Service Level Agreements on IP Networks.

Dinesh C. Verma
IBM T. J Watson Research Center
PO Box 704, Yorktown Heights, NY-10598, USA
Email: dverma@us.ibm.com

Abstract: *This paper provides an overview of service-level agreements in IP networks. It looks at the typical components of a service-level agreement, and identifies three common approaches that are used to satisfy service level agreements in IP networks. The implications of using the approaches in the context of a network service provider, a hosting service provider, and an enterprise are examined. While most providers currently offer a static insurance approach towards supporting service level agreements, the schemes that can lead to more dynamic approaches are identified.*

Keywords: SLA, Service Level Agreements, Network Management.

1. Introduction

A Service Level Agreement (SLA) is a formal definition of the relationship that exists between a service provider and its customer. A SLA can be defined and used in the context of any industry, and is used to specify what the customer could expect from the provider, the obligations of the customer as well as the provider, performance, availability and security objectives of the service, as well as the procedures to be followed to ensure compliance with the SLA. Service level agreements are often used when corporations outsource functions considered outside the scope of their own core competencies to third party service providers. The operation and maintenance of computer networks is outsourced by many companies to third-party network providers, making SLA support an important subject in the context of computer networks.

This paper looks at the different approaches used for supporting service level agreements in computer networks, specifically in the context of networks based on the Internet Protocol. In the next section of this paper, we look at the typical components of a service level agreement. This is followed in Section 3 by a description of some common environments of IP networks where service level agreements play an important role. Section 4 examines the different approaches that are used by different organizations in order to meet the obligations of their service level agreements in these environments. Finally, we summarize the status of service level agreement support in IP networks, and identify areas for further research.

2. Typical Components of SLAs

A service level agreement would typically contain the following information:

- *A description of the nature of service to be provided:* It includes the type of service to be provided, and any qualifications of the type of service to be provided. In the context of IP network connectivity, the type of service may specify the

maintenance of network connectivity, or it may include additional functions such as operation and maintenance of domain name servers, dynamic host configuration protocol servers, etc.

- *The expected performance level of the service, specifically its reliability and responsiveness*: Reliability includes availability requirements; when is the service available, and what are the bounds on service outages that may be expected. Responsiveness includes how soon the service would be performed in the normal course of operations.

- *The procedure for reporting problems with the service*: This includes information about the person to be contacted for problem resolution, the format in which complaints have to be filed, and the steps to be undertaken in order to quickly resolve the problem. The agreement would also typically describe a time-limit by which a reported problem would be responded to (someone would start to work on the problem) as well as how soon the problem would be resolved.

- *The time-frame for response and problem resolution*: This specifies a time-limit by which someone would start investigating a problem that was reported. The start of the investigation is typically marked by a representative of the supplier contacting the customer who reported the problem initially. There may also be a time limit by which the problem would be resolved. A SLA may specify that a failed link would be recommissioned within 24 hours.

- *The process for monitoring and reporting the service level* : This outlines how performance levels are monitored and reported, i.e., who will do the monitoring, what types of statistics will be collected, how often would they be collected, and how past/current statistics may be accessed. Some network providers may allow the customer to directly access part of the network through a network management tool. The customer would be typically provided access to monitoring and statistics information, but may not be allowed to modify the configurations or operation of the network.

- *The consequences for the service provider not meeting its obligations*: It is customary to extend some credits to the customers when the service expectations are not met. Other consequences of not meeting the obligation may include the ability of the customer to terminate its relationship, or to ask for reimbursement of part of the revenues lost due to loss of service. The consequences of not meeting the SLA may vary depending on the nature of the relationship between the customer and the supplier.

- *Escape clauses and constraints*: Escape clauses are conditions under which the service level does not apply, or under which it would be considered unreasonable to meet the requisite service level agreements, e.g. when the service provider's equipment have been damaged in flood, fire or war. They often also impose some constraints on the behavior by the customer. A network operator may void the

service level agreement if the customer is attempting to breach the security of the network.

Not all of the components of a SLA may be present in all contracts, but a good SLA would provide an overview of the different items that can go wrong with the provided service, and attempt to cover those situations as part of the SLA agreement.

3. IP Network Environments

Within the context of an IP networks, SLAs are typically provided for three common types of operating environments. Each of these environments consists of service providers offering a different type of service to their customers. The three common services provided in IP networks are network connectivity services, hosting services, and integrated connectivity and hosting services.

3.1 Network Connectivity Services

Network connectivity services are provided by several telecommunications companies to large enterprises. They provide the access links to the different sites of its customers, enabling customer sites to be connected to each other as an intranet, or provide the access to the global Internet from the customer's site. Customer networks are attached to the provider network via access routers that are present at the access points. A typical scenario is shown in Figure 1. For each customer, the network operator has defined performance and availability limits in the appropriate SLA signed with the customer.

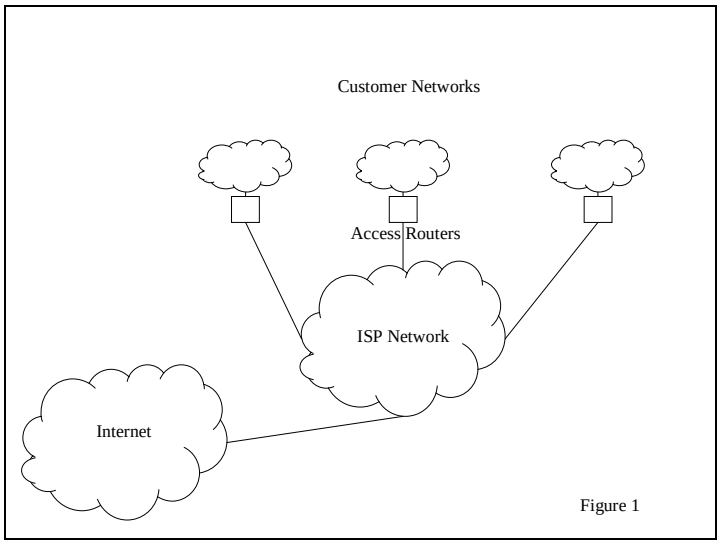

Figure 1

In this environment, typical clauses related to performance and availability may look like:

- The average delay measured monthly across the ISP network between any two access routers of the customer should be less than 200 ms.

- The average delay across the ISP network between any access router in New York City to any access router of the customer within the US would be less than 200 ms.
- The average delay across the ISP network on the transcontinental link between New York City access router of the customer and the London access router of the customer would be less than 250 ms.
- The customer will not have unscheduled connectivity disruption across the ISP network between any two access routers exceeding 5 minutes. Connectivity disruption would be defined as the loss of 100% of packets as measured by pinging an access router from a machine connected to another access router.

By offering better performance, availability and responsive customer service to its customers, a network provider would be better able to compete with its rivals. The resources within the network are provisioned so as to meet the desired performance and availability objectives, thereby reducing the operational cost without impacting the satisfaction of the customers receiving the connectivity service.

3.2 Hosting Services

Hosting services are offered by operators that host and support different types of servers on behalf of their customers. The most common case of these providers are web-hosting companies, e.g. Verio (www.verio.com) or SBC web-hosting (www.webhosting.com), that manage and provide servers to operate the web-sites for individual companies. Hosting operators provide a variety of services to their customers, and includes companies that simply provide cages in well-connected locations where customers may place their servers, companies that provide cages and maintain the uptime of the servers in the cages, as well as companies that take responsibility for running and ensuring the operation of the entire application hosted at the site. The last mode of providing hosting services, in which the provider is responsible for the overall operation of the application being hosted, is gaining ground as being more attractive to both the service provider and the customer.

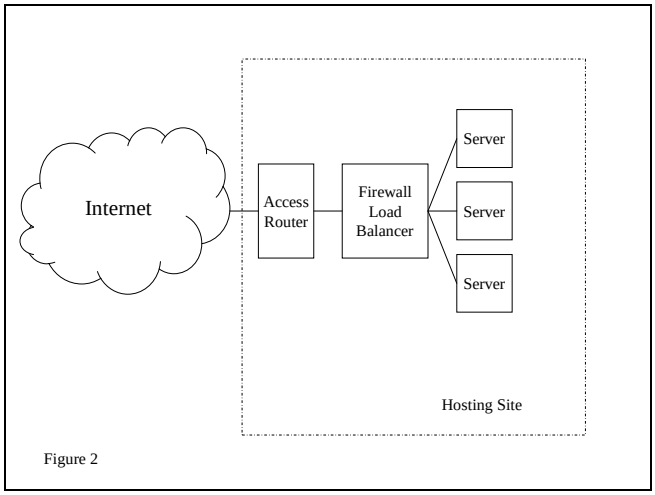

Figure 2

The typical scenario for hosting services is shown in Figure 2. The SLAs offered by hosting providers deal with the uptime and performance of the servers that are being hosted. These operators are only able to control the server side of the total

communication. In most cases, they have no control over the client side of the communication, nor over the performance of the network (usually the Internet). As a result, Service Provider SLAs usually specify the amount of sustained throughput or connection request rates that needs to be supported for a specific server. This determines the aggregated number of requests that must be handled by the server with acceptable performance. In the hosting environment, typical performance and availability clauses that are provided to the customer may look like:

- The hosted server will not be unavailable for a contiguous period exceeding 5 minutes in any 24 hour period. Unavailability is defined as the ability to ping the server from a machine with network connectivity to the hosting provider's access router.
- The hosted server will be able to handle inbound traffic of 30,000 web-requests per day.
- The hosted application will be provided access to the Internet at a bandwidth of 45 Mbps or more.
- The service provider will ensure that there are at least 5 servers available and running the application at all times.

A large hosting service provider may host multiple customers at the same site, and thus would be responsible for ensuring that the performance of one customer's server is not adversely affected by requests directed to other customers.

3.3 Integrated Services

A third type of service provides a consolidated service in which the service provider controls the network as well as the hosting infrastructure. Such a service is often provided by an enterprise information technology (IT) department that operates and maintains the intranet of an enterprise and the various applications that run within that environment. The customers of the IT department are the other departments of the enterprise. Often times, the IT department is in charge of clients as well as servers in the network, and needs to control end-to-end performance of the different applications.
In an integrated environment, customers would often expect performance and availability on the operation of the entire distributed system. Some examples of performance clauses that one may see within an enterprise IT context would look like:

- The time to perform an employee lookup on the corporate directory would not exceed 500 milliseconds.
- The average performance of a standard synthetic web-based transaction, as reported by probes located at selected locations, will not exceed 100 milliseconds.
- Unscheduled downtime of the mail server will not exceed a 30 minute period during the normal business day of 9 AM to 5 PM.

Another type of integrated service is offered by hosting service providers which would provide an integrated hosting and connectivity service to the provider. As an example, a networking service provider like AT&T also operates data-centers, and could offer a

consolidated service including network connectivity and data center hosting to its customers. An alternate way to provide an integrated service may be outsourcing some the hosting part or the networking part to another company. As an example, IBM Global Services may offer an integrated service to its customers, and obtain networking connectivity by outsourcing it to AT&T.

In all of the above operating environments, the nature of the service being provided and the performance/availability objectives, and the mechanism used to monitor the performance level would the service are different. However, the other components of the service level agreements would tend to be relatively similar in all of those environments.

The procedure for reporting problems with the service would typically describe whether the problem should be reported by calling a help-desk, or whether it can be reported via a web-based interface. The help-desk personnel or the web-interface may attempt to solve the problem using a set of known procedures, or open a problem ticket for support personnel to try to solve the problem. The time-frame for response and problem resolution clause would be similar across the different environments, and would dictate how soon action would be taken on the problem ticket. The escape clauses as well as the penalties to be paid for not monitoring the service would also be expressed in similar terms.

4. Approaches to SLA support

Supporting the appropriate level of performance and availability specified in a service level agreement is an important aspect of the operation of an enterprise. Since the inability to meet service level agreements can often result in monetary damages, the provider of services, regardless of the specific type of service being offered, attempts to meet the service level agreements to the best of their ability.

A service provider may sign SLAs with difference performance objectives with different customers. The network operator needs to identify the type of packets coming into the network, so that they can be dealt with appropriate urgency. Once the SLA has been agreed upon, the network operator needs to monitor the performance of the network. The SLA would detemine which network performance metrics ought to be monitored, as well as the operating ranges of the performance metrics.

The creation and filing of periodic reports is an important step in the process of supporting SLAs. The reports on monitored performance must be available for examination by the customer. A side-benefit of storing reports would be that the historic information can be used to extrapolate trends in network traffic, and thus be used as input to the service provisioning process.

If monitoring indicates that all the SLAs are being satisfied, there is no need for any further action. However, one may want to check if it would be possible to satisfy the same SLA constraints with a possibly cheaper or simpler configuration. If so, the

operational constraints of the network might need to be changed. A worse case would be when the SLA objectives are not being met. In this case, the network configuration must be changed, through the service provisioning process, so that the objectives can be successfully met.

As a last step of the customization process, one must examine if the agreed SLAs can be satisfied. If experience shows that the SLAs can not be met, one may want to revise the performance objectives to those that are feasible to meet. One can also revise SLA objectives to become more stringent, if that is likely to attract new customers and new streams of revenues.

Three common approaches are used to support and manage service level agreements within the three IP environments described in Section 3. The first approach takes the model of an insurance company towards monitoring and supporting SLAs. The second approach uses configuration and provisioning techniques to support SLAs within the network and the third approach takes a more dynamic and adaptive approach towards supporting service level agreements.

Insurance Approach:

In the insurance approach towards supporting SLAs, the service provider makes its best attempt to satisfy the performance, availability and responsiveness objectives that are specified in the service level agreement according to its normal operating procedures. Generally, the same level of service is offered to all of the customers. The service provider keeps on monitoring its service to check how well it is complying with the objectives set within the contract. The performance and availability parameters specified within the SLA are specified so that they are not likely to be violated during the course of normal operation of the system. When the limits are violated, the service provider would pay the penalty charges specified in the contract to the customers. Thus, the service provider is computing the financial risk associated with providing a given service level to a new customer, and revises the terms offered to customers when the financial risk associated with the SLA violation is unacceptable. The computation of such risk is what most companies in the insurance business do, and this approach can be viewed as self-insurance of a service provider against the risk of violating the service level objectives. It would not be unreasonable in the future to see service providers take out explicit insurance policies against the possible violation of their service level objectives.

In other words, the insurance approach towards supporting SLAs can be summarized as performing the following steps:

1. Identify service objectives (performance, availability, responsiveness) to be offered for the service.
2. Monitor agreed-upon service objectives.
3. Issue SLA reports, possibly including periodic meetings with the customers to discuss status of SLA compliance.
4. Issue appropriate credits to the customers if service levels are not being satisfied.

5. Periodically, modify service level objectives so that the probability of violating the objectives and the associated financial impact is acceptable.

A description of a network architectures that monitors SLA compliance can be found in [8].

Provisioning Approach:

In the provisioning approach towards satisfying service level agreements, the service provider typically signs different types of service objectives with different customers. The service provider would allocate the resources within the environment differently to each customer in order to be able to support the service level objectives for each of the individual customer. The determination of the configuration of the system is the most crucial step towards the support of service level objectives for each individual customer. Beyond this step, the service provider follows the same approach towards monitoring and payment of credits/penalties to the customers as in the insurance approach described earlier. The provisioning approach can be summarized as the following steps

1. Identify service objectives (performance, availability, responsiveness) to be provided to each customer.
2. Determine the right system configuration to be used for each of the customers.
3. Monitor agreed-upon service objectives.
4. Issue SLA reports, possibly including periodic meetings with the customers to discuss status of SLA compliance.
5. Issue appropriate credits to the customers if service levels are not being satisfied.

Adaptive Approach:

The third approach towards satisfying service level agreements adds on an additional aspect of adaptive configuration to the provisioning approach. In this approach, the service provider would dynamically modify the configuration of the system used to support the customer when monitoring indicates that the service objectives provided to the customer are in the danger of being violated. This step reduces the probability that the service will actually be violated, but does not eliminate it altogether. The steps involved in the adaptive approach are the following:

1. Identify service objectives (performance, availability, responsiveness) to be provided to each customer.
2. Determine the right system configuration to be used for each of the customers.
3. Monitor agreed-upon service objectives.
4. If monitoring indicates possible violation of objectives, reconfigure customer configuration to better server the service objectives.
5. Issue SLA reports, possibly including periodic meetings with the customers to discuss status of SLA compliance.
6. Issue appropriate credits to the customers if service levels are not being satisfied.

We can now map the above three approaches to each of the three types of services discussed in Section 3.

4.1 Network Connectivity Services

In the context of networking SLAs, the insurance approach to support service level agreements is the most prevalent one in the industry. Service providers (ISP) such as UUNet provide assurances on the availability and responsiveness of the connectivity services they offer to their enterprise customers. The availability and responsiveness can be defined in a variety of ways, the most common metrics being the delay between two access routers and the loss-rate between a pair of access routers.

In an insurance approach towards supporting service level agreements, the ISP will assure an upper bound on the delay and loss-rate between any pair of access routers, averaged across all of the access routers and over a reasonable duration of time. The ISP would install a monitoring scheme to measure the delay and loss-rate among the access-points. Common ways of measuring the performance include the use of IP pings to collect the delays between points in the network, as exemplified by the data collected by the Surveyor project (http://www.advanced.org/surveyor/) and the collection of network delays as measured by the protocol exchanges of NTP, the network time protocol. UUnet measures monthly averages of access point latencies, offering to return parts of service charges if the performance guarantees are not being satisfied [1].

As we move from the insurance approach to the provisioning approach of SLAs, networking connectivity providers would need to offer different levels of service and performance to different customers. For a networking services provider, it would mean offering a service with a lower latency and loss rate to some customers in comparison to others. The terms of SLAs are customized for different customers. In the context of network connectivity, it means that a service provider would need to provide for different latencies and/or loss-rates on the links that interconnect different customers on the same link. A customer running latency-critical applications, e.g. Voice over IP calls, on the IP network may desire a tighter assurance on the network latency than one running traditional computer applications. A monthly average for round trip delays is not likely to be useful for VoIP applications, which typically require an absolute maximum delay bound in the order of hundreds of milliseconds. In order to support the different requirements of different customers, network service providers would need to use techniques like Differentiated Services [2] or traffic engineering using MPLS [3] in order to plan the networks so that they would meet the requisite performance targets. Such provisioning of systems can be done currently using capacity planning tools and circuit establishment schemes that typically tend to be manual.

Future technologies on the horizon, e.g. dynamic bandwidth provisioning in optical networks [4] offers the ability for the carrier to do such provisioning in an automated manner. That would enable the network operator to dynamically provide more bandwidth on its circuits as traffic carried on specific segments of the network increase. This would

allow the network connective service providers to move towards the adaptive approach for providing service level agreements. Some of the algorithms to support an adaptive approach can be found in [5].

4.2 Hosting Services.

In the context of hosting services, the service provider is mostly concerned with operating a set of network connected machines that are hosting an application, e.g. a web-site or a mail-server. The main aspects of service level agreements that are provided in this context deal with the uptime and availability of the service. A typical SLA in this environment would offer the customer an assurance that service will not be down for more than a specific period of time. Generally, if the service is disrupted for more than a specific amount of time, the provider would issue credits to the customer. Some sample SLAs that correspond to the hosting environment can be seen in [9] and [10]. The uptime and availability constraints provided by the hosting service providers tend to follow the insurance model for supporting SLAs. The same level of uptime and reliability assurance is provided to all of the customers, and when unexpected events cause the performance objectives not to be met, credits are issued to the customer.

Performance based service level agreements are typically not found in the context of hosting services. This is because the performance of an application depends upon several factors that are no within the control of the hosting service provider. The hosting service provider has little control over the network latency connecting potential users to the service, or on the development of the actual application itself. As a result, the hosting service provider is not able to influence the performance of the overall application itself. As a result, many hosting services provider only offer their customer co-location services, i.e. the ability to place machines in a physical facility, and offer guarantees related to the availability of the network connectivity to their machines. Other hosting services providers could also take over the administration of the servers, and their services include the task of provisioning the appropriate configuration of machines needed to run the applications.

Co-location services are a type of hosting in which the provider is offering physical space (including power supply and network access) to its customer. In the context of co-location services, the only difference in contract that the provider can offer to different customers would be in terms of the physical bandwidth available on the access link to the Internet. The provider could obtain different physical access links for the different customers. Providers tend to follow an insurance model for such a physical access.

When more than one customer is preset at one of service provider’s premises, it is more economical to get a common access link and share it among all the different customers. The access link could be controlled by a rate control device which provides the ability to reserve different levels of bandwidth to different set of machine addresses. Such devices are available from many vendors, both small and large in the industry. The rate-control device is generally partitioned statically to provide for the different rates negotiated with

the different customers located at the service provider's premises. The static provisioning of a shared access link to provide different levels of network bandwidth access allows a service provider to support the provisioning approach for supporting SLAs.

With any shared resource, the static partitioning into shares of different customers independent of usage, is not the optimal use of the shared resource. For the best usage of the shared resource, the network resource must be partitioned dynamically so that the resource is used in the optimal manner. The rate-control algorithm used for regulating access link bandwidth can thus be crucial in influencing the right sharing of bandwidth among the different customers. Rate control algorithms which implement only a leaky-bucket model for sharing bandwidth would note be able to share the bandwidth among the different users, while rate control algorithms that allow the use of excess shared bandwidth among the heavy demand users would tend to use the bandwidth more effectively. This allows the service provider to move to an adaptive approach to support SLAs.

If the SLAs signed between the hosting services provider and the customer specify a fixed rate of access bandwidth, then the service provider would not be changing the parameters at the rate control device during run-time. However, if the performance parameters in the SLA contract specify other parameters such as the packet loss rate in the access network, then the service provider has to ensure that adequate bandwidth is allocated to meet the current demands of the customers. In that case, the service provider may want to monitor the usage of the access link and periodically update the shares allocated to different customers in order to keep the shared allocated to each customer in proportion to their current usage.

Some hosting providers offer services which manages the servers and applications of the customer. Most hosted systems tend to follow a tiered configuration, in which the application system consists of multiple tiers, each tier consisting of multiple servers performing the same task. As an example, most web-based applications tend to be implemented in three tiers, the first tier consisting of HTTP servers (e.g. Apache), the second tier consisting of application servers (e.g. IBM Websphere Application Server or tomcat), and the third tier consisting of the database server. At each tier, a load balancer may be used to distributed work among a number of different machines. The number of servers used at various tiers has a strong influence in determining the responsiveness and availability of the managed service.

Hosting service providers that offer management and provisioning of servers can use the shared resources to manage the performance and responsiveness of applications that are hosted at their sites. In this environment, the hosting service provider can follow an adaptive approach to support SLAs, dynamically modifying the number of servers provided to each of the customers in order to meet its performance objectives. However, adjusting the number of servers would require the service provider to maintain a shared pool of servers from which it would be able to carve out the appropriate configuration for each customer as needed. The creation of a shared pool of servers and providing customers free servers to access from that pool is the corner-stone of utility based

computing initiatives by hosting services providers such as HP [6] and the Oceano project [7].

4.3 Integrated Services.

Integrated services are usually offered by the IT (information technology) department of an enterprise and have the unique ability to control the hosting environment as well as the network which connects the clients to the servers.

SLAs offered by the IT department of most large corporations tend to follow an insurance approach. The IT department would offer appropriate SLAs and provision the network connectivity within the enterprise intranet and the servers running an application so as to provide the desired level of application performance.

To migrate from the insurance approach of supporting service levels to the provisioning or the reactive model, the integrated service provider has the ability to use a combination of the techniques from the networking services provider area as well as from the hosting services provider area. These techniques used for this purpose would include the task of maintaining a shared set of resources, and then trying to adapt the set of resources allocated to each application/customer in a manner best suited to support the service level agreements that are in place. A combination of the adaptation techniques within the network and ones at hosted application sites can be used.

5. Conclusions and Future Areas of work.

In this paper, we have provided an overview of the different techniques and approaches that can be used to support the notion of service level agreements in IP networks. We have identified the different types of agreements that are commonly used in IP networks, and examined the different models used in supporting the agreements in each of those contexts. Most of the industry currently tends to follow an insurance model for supporting service level agreements, although we would expect them to move towards the more dynamic schemes in the near future.

References

[1] UUNET Service Level Agreements Specifications, Available at URL http://global.mci.com/us/enterprise/customer/sla/servicessupported/index.xml.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, "An Architecture for Differentiated Services", Internet RFC 2475, December 1998. URL http://www.ietf.org/rfc/rfc2475.txt.

[3] J. Boyle, V. Gill, A. Hannan, D. Cooper, D. Awduche, B. Christian and W. Lai. "Applicability Statement for Traffic Engineering with MPLS", Internet RFC 3346, August 2002. URL http://www.ietf.org/rfc/rfc3346.txt.

[4] S. Sengupta and R. Ramamurthy, "From Network Design to Dynamic Provisioning and Restoration in Optical Cross-Connect Mesh Networks: An Architectural and Algorithmic Overview", IEEE Network Magazine, July/August 2001, pp 46-54.

[5] E. Bouillet, D. Mitra and K. Ramakrishnan, "The Structure and Management of Service Level Agreement in Networks," IEEE Journal on Selected Areas in Communications, Vol.20, NO.4, May 2002, pp 691-699.

[6] V. Turner, "HP Utility Data Center", White Paper available at http://www.hp.com/large/infrastructure/utilitycomputing/images/IDCWhitePaper.pdf

[7] K. Appleby, S. Fakhouri, L. Fong, G. Goldszmidt and M. Kalantar, "Oceano: SLA based Management of a Computing Utility", IFIP/IEEE International Symposium on Integrated Network Management, May 2001, pp. 855-868.

[8] E. Kim, J. Song, and C. Hong, "An integrated CNM architecture for multi-layer networks with simple SLA monitoring and reporting mechanism", Proceedings of IEEE Network Operations and Management Symposium, (NOMS) 2000, pp. 993 -994.

[9] Rackspace Fanatical Support Service Level Agreement, available at URL http://www.rackspace.com/infrastructure/service_levels.php.

[10] Verio Service Level Agreements, available at URL http://verio.com/about/legal/sla/.