience while consuming resources provided in the Cloud. Recall that a SLA describes an agreement on non-functional requirements between provider and customer. An SLA consists of service level objectives (SLOs) that are evaluated according to measurable Key Performance Indicators (KPIs). Automatic SLA protection enables further increase of the system utilization and system profit. In currently available systems only some basic SLAs like "uptime over a time period guarantee" are available. As a result of the dynamic features of a Cloud-based system, continuous monitoring of QoS attributes is necessary to enforce the SLA. Both the CSP and the end user must be able to monitor and assess the services being provided.

Other factors also affect CSLA enforcement. For example the complexity of a Cloud-based system necessitates the use of elaborate or automated methods to manage the CSLA such as through the use of WSLA which was discussed in Section III B. Trust also affects CSLA enforcement, especially when the customer outsources its critical data. The customer also may not trust the information contained in the monthly SLA reports if the monitoring and reporting is all done by the provider. Employing a third party to enforce and monitor the CSLA can address this issue. The third party would be responsible for the measurement of the QoS parameters as well as reporting violations of the CSLA by the provider or the end user. Tools such as WSLA even offer a third-party support option to efficiently monitor and enforce the CSLA.

Transitioning to the Cloud will bring with it new concerns and responsibilities. When considering the move to Cloud computing, organizations should weigh the cost savings with the additional risks incurred. The utilization of computing resources and the approach to creating, enforcing, and monitoring the SLA will change. Both the Cloud provider as well as the Cloud consumer must take on new responsibilities. Some of the Cloud consumer responsibilities include performing provider failure planning, adhering

to established acceptable use policy, and training the Cloud users. The Cloud consumer must also establish Cloud provider requirements to adhere to a given set of standards, logging requirements, licensing requirements, as well as information on how audits will be conducted and how the consumer's information will be managed. The Cloud provider is responsible for things like the physical infrastructure, applications, middleware and the hosting and transmitting the Cloud consumer's data.

When considering the three core goals of information security Confidentiality, Integrity and Availability or CIA as they are commonly referred to, responsibilities can differ not only by stakeholder but also based on the Cloud-service model in use. Figure 9 compares Cloud-security responsibilities for Cloud providers and Cloud consumers based on three different types of service models: IAAS, PAAS and SAAS. In this example Confidentiality refers to limiting data access to authorized users. Integrity refers to preventing data from being changed inappropriately as well as source and origin integrity. Availability refers to the availability of data and the services provided in the Cloud. Figure 9 shows the wide variance in responsibilities in these areas for both the provider and the consumer.

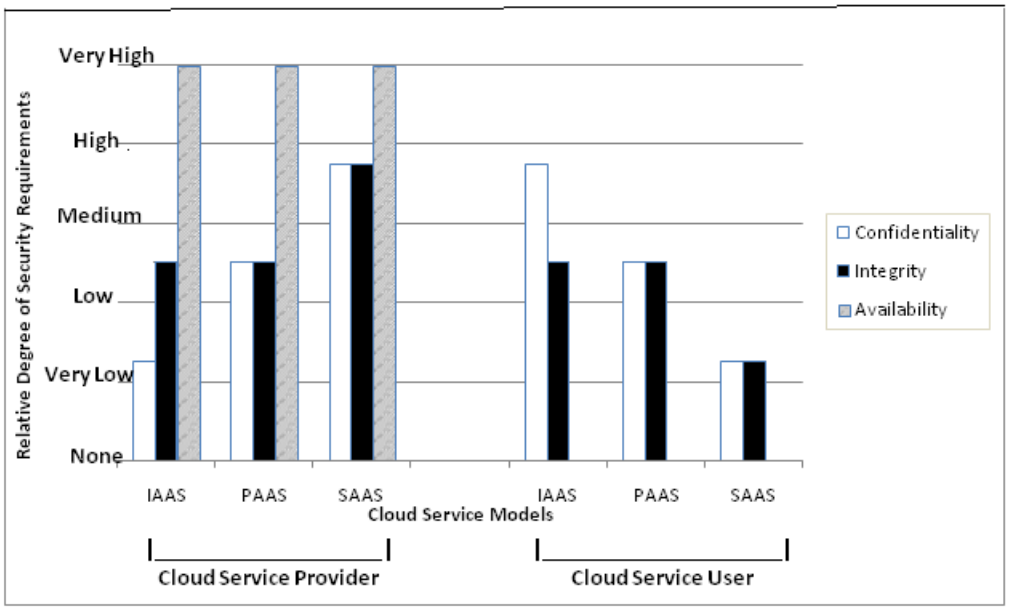

Figure 9. Cloud Security Responsibilities for Providers and Users. From [8]

Cloud providers are generally responsible for providing physical security, certification, and maintenance of Cloud resources as well as supporting portability, interoperability, and planning for redundancy. The Cloud provider must also ensure the SLA contains a limitations section to cover areas where they denote the limitations of the services offered and their liability. These limitations may include the following:

Force Majeure: Clauses which excuse a party from liability if some unforeseen event beyond the control of that party prevents it from performing its obligations under the contract.

SLA Changes: Terms which specify how and when the provider may change the terms of the SLA and whether or not there is any obligation to notify the customers affected.

Security: It is incumbent upon Cloud consumers to read carefully the de facto security included as a part of the service offering. Providers frequently disclaim any significant responsibility for security of user information created, transported, or processed within their services.

Scheduled Outages: Also referenced as Planned Downtime, Scheduled Maintenance, et al. These are service disruptions initiated by the provider to undertake system maintenance and/or upgrades. This type of outage is typically excluded from remedies offered to specified unscheduled or unplanned disruptions [12].

There are many complications to deal with in CSLA management. One of the most common problems with Cloud SLAs is they overlook network performance. In the Cloud, services are most often utilized over the Internet. It is difficult to guarantee Internet availability as it is a best-effort service. Cloud services may also be accessed through another company's network connection making it impossible for the CSP to

guarantee network performance. “It is difficult to justify negotiating a cloud SLA when you can’t guarantee the connection; it's also hard to prove that a cloud provider failed to meet an SLA when there's a component of the service–the network–between your QoE measurement point and the cloud. This particular issue also affects management connections to the cloud and the ability to write an SLA on management-level QoE” [16].

Another problem is that sometimes it is not easy to find the root cause of an SLA failure or violation. This is mainly due to the complexity of Cloud-based services. Amazon Elastic Compute Cloud (EC2), one of the largest Cloud providers, as well as Google App Engine and Rackspace, require their consumers to prove SLA violations have occurred by submitting a claim when they experience a network outage. Many Cloud providers agree that the customer should be responsible for both detecting and notifying of an SLA violation. This is one of the biggest problems pertaining to the current status of Cloud SLAs; it burdens the customer with not only the loss of productivity during an SLA violation but also with the tasks of documenting the violation and notifying the provider. This puts the customers at a disadvantage if they do not fully understand the complexity of the system enough to monitor it correctly. An alternative would be for the Cloud provider to take full responsibility for monitoring the system as well as automatically paying or offering a credit when the customers experience an outage or other non-compliance issue. A possibly more effective approach would be have both the customer and the provider play a role in the monitoring of the system. This would provide a check-and-balance situation and also relieve the customer of having to assume full responsibility for monitoring compliance with the SLA.

G. LESSONS LEARNED FROM AMAZON EC2 BLACKOUTS

Outages are always a problem with any type of system, especially in Cloud-based systems. For example, Amazon EC2 is divided into regions or data centers. Each of the regions are then divided into several availability zones (AZs) as shown in Figure 10. Elastic Block Store (EBS) which are network attached storage devices are used in the AZs. Amazon also uses the Relational Database Service (RDS) to permit the use of databases on EC2 that are backed by EBS. Their SLA guarantees 99.95%

uptime/availability within a region over a 365 day period; this is approximately 4.3 hours of nonscheduled downtime per year.

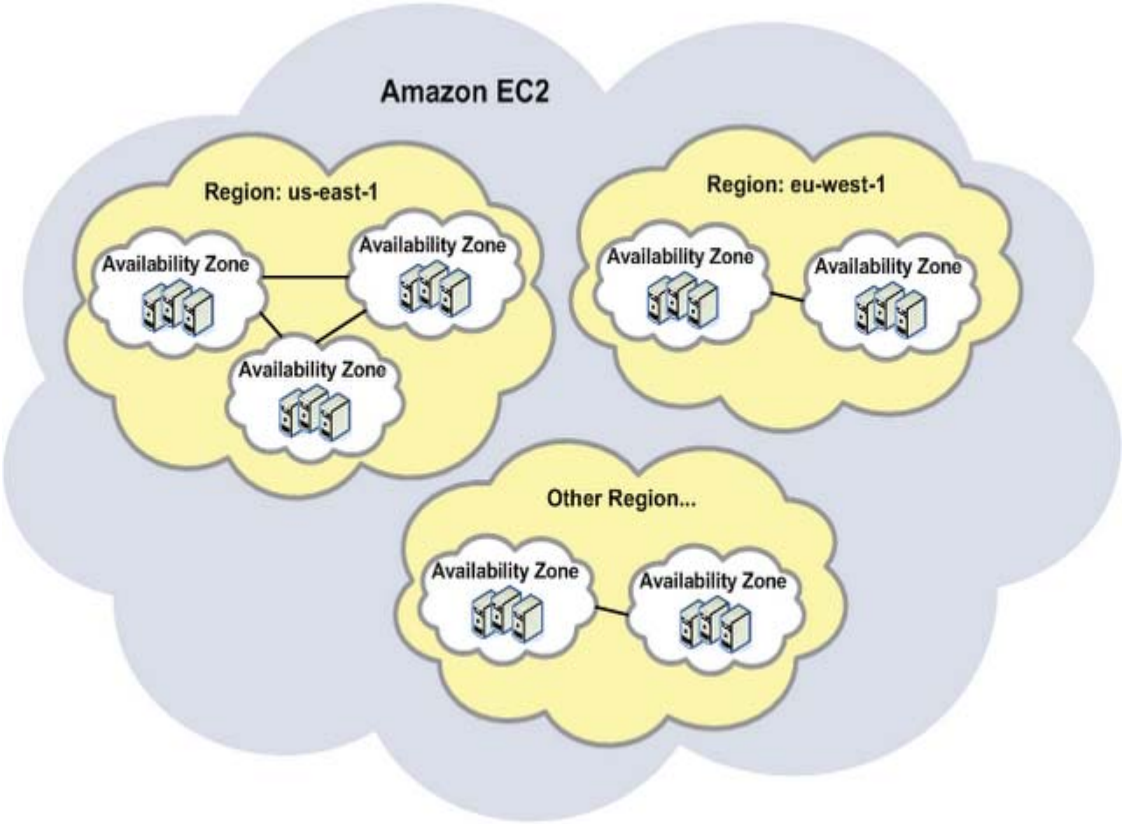

Figure 10. Availability Zone Concept. From [66]

In April 2011 and June 2012, Amazon EC2 experienced major service disruptions. The April 2011 incident occurred during a network configuration change to upgrade the capacity of the primary network in the US East-1 region. During this time traffic on the primary network was supposed to be shifted through one of the redundant routers to another network of the same capacity. This procedure did not execute correctly and traffic was routed through the wrong router onto a lower capacity network that could not handle the traffic level it was receiving. As a result, re-mirroring of a large number of EBS volumes was impacted. The RDS was also affected as it uses the same storage infrastructure. There was a simultaneous disconnecting of both the primary and secondary network that left the affected nodes completely isolated from each other.

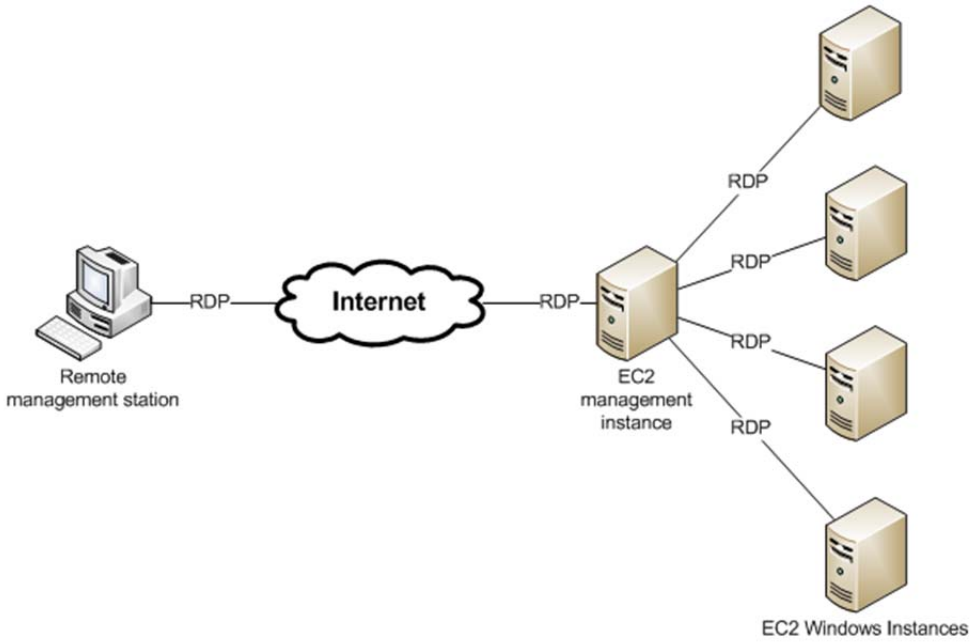

Figure 11. Preferred Amazon EC2 Management Flow. From [17]

The outage lasted for nearly four days, more than long enough to qualify for a service credit. However, Amazon stated that the availability guarantee applies only to the connectivity of EC2 instance and in this case it was actually the EBS and RDS services that failed. Amazon went ahead and gave a ten-day service credit to affected customers, equal to 100% of their usage of EBS Volumes, EC2 Instances and RDS database instances they were running in the affected Availability Zone.

The June 2012 outage was triggered by a large electrical storm in Northern Virginia. Backup systems had to respond due to power fluctuations. No loss of power was experienced until the electrical switching equipment in two of eastern datacenters supporting a single Availability Zone were affected by a large voltage spike. Then the equipment switched to generator power, except in one of the datacenters where the generators did not respond correctly and the servers ended up running on the Uninterruptable Power Supply (UPS) units which were eventually depleted causing the servers to lose power at 8:04 p.m. Pacific Daylight Time (PDT). It took about ten minutes for onsite personnel to stabilize the primary and backup power generators and get power restored. It was not until around 2:45am PDT that the affected data center was 90%

recovered due to the backlog of volumes to be processed. Once again Amazon stated that this was not an outage of the EC2, instead it was an outage of the Elastic Load Balancing (ELB) and the RDS caused by an unexpected bug discovered as the power was restored and the network began to come back up.

> "The bug caused the ELB control plane to attempt to scale these ELBs to larger ELB instance sizes. This resulted in a sudden flood of requests which began to backlog the control plane. At the same time, customers began launching new EC2 instances to replace capacity lost in the impacted Availability Zone, requesting the instances be added to existing load balancers in the other zones. These requests further increased the ELB control plane backlog. Because the ELB control plane currently manages requests for the US East-1 Region through a shared queue, it fell increasingly behind in processing these requests; and pretty soon, these requests started taking a very long time to complete" [18].

In a statement to their customers, Amazon apologized for the inconvenience this caused for some of its customers. Once again, Amazon stated that it had learned from the event and would be making every effort to improve its services. No mention of a service credit was made this time.

The primary lesson to be learned by both the customer and the provider from any outage is that the Cloud is not infallible and 100% unlikely. However, following a few rules can improve the Cloud experience. Make sure to read the SLA thoroughly and understand what it covers. The CSP as well as the customer should have a contingency plan in place to help alleviate some of the inconvenience of an outage. Also, keep in mind that every transaction is subject to failure. And, as is important in any computer system, always backup data to other locations to avoid a partial or total loss in the event of an outage.

THIS PAGE INTENTIONALLY LEFT BLANK

VI. CONCLUSIONS

Developing an effective SLA is a fundamental element to successful use of Cloud computing and traditional client-server systems. Some of the same principles apply to both types of SLAs; however, a Cloud SLA is more complex. An SLA must be specified according to the computing paradigm it will be applied to, such as a Cloud-based system or a traditional client-server system. The type of paradigm greatly affects the provider and the users' responsibilities, the levels of QoS, and the complexity of SLA creation. An SLA is a living document that must be continually monitored and updated as necessary.

End-to-End Service Level Management in a Cloud-based system is not a trivial simple process. Cloud-based systems are complicated by the varied technologies, networks, and provider services involved. Multiple vendors, domains, and technologies must be supported by SLM. This can be accomplished with the use of machine-readable SLAs that can be interpreted across many different platforms. Machine-readable SLAs are preferable and enable runtime interpretation of parameters that can then be used to determine the most efficient way to allocate the resources provided by the Cloud-based system. Standardization for defining and negotiating SLAs can also simplify the management of Service Levels across different networks and providers.

With both types of systems, customer perception and satisfaction are actually the ultimate measure of service-level performance. This can be difficult to measure because sometimes it takes more than meeting the SLA requirements to satisfy the customer. When first implementing the SLA, reports are instrumental in adjusting service levels to meet the requirements required to satisfy the customer. After the requirements are met, the reports may not be a good indication of the end-users' satisfaction as their needs or expectations may have changed or were misunderstood by the provider in the first place. The reports, however, can be used to improve delivery service based on the users' higher-level requirements once there is an understanding of these requirements. To do this there must be good communication between the provider and the users.

A. ISSUES AND LESSONS LEARNED

An effective SLA is a means for improving communications, managing expectations, and clarifying responsibilities. However, an SLA executed incorrectly or fails is of little value and will not benefit anyone in any way.

There are many reasons why an SLA may fail. The two key areas of failure are alignment and integrity. Alignment is critical on many levels and includes: misalignment of service performance to business strategy, alignment of commitments from your vendors to your commitments to your customers, and alignment of service component to business process." [19]

Integrity issues are also a concern. These are encountered when the SLA has not been clearly defined, SLA metrics are not captured correctly, or the provider takes advantage of unintentional loopholes in the SLA. These issues can make it appear as if the SLAs have been met when in reality they have not. A provider doing its own auditing may not be honest and provide information that could be used to penalize itself. One way to alleviate these problems is to use a third-party auditing provider.

Other issues that can lead to SLA failure include: lack of upper management support, misinterpreted information, SLA obligations not met, and an SLA that is not proactive enough.

1. Lack of Upper Management Support

Upper management has a key role in driving a project's progress. Permission and support from upper management are required to succeed. Without management it is difficult if not impossible to obtain the necessary personnel and financial resources to be successful.

2. Misinterpreted Information

Most SLAs are written using technical terms that are not always fully understood by all of the stakeholders. To avoid this problem ensure service level definitions are business based and meaningful to the users. SLDs should also be easily defined and measurable. On many occasions the example SLA discussed in this thesis brought about

disputes due to misinterpreted requirements mainly during the review of the monthly SLA reports required by the example company's contract. The disputes were settled during resolution meetings with a third party mediating.

3. SLA Not Proactive Enough

Customers most often want the provider to use a proactive approach to help them identify their business needs and proactively meet them. This requires a broader sharing of information that might not always be possible. Technical staff must also have the ability to elicit requirements from the customers. This can be difficult as the two groups may not be able to communicate on the same level [20]. It was found during SLA reviews for the example customer used in this thesis that by only penalizing the provider for performance below the minimum allowed standard, the providers had no incentive to strive for better than the minimum levels.

B. BENEFITS

Many benefits can be realized when incorporating an effective SLA:

- Improved communication between all stakeholders
- Improved software qualities by incorporating quality factors into the development effort
- Conflict prevention tool
- Documents service levels
- Provides standardized methods for communication of expectations
- Can be used to gauge service effectiveness
- Identifies areas that are working well and those that are not
- A clearer understanding of responsibilities by all
- Establishes a two way accountability for a service
- Provides a basis for improving service levels
- A living document which allows for changing as needed
- Ensures all parties are using the same criteria to evaluate service quality
- Provides requirements should something go wrong
- Creates standardized levels of service that are negotiated
- A type of insurance policy for the customer used to communicate requirements and obtain compensation when agreed upon service levels are not met

These benefits will only be realized if an effective Service Level Management process is followed. This requires ongoing communication between the provider and the customer. Regularly scheduled meetings to discuss SLA measurement results make it possible to discover areas that need worked on or improved and may lead to changes being made to the SLA. In this way SLA management and monitoring are continuous processes. Identification of customer needs, design, and implementation of a service process, and improving the service must continue throughout the lifecycle of the SLA.

Any SLA management strategy considers two well-differentiated phases: the negotiation of the contract and the monitoring of its fulfillment in real-time. Thus, SLA Management encompasses the SLA contract definition: basic schema with the QoS (quality of service) parameters; SLA negotiation; SLA monitoring; and SLA enforcement—according to defined policies [21].

Developing an effective SLA is a challenge. It involves a lot of time, research, and effort to get it right. But if done correctly the benefits can be seen in improved communication between the customer and the provider as well as improved service delivery. This can be accomplished by accurately allocating responsibilities and the associated risks among the parties involved, as well as by clearly defining specifications and techniques for verifying performance. These core elements will be used whether the SLA is for a traditional client-server system or for a Cloud-based system.

Communication and clear expectations of the customer and the service provider are the most important elements in making an SLA effective. An effective SLA not only documents what is important to the customer but also realistic expectations of the services provided by the service provider.

C. FUTURE WORK

Current Cloud SLA processes are not mature enough to adequately manage Cloud services. Therefore, we are challenged with the task of creating more complex SLAs which can effectively manage and monitor Cloud services. This will involve further research of several key issues including standardization, management and monitoring automation, and SLA formal specification and validation.

1. Standardization

Currently there is a lack of standardized templates, best practices and policies for creating and maintaining Cloud SLAs. A disadvantage this causes is it complicates the situation when a customer shops for a CSP; it is difficult to compare CSPs when they use different standards to measure compliance. While some work has been done towards developing Cloud standards, more work needs to be done and more businesses need to adopt the standards.

2. Management and monitoring automation

Today, the majority of SLA management is mostly performed manually; this is time consuming, expensive and prone to human error. Increasing automation of SLA management and monitoring for large enterprises will greatly enhance the success of Cloud SLAs. As enterprises become larger and more complex, meeting SLA requirements becomes more difficult. Large service providers need to manage thousands of SLAs for many different customers requiring many different services. In addition, business requirements and operational environments are constantly changing making it ineffective to use static QoS requirements and metrics in SLAs. Meeting or exceeding SLA requirements to avoid losses and penalties is also a major concern for service providers. For these reasons, automation and dynamic resource allocation at runtime based on QoS parameters are needed. Machine-readable SLAs are more effective than human-readable SLAs in providing these capabilities. It is also highly desirable to use machine-readable SLAs because only a few or no humans are needed in the loop, this is more efficient as it lessens or totally removes the chance of human errors. Improvements in machine-readable SLAs would also benefit by providing more efficient automatic SLA violation detection at runtime. This would remove the burden of placing the violation proof on the customer.

Current systems provide no integrated support for SLA service QoS specification and translation of QoS to configuration of low-level mechanisms for delivering the expected QoS. New approaches like those proposed in [22] and [23] are needed. In [22] Correia and Abreu were concerned with improving SLA specification, definition and compliance verification. They proposed a model-based approach to SLA specification and compliance verification using SLA specifications derived from domain specific

languages (DSL). The goal was "to implement a model where the dispatching of events will result from the conjunction of rules, namely settled in SLA contracts."

In [23]Freitas, Parlavantzas, and Pazat proposed an integrated SLA description, translation, and enforcement concept using WSLA, WS-Agreement, and Quality Assurance for Distributed Services (Qu4DS) framework (see Figure 12). WSLA and WS-Agreement were discussed earlier in this thesis. They presented a research prototype called the Qu4DS framework as a proof of concept tool to support the development and management of Cloud services. Qu4DS provides automatic SLA management functions such as service negotiation, instantiation, SLA translation, and QoS assurance.

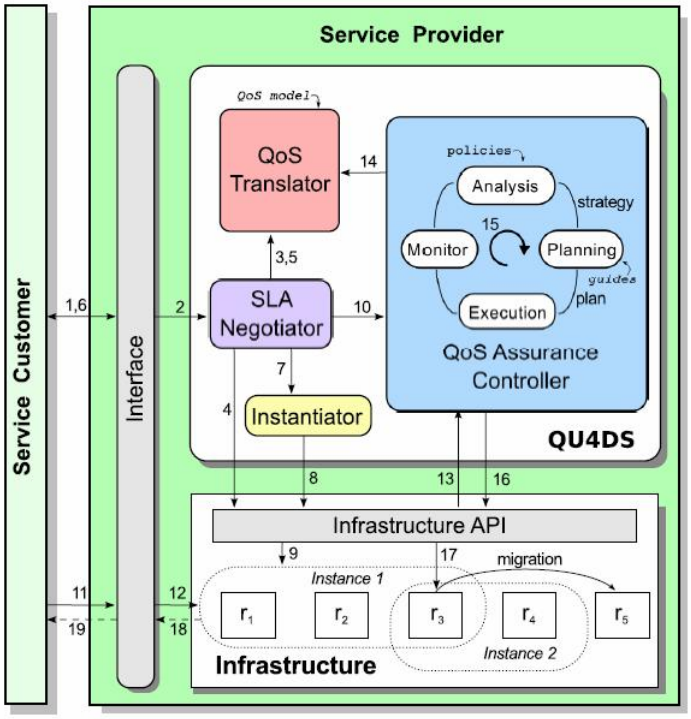

Figure 12 The Qu4DS Framework. From [24]

3. SLA formal specification and validation

Current SLA specification languages used for creating machine-readable SLAs are not robust enough to handle the complex needs of a Cloud SLA. Specification languages, like SLALOM, SLAng, RBSLA, and WSLA, are based on XML; limiting their ability to match composition metrics to syntactical [25]. Research of SLA

specification languages to produce more effective machine-readable SLAs is ongoing. In [26] Paschke and Bichler, they proposed using a logical framework combining different logical formalisms like Horn Logic, Event Calculus, Deontic Logic, and Event and ECA rules to address the need for automatic SLA management. Their goal is to construct a logical framework for specifying complex business rules and policies, detecting contract violations, authorization control, and conflict detection. "The particular advantages of their logical approach in contrast to traditional procedural programming approaches is its high flexibility, its dynamic extendibility and its high potential for the automation of contract enforcement processes such as the detection of contract violations, authorization control, conflict detection, service billing and reporting" [26].

Tools to translate machine-readable SLAs into more readable formats are also being developed. This is beneficial because the translated SLAs can be included in the actual service contract. On the other hand, processes are also needed for the manual translation of natural language into machine-readable specifications to make SLA development less difficult, less time consuming and less expensive, especially for very large enterprises.

With additional research, machine-readable SLAs will be able to provide for more flexible automatic management, execution, and maintenance of SLAs for larger complex systems. Researchers continue studying ways to make a specification language with a more powerful matching capability and to improve machine-readable SLAs to allow for more complex automatic, dynamic resource allocation at runtime. Advancements in these areas will not only increase the service provider's ability to meet or exceed SLA expectations but will also improve the customer's experience. While research has shown that formal specifications and methods help improve the clarity and precision of requirements specifications (for example, see the work of Steve Easterbrook and his colleagues [27]), formal specifications are useful only if they match the true intent of the customer's requirements. We need more effective means to validate the correctness of the formal SLA specifications.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX

TRADITIONAL CLIENT/SERVER SLDS

SERVICE NAME: HELP DESK		**SLA: 101**
Service Description: Provides technical assistance to users of the network. Assistance can be provided through email or telephone calls.		
Performance Category: Average Speed of Answer - Telephone Calls		**Increment 1** **SLA: 101.1**
Performance Category Description: The Average Speed to Answer (ASA) is the monthly average of the amount of time that a caller will wait, after choosing the last voice menu prompt, before a live agent answers. A user will be offered the option to leave a voice mail or continue to wait for a live agent. If a user chooses to leave a voicemail, the amount of time calculated will be the time between choosing the last prompt on the initial voice menu and the time that the customer selects the voicemail option.		
Measurement CONOPS: This SLA is the measure of time, following a call to the Help Desk, between the selection of the last prompt on the Automated Call Distribution (ACD) system [call is considered at this point to be in queue] and the call being answered by a help desk agent, or the user choosing to leave a voicemail. Abandoned calls, either prior to listening to the phone prompts or after selected the last phone prompt will be excluded from this calculation. Calls will not be segmented by seat type nor by prime time vs. non-prime time. On a monthly basis, the summary field in the ACD system that is titled ANSTIME will be added to the summary field in the ACD system that is titled OUTFLOWTIME. These two fields represent the total number of seconds associated with calls waiting in the queue and the queue time for calls that go to voicemail, respectively. The sum of these two figures is divided by the total calls offered to the queue minus any abandoned calls. This value is the average speed to answer.		
Who: Contractor	**Frequency**: Monthly	
Where: Client Site **User Population:** All Network Users **Sample Size:** All calls **Sample Unit:** End User calls to Help Desk **Where Measured:** Help Desk automated call distribution system	**How Measured (i.e., captured):** End user calls to the Help Desk **Measurement Formula:** Total number of seconds from the last voice menu prompt until a live agent answers for all answered calls to Help Desk / Total number of calls answered by Help Desk **Frequency of Measure:** Continuous **Weighting (as applicable):** Equal weighting for all calls	
Aggregation of Data:	Sites will be aggregated at the user population level.	
SLA Success Criteria	All targets must be met, to pass the SLA	
SLA Target	**Average Speed to Answer**	<= 40.0 seconds

Performance Category: Average Speed of Response – Voice Mail/E-mail		**Increment 1 SLA: 101.2**
Performance Category Description: If a user elects to leave a voice mail or e-mail message with the help desk instead of speaking with a live agent, the help desk will contact the user regarding the voice mail or e-mail. The user must provide in the voice mail or e-mail accurate contact information (i.e., name and phone number). The Contractor has notified the End User that the required technology is not currently available. Upon availability of the technology, the End User the Contractor shall, within six months develop the measurement CONOPs and SLA targets to implement this SLA.		
Measurement CONOPS: The receipt time/date stamp of the voice mail or e-mail will be the start time; the creation of the trouble ticket with e-mail or voice reply will end the SLA measurement.		
Who: Contractor	**Frequency**: Monthly	
Where: Client Site **User Population:** All End Users **Sample Size:** All calls and emails **Sample Unit:** End user calls and emails to Help Desk **Where Measured:** Help Desk Trouble Ticket System	**How Measured (i.e., captured):** End user calls and emails sent to the Help Desk **Measurement Formula:** Voice Mail = Total response time (in minutes) of all Voice Mail tickets / Total number of Voice Mail tickets E-mail = Total response time (in hours) of all E-mail tickets / Total number of E-mail tickets **Frequency of Measure:** TBD **Weighting (as applicable):** Equal weighting for all responses. Both performance elements must be separately passed in order to meet the SLA performance.	
Aggregation of Data:	Sites will be aggregated at the user population level.	
SLA Success Criteria	All targets must be met, to pass the SLA	
SLA Target	**Average Speed of Response Voice mail**	(goal 60.00 min) TBD
	Average Speed of Response E-mail	(goal 4.00 hrs) TBD

Performance Category: Call Abandonment Rate	**Increment 1 SLA: 101.3**
Performance Category Description: The Call Abandonment Rate is the percentage of calls that are terminated by the customer following the selection of the last voice menu prompt and prior to a live agent answering the call.	

Measurement CONOPS: Call abandonment rate is measured against phone calls placed by users to the help desk and received by the automated call distribution system (ACD).

For the purposes of this SLA, a user is identified as a caller who selects all of the ACD menu prompts applicable to their problem type. Callers that select the final ACD menu prompt are placed in the ACD queue and offered the opportunity to communicate with the Help Desk. Their calls are then characterized as offered in one of the following three ways:

1. Users in queue may wait for a live Help Desk agent to answer and handle their call.
2. Users in queue may choose to be transferred to Voice Mail that will then handle their call.
3. Users in queue may abandon the call either before a live agent answers or they choose to transfer to Voice Mail. These calls are not handled.

Callers that hang up before the final ACD menu prompt are not included in the abandonment calculation.

Who: Contractor	**Frequency:** Monthly	
Where: Client Site **User Population:** All End Users **Sample Size:** All calls **Sample Unit:** End user calls to the Help Desk **Where Measured:** Automated Call Distribution System	**How Measured (i.e., captured):** End user calls to the Help Desk **Measurement Formula:** Number of calls abandoned / Offered calls **Frequency of Measure:** Continuous **Weighting (as applicable):** Equal weighting for all abandoned calls	
Aggregation of Data:	Sites will be aggregated at the user population level.	
SLA Success Criteria	All targets must be met, to pass the SLA	
SLA Target	**Call Abandonment Rate**	<= 5.00%
Performance Category: First Call Resolution		**Increment 1** **SLA: 101.4**

Performance Category Description: The percentage of answered calls to the help desk that are resolved on the initial call in the following scenarios:

a) Problems and/or issues resolved within 30.0 minutes of the initial call to the Help Desk while the user remains on the phone line.
b) Problems and/or issues resolved within 30.0 minutes of a return call to the customer from a Help Desk agent in response to an e-mail/voicemail.
c) Problems and/or issues resolved within 30.0 minutes of the initial call by the NOC or other Help Desk Subject Matter Expert due to a warm transfer, which results in problem being resolved while the user remains on the phone line.
d) Cases in which the end user is redirected to another support center after determining that responsibility for resolution lies outside of the Contractor.

Measurement CONOPS: This SLA is the percentage of tickets called or emailed into the Help Desk that were resolved on the first contact with the help desk agent. The 30.0 minute time measure is based on the difference between the Create Date/Time of the Remedy (help desk ticketing) system and the Resolved Date/Time of the same ticket.

On a monthly basis, all user-facing tickets, which were closed in the given reporting month will be collected and assessed for SLA reporting purposes. Those tickets that have been resolved by the Help Desk or Network Operation Center (NOC), will be reviewed based on the field titled Total Time Open in Seconds. This field represents all work time associated with the ticket and will have excluded any time the ticket was pending input from the customer. The field First_Call_Resolution is then reviewed. The number of tickets that have "Yes" in the First_Call_Resolution field is divided by the total number of tickets resolved by the Help Desk and NOC.

Who: Contractor	**Frequency:** Monthly
Where: Client Site **User Population:** All Network Users **Sample Size:** All tickets **Sample Unit:** Closed External Incident Ticket **Where Measured:** Help Desk Trouble Ticket System	**How Measured (i.e., captured):** End user incident reports to the Help Desk **Measurement Formula:** For all closed tickets during the reporting period, The number of tickets resolved* on the first call / Closed tickets *"tickets resolved" must meet Description criteria (a-d) above. **Frequency of Measure:** Continuous **Weighting (as applicable):** Equal weighting for all calls
Aggregation of Data:	Sites will be aggregated at the user population level.
SLA Success Criteria	All targets for each LOS must be met to pass the SLA for that LOS.

SLA Target	**Level of Service**	**Percentage Complete**
	LOS 1 & 2	>= 65.00%
	LOS 3	>= 80.00%

Service Name: E-MAIL SERVICES	**SLA: 102**

Service Description: Access and storage of email messages for users of the network.

Performance Category: E-mail Services - User E-mail Availability	**Increment 1** **SLA: 102.1**

Performance Category Description: E-mail is the Contractor provided user service for sending and receiving E-mail and attachments. This SLA applies to a network-connected workstation at an End User site, and the shared network storage assigned to that site. This SLA excludes RAS and web-based activities.

> **E-mail Services - User E-mail Availability:** Percentage of time E-mail service is available at the end-user workstation. User E-mail Availability is measured by synthetic E-mail transaction generated from the end user workstation across the full connection path of the network infrastructure, to include the user LAN, Base Area Network, WAN, and Server Farm connectivity. A small site sampling is used and the measured performance is assumed to be representative of all users at that site.

The transactions conducted for User E-mail Availability are conducted using synthetic scripts that replicate the actions of a standard E-mail. The intent of these measures is to verify that the supporting network, domain name server, directory, security boundaries, E-mail servers, remote procedure calls, and E-mail applications are available and functioning satisfactorily. End-to-End measurement will be a representative sampling of local, regional, and enterprise infrastructure performance.

The Contractor has notified the End User that automated synthetic transactions will not be available until the 1st Quarter calendar year 2014 timeframe incident to the upgrade to Microsoft Server 2008. Until synthetic transaction measurements are available, User Email Availability will rely on existing E-mail Availability and Performance measures defined in Attachment 2B, Transition Service Level Agreements. These improved measures using automated synthetic transactions (vice manual methods) are a priority for the End User. Upon availability, the End User and Contractor shall, within six months, revise the measurement CONOPs and SLA targets.

User E-mail Availability is measured by sampling. For E-mail Availability, all sites with 24 or greater seats will be configured and conduct synthetic transactions from at least two on-site representative point to ensure reportable data from at least one on-site representative point. If the site receives service from multiple servers than each probe will test a different server. Sites with fewer than 24 seats will not be measured unless mutually determined by the End User and the Contractor.

For E-mail Availability the End User will approve the location of the measurement points.

Who: Contractor	**Frequency:** Monthly

<table>
<tr><td>Where: Client Site

User Population:
All Network Users

Sample Size:
• Sites>=24 seats will have at least two on-site representative points
• Sites<24 seats will not be measured unless mutually determined by End User and Contractor

Sample Unit:
Client

Where Measured:
Client (representative of site)</td><td colspan="2">How Measured (i.e., captured):
With an automated tool.

Measurement Formula:

For sites that have not achieved full performance:
Total available minutes derived from the representative point on an end-user workstation at the site/ Total minutes in the month

For sites that have achieved full performance:
Sum of (Total available minutes derived from the representative point on an end-user workstation at the site x number of seats at the site)/ Sum of the seats at all sites

Frequency of Measure:
(goal 5 min) TBD

Weighting (as applicable): Weighted Average (by seat count)</td></tr>
<tr><td>Aggregation of Data:</td><td colspan="2">Performance data for sites that have not yet achieved Full Performance will be aggregated at the site level and the SLA targets will apply at the site level.
Performance data for sites that have achieved Full Performance will be aggregated at the user population level and the SLA targets will apply at the user population level.</td></tr>
<tr><td>SLA Success Criteria</td><td colspan="2">All targets must be met, to pass the SLA</td></tr>
<tr><td>SLA Target</td><td>User E-mail Availability</td><td>(goal >= 99.7%) TBD</td></tr>
</table>

Performance Category: E-mail Services - E-Mail End-to-End (Client-Server-Server-Client Performance	**Increment 1** **SLA: 102.2**

Performance Category Description: E-mail is the Contractor provided user service for sending and receiving E-mail and attachments. This SLA applies to a network-connected workstation at an End User site, and the shared network storage assigned to that site. This SLA excludes RAS and web-based activities.

> **E-Mail End-to-End (Client-Server-Server-Client) Performance:** Percentage of synthetic E-mail and 10K attachment transactions successfully processed and returned in the required time, stated in minutes roundtrip. Transactions are generated at the client, processed by the host server, forwarded to an appropriate destination server, responded to via an auto-reply generated by the destination server, and returned to the client.

The transactions conducted for End-to-End performances are conducted using synthetic scripts that replicate the actions of a standard E-mail. The intent of these measures is to verify that the supporting networks, domain name server, directory, security boundaries, E-mail servers, remote procedure calls, and E-mail applications are available and functioning satisfactorily. End-to-End measurement will be a representative sampling of local, regional, and enterprise infrastructure performance.

The Contractor has notified the End User that automated synthetic transactions will not be available until the 1st Quarter calendar year 2014 timeframe incident to the upgrade to Microsoft Server 2008. Until synthetic transaction measurements are available, End-to-End Performance will rely on existing E-mail Availability and Performance measures defined in Attachment 2B, Transition Service Level Agreements. These improved measures using automated synthetic transactions (vice manual methods) are a priority for the End User. Upon availability, the End User and Contractor shall, within six months, revise the measurement CONOPs and SLA targets to incorporate the defined automated synthetic transactions described above.

For E-Mail End-to-End Performance, the End User will approve the location of the measurement points.

End-to-end performance will utilize the actual value received or (TBD goal 30 minutes) for any failed test without an associated network or server availability outage documented in another SLA measurement.

Measurement CONOPS: TBD

Who: Contractor	**Frequency:** Monthly
Where: Client Site **User Population:** All Network Users **Sample Size:** TBD **Sample Unit:** Client **Where Measured:** Client (representative of site)	**How Measured (i.e., captured):** With an automated tool **Measurement Formula**: Number of attempts successful within the required Time Interval / Total number of attempts **Frequency of Measure:** (goal 5 min) TBD **Weighting (as applicable):** Weighted Average (by seat count)

Aggregation of Data:	Performance data for sites that have not yet achieved Full Performance will be aggregated at the site level and the SLA targets will apply at the site level. Performance data for sites that have achieved Full Performance will be aggregated at the user population level and the SLA targets will apply at the user population level.		
SLA Success Criteria	All targets must be met, to pass the SLA		
SLA Target	**E-Mail End-to-End Performance**	**Time Interval**	**Percentage Complete**
		(goal <= 5.00 min) TBD	(goal >= 95.0%) TBD
		(goal <= 10.00 min) TBD	(goal >= 99.5%) TBD

Performance Category: E-mail Services - E-Mail Server Service Availability	**Increment 1 SLA: 102.3**

Performance Category Description: E-mail is the Contractor provided user service for sending and receiving E-mail and attachments. This SLA applies to a network-connected workstation at a User site, and the shared network storage assigned to that site. This SLA excludes RAS and web-based activities.

E-mail Services - E-Mail Server Service Availability: Percentage of time the Mail Transfer Service at the E-mail server is online, running, and the Mail Queue is processing or available for processing mail. The terms “active” and “processing” are defined to mean that user-generated E-mail is capable of or is being received and delivered. Server Service Availability is measured at every E-mail “service” at the Server Farm. The term “service” indicates that there may be more than one server identified for processing E-mail for a given user, and availability of any one meets the requirement for the associated set of users.

Measurement CONOPS: All E-mail servers are monitored by Tivoli TEC. If there is an outage, a TEC event will be detected and a Remedy ticket will be created with the start time of the event.

On a monthly basis, all E-mail Server Service customer-impacting tickets, which were closed in the given reporting month, will be collected and assessed for SLA reporting purposes. Those tickets that have been categorized with a Category/Type/Item combination that relates to this SLA will be reviewed based on the field titled Total Time Open in Seconds. This field represents all work time associated with the ticket and will have excluded any time the ticket was pending due to input/access needed from customer. The Total Time Open in Seconds fields will be combined and calculation will be performed.

The following will be excluded from measurement:

- Non-active servers (e.g., backup servers in server clusters) do not count if multiple servers provide the service.

Who: Contractor	**Frequency:** Monthly
Where: Client Site **User Population:** All Network Users **Sample Size:** All servers **Sample Unit:** Server **Where Measured:** Server	**How Measured (i.e., captured):** With an automated tool and end user trouble calls to Help Desk **Measurement Formula:** For sites that have not achieved full performance:: Total available minutes of active email servers at the server farm/ Total minutes in the month x total number of email servers at the server farm For sites that have achieved full performance: Total available minutes of active email servers / Total minutes in the month x total number of email servers at the server farm **Frequency of Measure:** Continuous **Weighting (as applicable):** Equal Weighting
Aggregation of Data:	Performance data for sites that have not yet achieved Full Performance will be aggregated at the site level and the SLA targets will apply at the site level. Sites shall inherit the performance level of the email servers that provide the service to them. Performance data for sites that have achieved Full Performance will be aggregated at the user population level and the SLA targets will apply at the user population level.
SLA Success Criteria	All targets must be met, to pass the SLA.
SLA Target	**Server Service Availability** >= 99.70%

Performance Category: E-mail Services - E-mail Client Responsiveness	Increment 1 SLA: 102.4

Performance Category Description: E-mail is the Contractor provided user service for sending and receiving E-mail and attachments. This SLA applies to a network-connected workstation at an End User site, and the shared network storage assigned to that site. This SLA excludes RAS and web-based activities.

E-mail Services - E-Mail Client Responsiveness: Percentage of transactions sent by the users that fall within the response time to successfully open an e-mail with a 10K attachment. This measure provides a host server response time to an end user initiated request and is measured at the user workstation.

The transactions conducted for Client Responsiveness are conducted using synthetic scripts that replicate the actions of a standard E-mail. The intent of these measures is to verify that the supporting networks, domain name server, directory, security boundaries, E-mail servers, remote procedure calls, and E-mail applications are available and functioning satisfactorily. End-to-End measurement will be a representative sampling of local, regional, and enterprise infrastructure performance.

The Contractor has notified the End User that automated synthetic transactions will not be available until the 1st Quarter calendar year 2014 timeframe incident to the upgrade to Microsoft Server 2008. Until synthetic transaction measurements are available, E-mail Client Responsiveness will rely on existing E-mail Availability and Performance measures defined in Attachment 2B, Transition Service Level Agreements. These improved measures using automated synthetic transactions (vice manual methods) are a priority for the End User. Upon availability, the End User and Contractor shall, within six months, revise the measurement CONOPs and SLA targets to incorporate the four defined automated synthetic transactions described above.

E-mail Client Responsiveness is measured by sampling. All sites with 24 or greater seats will be configured and conduct synthetic transactions from at least two on-site representative points to ensure reportable data from at least one on-site representative point. If the site receives service from multiple servers than each probe will test a different server.

E-mail Client responsiveness will utilize the actual value received for any failed test without an associated network or server availability outage documented in another SLA measurement. The Contractor can select the best response from any of the site probes for any particular time measurement to account for individual seat issues, the expressed intent is to ensure the availability of an appropriate measurement for each time interval at each site.

For Client Responsiveness, the End User will approve the location of the measurement points.

End-to-end performance will utilize the actual value received or (TBD goal 30 minutes) for any failed test without an associated network or server availability outage documented in another SLA measurement.

Measurement CONOPS: All E-mail servers are monitored by Tivoli TEC. If there is an outage, a TEC event will be detected and a Remedy ticket will be created with the start time of the event.

On a monthly basis, all E-mail Server Service customer-impacting tickets, which were closed in the given reporting month, will be collected and assessed for SLA reporting purposes. Those tickets that have been categorized with a Category/Type/Item combination that relates to E-mail Server Service, will be reviewed based on the field titled Total Time Open in Seconds. This field represents all work time associated with the ticket and will have excluded any time the ticket was pending due to input/access needed from customer. The Total Time Open in Seconds fields will be combined and calculation will be performed.
The following will be excluded from measurement:

- Non-active servers (e.g., backup servers in server clusters) do not count if multiple servers provide the service.

Who: Contractor	**Frequency:** Monthly
Where: Client Site **User Population:** All Network Users **Sample Size:** - Sites>=24seats will have at least two on-site representative points - Sites<24 seats will not be measured unless mutually determined by End User and Contractor **Sample Unit:** Client **Where Measured:** Client (representative of site)	**How Measured (i.e., captured):** With an automated tool. **Measurement Formula:** Number of attempts successful within the required Time Interval / Total number of attempts **Frequency of Measure:** (goal 5 min) TBD **Weighting (as applicable):** Weighted Average (by seat count)

Aggregation of Data:	Performance data for sites that have not yet achieved Full Performance will be aggregated at the site level and the SLA targets will apply at the site level. Performance data for sites that have achieved Full Performance will be aggregated at the user population level and the SLA targets will apply at the user population level.
SLA Success Criteria	All targets must be met, to pass the SLA

SLA Target		**Time Interval**	**Percentage Complete**
	Client Responsiveness	(goal <= 2.00 sec) TBD	(goal >= 95.0%) TBD

		(goal <= 4.00 sec) TBD	(goal >= 99.5%) TBD

Service Name: WEB AND PORTAL SERVICES	**SLA: 103**
Service Description: Web site or service offering an array of resources and services to the network users.	
Performance Category: Web and Portal Services	**Increment 1** **SLA: 103.1**

Performance Category Description: Web and Portal Services are the Contractor provided services that allow end users to access web content as supported by the network. This SLA applies to web/portal services obtained through a network-connected user workstation and excludes services obtained through RAS. The performance measure for Web Services is End-to-End Performance.

> **End-to-End Performance:** Percentage of synthetic web transactions successfully processed and returned in the required time (i.e., Time Interval (*x*) seconds roundtrip). Web-access transactions are generated at the client, processed through the network (including PKI infrastructure), resulting in an authenticated website displayed on the client Internet browser.
>
> The measurement of end-to-end performance will include validation of:
> - Supporting PKI services (excludes initial authentication of a DoD PKI certificate)
> - A representative PKI-enabled, User Network-hosted static website
> - A B1 and/or B1 DMZ security suite
> - Supporting Domain Name Services

The intent of this measure is to provide indication of the performance of the end-to-end set of service components required for the end-user to access Contractor-hosted web and portal services located in the DMZ. It is targeted at providing indication of the services obtained from the network operations center (NOC) where the B1 and network portal are located.

The transactions for Web and Portal Services End-to-End performance, are conducted using synthetic scripts that replicate the actions of a web request. The intent of the measures is to verify that the supporting User networks, domain name server, directory, security boundaries, web servers, remote procedure calls is available and functioning satisfactorily. End-to-End measurement will be a representative sampling of local, regional, and enterprise infrastructure performance.

The Contractor has notified the End User that automated synthetic transactions will not be available until the 1st Quarter calendar year 2014 timeframe incident to the upgrade to Microsoft Server 2008. Until synthetic transaction measurements are available, the defined Web and Portal measurement -- End-to-End Performance will rely on existing Web Access Services Availability and Performance measures defined in Attachment 2B, Transition Service Level Agreements. These improved measures using automated synthetic transactions (vice manual methods) are a priority

for the End User. Upon availability, the End User and Contractor shall, within six months, revise the measurement CONOPs and SLA targets to incorporate the defined automated synthetic transactions described above.

All sites with 24 or greater seats selected for sampling will be configured and conduct synthetic transactions from at least two on-site representative points to ensure reportable data from at least one on-site representative point. Each probe will test a different server. Client responsiveness will utilize the actual value received or (TBD goal 30 sec) seconds for any failed test without an associated network or server availability outage documented in another SLA measurement. The Contractor can select the best response from any of the probes at a given site for any particular time measurement to account for individual seat issues, the expressed intent is to ensure the availability of an appropriate measurement for each time interval at each site sampled.

For End-to-End Performance, the End User will approve the location of the measurement points.

Measurement CONOPS: TBD

Who: Contractor	**Frequency:** Monthly
Where: Client Site **User Population:** All Network Users **Sample Size:** TBD **Sample Unit:** Client **Where Measured:** Measured at all sites using a representative client workstation to a B1 DMZ web server	**How Measured (i.e., captured):** Automated tool **Measurement Formula:** Number of attempts successful within the required time interval / Total number of attempts **Frequency of Measure:** TBD **Weighting (as applicable):** Equal weighting for all sites.
Aggregation of Data:	Performance data for sites that have not yet achieved Full Performance will be aggregated at the site level and the SLA targets will apply at the site level. Performance data for sites that have achieved Full Performance will be aggregated at the user population level and the SLA targets will apply at the user population level.
SLA Success Criteria	All target s must be met to pass the SLA

SLA Target	**End-to-End Performance**	**Time Interval (x)**	**% Complete**
		(goal <= 5.00 sec) TBD	(goal >= 95.0%) TBD
		(goal <= 8.00 sec) TBD	(goal >= 99.8%) TBD

Service Name: FILE SHARE SERVER SERVICES	**SLA: 104**
Service Description: Provides access to digital information stored on the network.	
Performance Category: File Share Services – Server Availability	**Increment 1** **SLA: 104.1**

Performance Category Description: File Share Services is the Contractor provided service that allows end users to store and retrieve files on shared, controlled access storage media. This SLA applies to a network-connected user, at his/her assigned normal User workstation site, and the shared network storage assigned to that site. The performance measures for File Shared Services are Server Availability and Client Responsiveness.

Server Availability: Percentage of time the end user's File Share Service is active and available for transfer. Server Availability is measured at every File Share server. The availability measure does not include any supporting network infrastructure.

Note: Server Availability for file share is not an end-to-end measure and depends on the companion SLA for E-mail to provide indication of the availability of the intervening user to server end-to-end availability.

Measurement CONOPS:

All file servers are monitored using Tivoli TEC. If there is an outage, a TEC event will be detected and a Remedy ticket will be created with the start time of the event.

On a monthly basis, all File Server Availability customer-impacting tickets, which were closed in the given reporting month, will be collected and assessed for SLA reporting purposes. Those tickets that have been categorized with a Category/Type/Item combination that relates to SLA 103.3.1, File Server Availability, will be reviewed based on the field titled Total Time Open in Seconds. This field represents all work time associated with the ticket and will have excluded any time the ticket was pending due to input/access needed from customer. The Total Time Open in Seconds fields will be combined and calculation will be performed.

Who: Contractor	**Frequency:** Monthly
Where: Client Site **User Population:** All Users (measured separately) **Sample Size:** All servers **Sample Unit:** File Share Server **Where Measured:** Server	**How Measured (i.e., captured):** Automated tool and end user calls to Help Desk **Measurement Formula:** For sites that have not achieved full performance: Total available minutes of file share servers at the server farm/ Total minutes in the month x total number of file share servers For sites that have achieved full performance: Total available minutes of file share servers / Total minutes in the month x total number of file share servers **Frequency of Measure:** Continuous **Weighting (as applicable):** Equal weighting

Aggregation of Data:	Sites that have not yet achieved Full Performance shall meet the requisite target(s) at the site level. Sites shall inherit the performance level of the file share servers that provide the service to them. Sites having achieved Full Performance will be aggregated at the user population level.	
SLA Success Criteria	All target levels within each performance measure must be met, to pass the SLA.	
SLA Target	**Server Availability** (for sites that have not achieved full performance)	>= 99.50%
	Server Availability (for aggregation of sites that have achieved full performance)	>= 99.80%

Performance Category: File Share Services – Client Responsiveness	**Increment 1** **SLA: 104.2**
Performance Category Description: File Share Services is the Contractor provided service that allows end users to store and retrieve files on shared, controlled access storage media. This SLA applies to a network-connected user, at his/her assigned normal User workstation site, and the shared network storage assigned to that site. **File Share Services - Client Responsiveness**: This key SLA measures the network responsiveness to the end user by demonstrating the data transfer time of the host server, both to pull a file from the server and to push a file to the server. It is the average time the synthetic file transactions take to successfully transfer a scripted 1MB file between a File Share server and a network user. Client Responsiveness is measured at the user workstation. Client Responsiveness is measured by sampling. All sites with 24 or greater seats will be configured and conduct synthetic transactions from at least two on-site representative points to ensure reportable data from at least one on-site representative point. If the site receives service from multiple servers than each probe will test a different server. Sites with fewer than 24 seats will not be measured unless mutually determined by the End User and the Contractor. Client responsiveness will utilize the actual value received or 10 seconds for any failed test without an associated network or server availability outage documented in another SLA measurement. The Contractor can select the best response from any of the site probes for any particular time measurement to account for individual seat issues, the expressed intent is to ensure the availability of an appropriate measurement for each time interval at each site.	
Measurement CONOPS: The file transfer tests are performed using a Visual Basic Intrinsic. The shared disk environment is setup prior to starting the measurement timer. A 1-megabyte file of random characters is used for the network-attached test. The 103.3.2 SLA measurement of LAN connected workstations copies the 1–megabyte file once every 5 minutes to the fileserver (KAPM_FILECOPYUP). The second part of the test reverses the process and workstation copies the 1-megabyte file from the fileserver to the local disk once every 5 minutes (KAPM_FILECOPYDOWN).	
Who: Contractor	**Frequency:** Monthly

Where: Client Site **User Population:** All Network Users **Sample Size:** • Sites >=24seats will have at least two on-site representative points • Sites <24seats will not be measured unless mutually determined by End User and Contractor **Sample Unit:** Client **Where Measured:** Client (Representative of site.)	**How Measured (i.e., captured):** Automated tool **Measurement Formula:** Sum of all non-excluded client responsiveness measured values / Total number of non-excluded attempts **Frequency of Measure:** Client Responsiveness- Every 5 minutes **Weighting (as applicable):** Equal weighting	
Aggregation of Data:	Performance data for sites that have not yet achieved Full Performance will be aggregated at the site level and the SLA targets will apply at the site level. Sites shall inherit the performance level of the file share servers that provide the service to them. Performance data for sites that have achieved Full Performance will be aggregated at the user population level and the SLA targets will apply at the user population level.	
SLA Success Criteria	All targets must be met, to pass the SLA	
SLA Target	**Client Responsiveness**	**Average Time Interval**
		<= 2.00 sec

Service Name: PRINT SERVICES	SLA: 105
Service Description: Collection of software components residing on a server or servers that provide network printing services for client computers.	
Performance Category: Print Services	**Increment 1** **SLA: 105.1**
Performance Category Description: Print Services is the Contractor provided service that allows end users to produce black & white and color hard copies of electronic documents and transparencies. This SLA applies to a network-connected user, at his/her assigned normal user workstation site, and the shared network print server assigned to that site. The performance measure for Print Services is Server Availability. **Server Availability:** Percentage of time that Print queues are active and available at the Print Server for transferring a print job to a local printer. Server Availability is measured at every Print server.	
Measurement CONOPS: All print servers are monitored using Tivoli TEC. If there is an outage, a TEC event will be detected and a Remedy ticket will be created with the start time of the event. On a monthly basis, all Print Server Availability customer-impacting tickets, which were closed in the given reporting month, will be collected and assessed for SLA reporting purposes. Those tickets that have been categorized with a Category/Type/Item combination that relates to this SLA, Print Server Availability, will be reviewed based on the field titled Total Time Open in Seconds. This field represents all work time associated with the ticket and will have excluded any time the ticket was pending due to input/access needed from customer. The Total Time Open in Seconds fields will be combined and calculation will be performed.	
Who: Contractor	**Frequency:** Monthly
Where: Client Site **User Population:** All Network Users **Sample Size:** All Servers **Sample Unit:** Print Server **Where Measured:** Server	**How Measured (i.e., captured):** Automated tool and end user calls to Help Desk **Measurement Formula:** For sites that have not achieved full performance: Total available minutes of print servers at the server farm/ Total minutes in the month x total number of print servers For sites that have achieved full performance: Total available minutes of print servers / Total minutes in the month x total number of print servers **Frequency of Measure:** Continuous **Weighting (as applicable):** Equal weighting
Aggregation of Data:	Performance data for sites that have not yet achieved Full Performance will be aggregated at the site level and the SLA targets will apply at the site level. Sites shall inherit the performance level of the print servers that provide the service to them. Performance data for sites that have achieved Full Performance will be aggregated at the user population level and the SLA targets will apply at the user population level.

SLA Success Criteria	All targets must be met, to pass the SLA	
SLA Target	**Server Availability**	>= 99.50%

SERVICE NAME: RAS SERVICES	**SLA: 106**
Service Description: Enables users to log into the network remotely	
Performance Category: RAS Services – Service Availability	**Increment 1 SLA: 106.1**

Performance Category Description: RAS is the Contractor-provided service that allows users to remotely and securely connect to the network. A remote user accesses the network by connecting a laptop to an analog phone line and launching an application to connect to the Contractor Dial Access Network (DAN). The Contractor DAN has filters to route traffic to the RAS Transport Boundary. The user then launches a VPN application to create a secure data tunnel into the network to gain access to network services.

> **RAS Service Availability:** Percentage of time that the RAS Dial-up Service is active at the NOC and available for access. WAN access circuits at each RAS Access Points are the transport for the network destined traffic once a user successfully connects to the Contractor DAN. The availability measure excludes any supporting network infrastructure not controlled by or contracted for by the network.

For RAS Availability, as long as any one of the test sites in each COI (e.g. one of the RAS Access Points) is meeting the test for that 5 minute test cycle, the test is successful.

Note: All RAS modems will operate at the current industry standard connectivity rate, and support automated selection of lower rates based on geographic distance, and modem and line quality. This measurement is based on the use of an industry standard modem -- currently 56Kb/sec.

Measurement CONOPS:

RAS Availability: The KAPM script collects data every 5 minutes; every 6 hours the data is uploaded to an Oracle DB via a Tivoli custom inventory scan.. The data extracted by Business Objects are KAPM probes fired 24 X 7 excluding the hours of 0900, 1900, and 2300 local time when the KAPM probe is used to measure SLA 103.7.2.

For the RAS portion of the SLA measurement, the connection speed is collected into the MIF file and subsequently into the Oracle database. On the initial release of KAPM, a single UUNET access point number is dialed and left connected. The number is recorded in the MIF file for each measurement and subsequently placed in the Oracle database.

Who: Contractor	**Frequency:** Monthly
Where: Client Site **User Population:** All Network Users	**How Measured (i.e., captured):** **Measurement Formula:**

Sample Size: One representative user per RAS access point, using a representative standard user laptop, dialing into a local Contractor DAN POP and authenticating with a VPN gateway **Sample Unit:** RAS Access point **Where Measured:** RAS Access Point, or other End User Facilities	Total available hours of RAS Connectivity for the test period / (1260 minutes x numbers of days in the month) **Frequency of Measure:** Every 5 Minutes, excluding the hours of 0900, 1900, and 2300 local time when the KAPM probe is used to measure this SLA. **Weighting (as applicable):** Equal weighting	
Aggregation of Data:	Sites will be aggregated at the user population level.	
SLA Success Criteria	All targets must be met, to pass the SLA	
SLA Target	**RAS Service Availability**	>= 99.00%

Performance Category: RAS Services – Client Responsiveness	**Increment 1** **SLA: 106.2**

Performance Category Description: RAS is the Contractor-provided service that allows users to remotely and securely connect to the network. A remote user accesses the network by connecting a laptop to an analog phone line and launching an application to connect to the Contractor Dial Access Network (DAN). The Contractor DAN has filters to route network traffic to the RAS Transport Boundary. The user then launches a VPN application to create a secure data tunnel into the network to gain access to network services.

> **Client Responsiveness:** Percentage of synthetic file transactions that fall within the required response time to successfully transfer a scripted 100KB file between a File Share server and a RAS client. This measurement will be taken during a connection to the Contractor DAN of at least 52.3 Kb/sec. This key SLA measures the RAS responsiveness to the end user by demonstrating the data transfer time of the RAS connectivity, both to download a file from the server and to upload a file to the server during one session. RAS Dial-up Client Responsiveness is measured at the representative user laptop. The measure is structured to factor out the effects of the dial in line.

For RAS Availability, as long as any one of the test sites in each COI (e.g. one of the RAS Access Points) is meeting the test for that 5 minute test cycle, the test is successful.

Note: All RAS modems will operate at the current industry standard connectivity rate, and support automated selection of lower rates based on geographic distance, and modem and line

quality. This measurement is based on the use of an industry standard modem -- currently 56Kb/sec.

Measurement CONOPS: The file transfer tests are performed using a Visual Basic Intrinsic. This SLA measurement copies 100 KB file thirty-one times from the local disk to the shared drive from the file server (KAPM_RAS_FILECOPYUP) per hour. The second part of the test reverses the process and thirty-one copies are made from the file server share to the local drive (KAPM_RAS_FILECOPYDOWN) per hour. After the test is complete, the shared disk environment is torn down. The test result is recorded in the MIF file and subsequently into the Oracle database.

Who: Contractor	**Frequency:** Monthly
Where: Client Site **User Population:** All Network Users. **Sample Size:** One representative user per RAS access point, using a representative standard User laptop, dialing into a local Contractor DAN POP and authenticating with a VPN gateway **Sample Unit:** Client **Where Measured:** From the Client to a supporting File Server not collocated with the Service Access Point.	**How Measured (i.e., captured):** **Measurement Formula:** Number of attempts successful within the required time interval/ Total number of attempts **Frequency of Measure:** 0900, 1900, 2300 local time, 7 days/week, for one hour each, Thirty-one copies are made from the local disk to the shared drive from the file server. The second part of the test reverses the process and thirty-one copies are made from the file server share to the local drive. **Weighting (as applicable):** Equal weighting
Aggregation of Data:	Sites will be aggregated at the user population level.
SLA Success Criteria	All targets must be met, to pass the SLA

SLA Target			**Time Interval**	**Percentage Complete**
	Client Responsiveness (100KB file transfer)	**Upload**	<= 40.0 sec	>= 90.00%
		Download	<= 22.0 sec	>= 90.00%

LIST OF REFERENCES

[1] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," National Institute of Standards and Technology Computer Security Division, Gaithersburg, MD, Rep: 800–145, pp. 1–7, 2011.

[2] R. Ghannoum, Swedish Institute of Computer Science, "Service Level Agreements," Dec. 2005, http://www.sics.se/~rabih/.

[3] I. S. Hayes, Clarity Consulting, Inc., "Metrics for IT Outsourcing Service Level Agreements," 2011, www.clarity-consulting.com/metrics_articl.htm.

[4] P. Bianco, G. A. Lewis, and P. Merson, Service Level Agreements in Service-Oriented Architecture Environment, Sept. 2008, http://www.sei.cmu.edu.

[5] A. Andrieux, "Web Services Agreement Specification (WS-Agreement)," Mar. 14, 2007, http://forge.gridforum.org/sf/projects/graap-wg.

[6] R. Matlus, and K. Brittain, "Creating a Service-Level Agreement for the IS Organization," Gartner Research, Tech. Rep. 15–1751, pp. 1–5, Jan. 2002.

[7] M. Macias, G. Smith, O. Rana, J. Guitart, and J. Torres, *Enforcing Service Level Agreements Using an Economically Enhanced Resource Manager, S*witzerland: Birkhauser Verlag Basel, 2009, pp. 109–127.

[8] "Best Practices: Service Level Management–Life Cycle Overview," white paper, Compuware Corporation, 2008.

[9] O. Rana, M. Warnier, T. B. Quillinan, F. Brazier, and D. Cojocarasu, "Managing Violations in Service Level Agreements," M.S. thesis, Cardiff University, UK, April 12, 2012.

[10] Cloud Standards Wiki, "Main Page," June 14, 2012, http://Cloud-standards.org/wiki/index.php?title=Main_Page.

[11] General Services Administration, "FedRAMP CONOPS," Feb. 7, 2012, http://www.gsa.gov/graphics/staffoffices/FedRAMP_CONOPS.pdf.

[12] NIST, Cloud Computing Reference Architecture Contracts and SLA - DRAFT Recommendations of the National Institute of Standards and Technology (NIST) Public Working Group, Jan. 2012, http://collaborate.nist.gov/twiki-Cloud-computing/pub/CloudComputing/RATax_Jan20_2012/NIST_CC_Public_WG_ ContractSLA_Deliverable_Draft_v16.pdf.

[13] "IT Control Objectives for Cloud Computing: Controls and Assurance in the Cloud," Whitepaper, ISACA, October 2011.

[14] "Cloud Computing: Business Benefits With Security, Governance and Assurance Perspectives," Whitepaper, ISACA, October 2009.

[15] T. Mather, S. Kumaraswamy, and S. Latif, *Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance,* Farnham: O'Reilly, 2009.

[16] T. Nolle, "Is There Any Hope for the Cloud SLAs?" http://searchCloudcomputing.techtarget.com/tip/Is-there-any-hope-for-Cloud- SLAs, October 2011.

[17] J. M. Johansson, "Windows on Amazon EC2 Security Guide," Sept. 4, 2012, http://aws.amazon.com/articles/1767.

[18] Amazon Web Services, "Summary of the AWS Service Event in the US East Region," Aug. 31, 2012, http://aws.amazon.com/message/67457/.

[19] A. Weissberger, "What Should Cloud Computing Users and Providers Consider for SLAs? Jan. 13, 2011, http://viodi.com/2011/01/13/what-should- Cloud-computing-users-and-providers-consider-for-slas/.

[20] Cloud Computing Use Case Discussion Group. "Cloud Computing Use Cases White Paper – Version 4.0," July 2, 2010, http://Cloudusecases.org.

[21] "Service-level Agreement," in *Wikipedia, the Free Encyclopedia* [Online], August 2009, Available: http://en.wikipedia.org/wiki/Service-level_agreement [April 11, 2012].

[22] A. Correia and F. Abreu, "Defining and Observing the Compliance of Service Level Agreements: A Model Driven Approach," in *Proceedings of the Seventh International Conference on the Quality of Information and Communications Technology,* 2010.

[23] A. Freitas, N. Parlavantzas, and J. Pazat, "An Integrated Approach for Specifying and Enforcing SLAs for Cloud Services," in *Proceedings of the IEEE 5th International Conference on Cloud Computing (CLOUD),* 2012.

[24] A. Freitas, N. Parlavantzas, and J. Pazat, "A QoS Assurance Framework for Distributed Infrastructures," in *Proceedings of the 3rd International Workshop on Monitoring, Adaptation and Beyond,* 2010, pp. 1–8.

[25] N. Oldham, K. Verma, A. Sheth and F. Hakimpour, "Semantic WS-agreement Partner Selection," in *Proceedings of the 15th International Conference on World Wide Web*, 2006, pp. 697–706.

[26] A. Paschke and M. Bichler, "SLA Representation Management and Enforcement," in *Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service*, 2005, pp. 158–163.

[27] S. Easterbrook et al., "Experiences Using Lightweight Formal Methods for Requirements Modeling," IEEE Trans. Software Eng., vol. 24, no. 1, 1998, pp. 4 –11; doi:10.1109/32.663994.

[28] W. Ziegler. "OPTIMIS: Improving Cloud Management With Dynamic SLAs," workshop on *Science Agency Uses Of Clouds And Grids*, July 18, 2011, Salt Lake City.

[29] M. A. Bochicchio and A. Longo. "Modelling Contract Management for Cloud Services," in *Proceedings of the 2011 IEEE 4th International Conference on Cloud Computing*, 2011, pp. 332–339.

[30] H. Goudarzi and M. Pedram. "Multi-dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems*,"in Proceedings of 2011 IEEE 4th International Conference on Cloud Computing*, 2011, pp. 324–331.

[31] T. Schaaf. "Frameworks for Business-driven Service Level Management - A Criteria–based Comparison of ITIL and NGOSS," in *Proceedings of 2nd IEEE/IFIP International Workshop on Business–Driven IT Management,*, 2007, pp. 65–74.

[32] L. T. Gaines, "Improving software quality and management through use of service level agreements," Diss. Naval Postgraduate School, Monterey, CA, 2005.

[33] D. Ameller, and X. Franch. "Service Level Agreement Monitor (SALMon)," *Proceedings of Seventh International Conference on Composition-based Software Systems*, February 2008, pp. 1–4.

[34] M. Comuzzi, K. C. Kotsokalis, G. Spanoudakis, and R. Yahyapour. "Establishing and Monitoring SLAs in Complex Service Based Systems," in *Proceedings of 2009 IEEE International Conference on Web Services*, 2009.

[35] F. Shulz. "Towards Measuring the Degree of Fulfillment of Service Level Agreements," in *Proceedings of Third International Conference on Information and Computing*, 2010, pp. 273–276.

[36] M. Vael, “Cloud Computing–An Insight in the Governance & Security Aspects,” *ISACA Belgium Chapter Meeting*, May 2010.

[37] Cisco Sytems, Inc. *User Guide for Cisco Unified Service Statistics Manager*, Cisco Systems Inc., 2009.

[38] A. Freitas, N. Parlavantzas, and J. Pazat, An Integrated Approach for Specifying and Enforcing SLAs for Cloud Services,” in *Proceedings of IEEE Fifth International Conference on Cloud Computing*, 2012, pp. 376–383.

[39] S. Overby, “Tips for Outsourcing Incentives and Penalties that Work,” www.cio.com.au. CIO, 2009, http://www.cio.com.au/article/301302/tips_outsourcing_incentives_penalties_w ork/.

[40] R. Sturm, “SLA Penalties for Outsourcers: What Are Their Pain Points?” *Network World*, Feb. 14, 2005, http://www.networkworld.com/newsletters/2005/0214out1.html..

[41] B. Golden, *“Cloud Computing and the Truth About SLAs,”* www.cio.com. CIO, 2011, http://www.cio.com/article/693535/Cloud_Computing_and_the_Truth_About_ SLAs.

[42] S. Zhang, S. Zhang, X. Chin, and X. Huo, “The Comparison Between Cloud Computing and Grid Computing,” in *Proceedings of 2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*.

[43] D. Chappell, “A Short Introduction To Cloud Platforms,” Aug. 2008, http://www.davidchappell.com/CloudPlatforms--Chappell.pdf.

[44] F. Ohihorst, “Making sense of Cloud-based SLAs,” June 2009. http://searchCloudcomputing.techtarget.com/tip/Making-sense-of-Cloud-Based- SLAs.

[45] T. Nolte, “Meeting performance standards and SLAs in the Cloud,” May 2009, http://searchCloudcomputing.techtarget.com/tip/Meeting-performance- standards-and-SLAs-in-the-Cloud.

[46] P. Patel, A. Ranabahu, and A. Sheth, “Service Level Agreement in Cloud Computing,” M.S. thesis. Knoesis Center, Wright State University, USA, 2009.

[47] Cloud Computing Use Case Discussion Group, “Cloud Computing Use Cases,” White Paper, July 2, 2010.

[48] T. J. Trappler, "If It's in the Cloud, Get It on Paper: Cloud Computing Contract Issues," *EDUCAUSE Quarterly Magazine,* 2 Nov. 2010.

[49] A. Mehendale, "Importance of an SLA for Any Organization," June 5, 2008, http://blog.maia-intelligence.com/2008/06/05/importance-of-an-sla-for-any- organization/.

[50] J. Trienekens, M. Z.wan, and J. Bouman, "Specification of Service Level Agreements, clarifying concepts on the basis of practical research, *Proceedings of the Software Technology and Engineering Practice*, 1999.

[51] M. Smit and E. Stroulia, "Maintaining and Evolving Service Level Agreements: Motivation and Case Study," in *Proceedings of the International Workshop on the Maintenance and Evolution of Service-Oriented and Cloud- Based Systems (MESOCA),* Sept., 2011.

[52] H. Eliadis and A. Rand, "Setting Expectations in SAAS: The Importance of the Service Level Agreement to SAAS Providers and Consumers,"*Software & Information Industry Association (SIIA) Software as a Service Working Group*. 2007.

[53] G. Peterson, "Federal Cloud Security Challenges and Solutions," Artec Group, White Paper, 2011.

[54] J. Brown and P. Robinson, "PKI Reborn in the Cloud," in *Proceedigs of RSA, Conference Europe*, pp. 1–61, 2011.

[55] M. Maxim, "PKI Still Matters, Especially in the Cloud," July 15, 2011, https://blog.cloudsecurityalliance.org/2011/07/15/pki-still-matters-especially-in- the-cloud/.

[56] K. Fakhfakh, S. Tazi, K. Drira, T. Chaari, and M. Jmaiel, "Semantic Enabled Framework for SLA Monitoring," *International Journal on Advances in Software 2.1,* pp. 36–46, 2009.

[57] A. Zhdankin, "SLA Management in Next Generation Networks," 2007, http://dmtf.org/sites/default/files/files/SLA_Management_in_Next_Genera tion _Networks_FINAL.pdf.

[58] S. Ried, "Trust.platform.com–How To Communicate SLAs In The Cloud?" Sept. 15, 2010, http://blogs.forrester.com/stefan_ried/10–09–15-trustplatformcom_%E2%80%94_how_communicate_slas_Cloud.

[59] K. Phaltankar, "Automating Cloud Security Authorizations," *Journal of Software Technology*, 2011.

[60] K. Goertzel, K. M. Holly, L. M. Schmidt, T. Winograd, and K. Mosteller, Cloud Computing Security, 2009, https://www.boozallen.com.

[61] C. Fortuna and M. Mohorcic, "Dynamic Composition of Services for End-To-End Information Transport," IEEE Wireless Communications, vol. 16, pp. 56–62, Aug. 2009.

[62] D. Lamanna, J. Skene, and W. Emmerich, "SLAng: A Language for Defining Service Level Agreements," in Proceedings of *The Ninth IEEE Workshop on Future Trends of Distributed Computing Systems*, May 2003, pp. 100–106.

[63] A. Paschke, "RBSLA A Declarative Rule-based Service Level Agreement Language Based On RuleML," in *Proceedings of International Conference on Computational Intelligence for Modelling, Control and Automation*, 2005.

[64] A. Keller and H. Ludwig, "Defining and Monitoring Service Level Agreements for Dynamic E-Business," in *Proceedings of the 16th System Administration Conference (LISA 2002)*, November 2002.

[65] A. Zhdankin, "SLA Management in Next Generation Networks," https://dmtf.org/sites/dafault/files/files/, 2007.

[66] NIST, "Cloud Computing Forum & Workshop V," http://www.nist.gov/itl/Cloud/Cloudworkshopv.cfm, June 14, 2012.

[67] General Services Administration (GSA), "FedRAMP," http://www.gsa.gov/portal/category/102371, June 14, 2012.

[68] General Services Administration (GSA), "FedRAMP Processes," http://www.gsa.gov/portal/content/131919, June 14, 2012.

[69] L. T. Gaines, "Improving software quality and management through use of service level agreements," Ph.D Dissertation, Naval Postgraduate School, Monterey, CA, March 2005.

[70] Amazon Web Services, "Summary of the Amazon EC2 and Amazon RDS Service Disruption," Apr. 29, 2011, http://aws.amazon.com/message/65648/.

[71] H. Ganesan, "Cloud, Big Data and Mobile," Sept 4, 2012, http://harish11g.blogspot.com/2012/07/amazon-availability-zones-aws-az.html.

[72] J. Nisha, “Cloud Computing–An Overview on Cloud Computing Concepts,” Sept. 6, 2011, http://www.indiastudychannel.com/resources/144808-Cloud- Computing-An-Overview-Cloud-Computing.aspx.

[73] Verizon Business, *Managed PKI for Remote Access*, Sept. 14, 2012, http://www.verizonbusiness.com/terms/us/products/security/managedPKI/.

[74] K. Hwang, “Security, Privacy, and Data Protection for Trusted Cloud Computing,” in Proceedings of *International Conference on Parallel and Distributed Computing and Systems,* 2010.

[75] P. Shafer, International Association for Contract & Commercial Management, “How SLAs drive, and don’t drive, performance: strategic, technical and process limitations,” http://www.iaccm.com/news/contractingexcellence/?storyid=514.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Man-Tak Shing
   Naval Postgraduate School
   Monterey, California

4. James Bret Michael
   Naval Postgraduate School
   Monterey, California

5. CDR Kurt Rothenhaus
   PMW/A 170 Deputy Program Manager
   Communications and GPS Navigation
   San Diego, California

6. Charles Suggs
   SPAWAR PEOC4I
   San Diego, California

7. Nancy J. Kelley
   SPAWAR Sytems Center - Pacific
   San Diego, California