RL Consulting

People • Process • Technology
Organization • Integration

IT Services Management

A Description of Service Level Agreements

White Paper

Prepared by:
Rick Leopoldi
May 25, 2002

Introduction

The primary objective of an organization's Information Technology (IT) department is to provide computing services that meet the business requirements of its users. To achieve this, IT needs to understand those requirements and translate them into its own business objectives. Against those objectives IT can measure the service delivered as well as its own capability and performance in providing the services. To enable this to occur, the critical services and level of delivery required must be identified and agreements reached. Typically this is done by Service Level Agreements (SLA's), which are formal written contracts developed jointly by the provider of services (IT) and its users.

SLA's between users and IT have been around for many years, sometimes resulting from an audit requirement but often filed away and almost forgotten. The users may not even remember that such a document exists and the performance of IT may not be governed by it. However, for the SLA process to work, each agreement must define the current relationships, responsibilities, and expected performance of the organization's information systems and services for its users. They are living documents; that is, they are not cast in stone because conditions driving them are frequently changing and must therefore be reviewed and kept current on an ongoing basis.

Service Level Management is much more than writing SLA's. The process, once undertaken with commitment, will change the way that IT views the user and in its turn, puts IT in the spotlight. Even from the outset, when there are only one or two SLA's in effect, there are obvious benefits to be derived from the regular communications, reports and reviews of the services. As more information is gathered on business profiles, transaction rates, response times and deadlines, areas such as capacity planning, chargeback and performance analysis are able to track and predict resource usage much more accurately. The successful implementation of Service Level Agreements will require total organizational commitment.

Why Have Service Level Agreements?

Frequently IT managers say that they know when their users are really happy by the number of phone calls they receive. No calls mean a satisfied customer since they don't call to complain about poor service. Often this has become the de-facto measure of customer service. In reality, it is typical that the IT manager understands little about their customers except for the ones who were willing to pick up the phone. In this scenario, it is difficult for IT managers to realistically

measure the general level of satisfaction or dissatisfaction with the service provided by IT.

From the customer viewpoint, their expectations of what service really is and how quickly the data center could respond, of what levels of service were reasonable, are rarely based on the business realities of managing a technology infrastructure.

This lack of understanding on both sides means that objectives and goals are not defined. As a result, the customer is dissatisfied with the service and typically has either looked elsewhere (PC's, departmental computing or outside services) or escalated the pressure on IT to deliver. Both of these approaches are costly to the enterprise. When unmanaged decentralization occurs, there is duplication of services and the benefits of the economies of scale of centralized management processes are lost. When unmanaged pressure is placed on IT, there may be an increase in costs and/or a loss in productivity as hardware and people are thrown at the problem.

So, very simply, the purpose of the SLA is to identify the shared goals and objectives of the concerned parties.

- Information Services: "We agree to provide you this level of service based on agreed-to set of guidelines."
- Customer: "We agree to abide by your guidelines in anticipation that you will provide us this level of service."

What Are Service Level Agreements?

The literature about Service Level Agreements (SLA's) defines them in a number of different ways, but most succinctly, they are agreements between the supplier of service and the user of that service that define:

1) The availability of the service to the user
2) The performance target of various components of the user's workloads
3) The bounds of guaranteed performance and availability
4) The measurement and reporting mechanisms
5) The cost of the service (where cost is an agreed to component of the SLA)

There are many different types of SLA's, ranging from the very basic broad application of performance and availability standards to precise, focused SLA's that vary from user to user in the same enterprise. For most large data processing centers, with multiple computing environments, various users with differing requirements, "mission-critical" workloads, time dependent workloads, a general-purpose SLA has little meaning and provides little benefit.

The premises behind a more complete Service Level Agreement are:

a) The data center provides multiple services (networking, online systems for both production and development, end-user direct support, imaging, management of data through multiple databases and DBMS's, etc.)
b) The using organizations have different demands for services (not all organizations need all services)
c) The using organizations have differing requirements for the same service (Dept. A needs online support from 7am-3pm, Dept. B from 7am-7pm)
d) There is value associated with faster, more desirable service (.5 second response is worth more than 10 second response, 1 second response is worth more at 11am than .5 second response at 11pm)
e) The cost of the system is shared by the users of the system; that cost is allocated according to when and how the system is used

So SLA's appropriate for a large data processing enterprise would reflect all of the above and to structure one, the following can be used as a guideline (from the excellent article "Service Level Agreements: Good Fences Make Good Neighbors" by G. W. Miller):

1) Identify the parties to the agreement - Depending on the scope of the agreement, the parties to the SLA may be groups within the same department, between major departments, or with customers outside the organization. In any case, agreements are best set out between individuals, either by name or by title, along with the physical location. All personnel involved must sign the document.
2) Describe the service to be provided - For purposes of discussion, it seems useful to define some terms:
    - Computer Resource Units (CRU's) are the basic units representing computer hardware, etc. activity. Examples would be processor time, various I/O counts, tape mounts, DASD storage, etc.
    - Computer Workload Units (CWU's) are the basic workload elements within the computing environment. Examples are background jobs and interactive transactions.
    - Business Work Units (BWU's) are the business related activities that are the purpose of all this computer activity. Examples are bank deposits, claims received, units manufactured or passengers boarded.

    The user of IT service may wish to describe activity in terms of BWU's of passengers boarded, etc. The provider of the service sees the activity generated by the conduct of the users business in terms of a number of the CWU's of jobs, transactions or messages required to accomplish a passenger boarding. These computer related work units can in turn be translated into

CRU's of processor time, peripheral usage, etc. The CWU's provide the necessary bridge between work as seen by the customer and demand as seen by the provider.

3) Specify the volume of demand for service, over time - It is not reasonable to expect that the provider of a service be required to agree to a specific quality of service without knowing the quantity of service which is to be demanded.

   Both the expected volume and the time distribution of that volume must be considered and specified. Requirements should specify the average and peak rates at which this work will enter the system and the times during which this demand is expected to occur. The user could be provided with the incentive of better service at a reduced cost by avoiding peak usage periods.
   It should be clear that the volume of demand is as much a part of the binding SLA as any other element. The quality or timeliness of service is agreed to be provided for a certain volume of demand at a certain time of day. If the actual volume of demand exceeds these estimates, then it is very appropriate to describe a reduced requirement for quality.

   In addition, the customer should be required to define the expectation for the volume of service to be required over some significant period of time into the future. It is not unreasonable to ask for an estimate of volume of demand by month for the next 24 months. The provider cannot be expected to provide a high level of quality service over an indefinite period of time without knowing what is coming next. One of the most tangible benefits to be derived from the development of comprehensive SLA's is the information on expected workload growth that results.

4) Define the timeliness requirements for the service - The timeliness metric defined in the SLA will depend on the nature of the work. Discussion with the customer and analysis by the provider will have determined the number and kinds of Computer Workload Units (CWU's) that will be required to support the customer's Business Work Units. These CWU's may be background jobs or on-line transactions or input messages. Timeliness will probably be defined in terms of the CWU's that is in terms of background turnaround time or transaction or message response time.

   Job turnaround time may be stated as job start to job stop or job submitted to job complete. For applications development and user submitted, non-scheduled production, distributions of the form "90% within 2 hours" are appropriate. Specific times will vary by class. Additionally, for production jobs, some times outside the host may be needed, e.g., input available to job submit, or print

complete to deliver to the user. Scheduled production jobs may define service in terms of whether the scheduled completion time has been met, e.g., if input is available by 8:00pm, then output will be delivered to the user by 8:00am, etc.

Transaction or input message response time as perceived by the user is generally the time from when the enter key is pressed until the keyboard unlocks and a response is received. Measurement of times outside the host processor present measurement challenges although a variety of tools and techniques exist for measuring or estimating these times.

Transactions with consistent characteristics are adequately described by a "90% within 'X' seconds" distribution. Some transactions, while consistent in their characteristics, may be highly data-dependent. In such cases, it may be necessary to link response times to the amount of data being processed and define response time limits that relate to ranges of volume.

5) Specify the availability of the service required - As an element of a SLA, availability is that time window during which service is to be provided to the customer. A frequently complicating factor is that a system may be up and accessible, but operating in a degraded manner. The customer's notion of availability is that if work is going in and coming out at the usual rate, the system is available; otherwise, it is not. In fact, availability is the product of a series of system components, any one of which can affect end user availability. These components include:
   - network equipment
   - front end processors
   - network control software
   - workstations and terminals
   - servers
   - mainframe processors
   - DASD
   - I/O devices
   - operating systems software
   - operational support software
   - applications programs
   - individual databases

   For each of these there is the possibility that a failure will result in loss of availability to at least some users. Obviously the probability and impact of component failure will vary with component. In the event that significant outages do occur, consideration must be given to the most appropriate use of the available service. The SLA can establish the fallback priority and method for

a particular application or user. Depending on the resource available and the needs of the user, this priority might be part of a pricing discussion.

6) Define the reliability of the service provided - There is often some confusion as to just what is meant by the terms availability and reliability. Reliability has to do with the comparison of actual availability to planned availability. There needs to be an acceptable definition of the conditions that accurately indicate whether or not the system is or is not fully available. Then reliability might most simply be stated as the ratio of the time that the system was actually available to the time during which the system was planned to be available. We would probably state this as a percent, and set a standard for an acceptable degree of reliability.

7) Quantify the compensation for providing the service - If Service Level Agreements are expected to be taken seriously, they should have some teeth. Cost allocation activities that go to great lengths to determine a customer's share of the cost of the installation and then report that cost for the customer's information only are relatively worthless. High-level summaries reported to corporate management can get some temporary attention, but this is a crude way to manage costs. The way to keep everyone honest requires a real chargeback system. Chargeback is a procedure for not just attributing but billing the cost of providing computer services to the users of those services.

   As part of the chargeback system used, certain options can assist the SLA efforts. Relating appropriate charges to both the volume and quality of that service gives the user an incentive to utilize service as effectively as possible and gives the provider an incentive to improve quality. For this reason it is possible to tie service level costing to both volume and service. If the user exceeds the contracted volume of data processed, he should pay a higher cost and if the user is provided with a level of service that is less than that for which he contracted, he is entitled to pay a lower cost. Since the administration of SLA's should be kept as simple as possible, this refinement of chargeback should be adopted after initial SLA's are functioning properly.

8) Describe the measurement procedures to be used - Comprehensive monitoring procedures are central to the establishment and maintenance of SLA's and to the effective management of the data center as a whole. Proper monitoring procedures begin with the means for routinely obtaining installation performance information. Sources of such information include SMF files, RMF data, IMS logs, hardware monitors, network monitors, and manual logs. There are many tools such as the MICS Accounting Module or its equivalent that can be used to process this information. IT must do the measurements!

The large volume of this very detailed information requires it be summarized for historical purposes. This summarized history is one of the single most valuable resources possessed by IT. Analysis of this information makes possible:

- The intelligent establishment of achievable Service Level Agreements
- The accurate assessment of charges to the user
- The performance tuning information required to ensure the environment makes the maximum resources available
- The historical trending that allows reliable forecasting
- The detailed modeling information required to support hardware planning

The importance of this historical information to the corporation implies that the primary tapping of the raw sources and their initial summarization should be a consolidated activity rather than a separate effort for each area of interest. There should be a Performance Data Base to serve as a source for this information by those requiring it. The consolidation of these various activities requires a high degree of coordination to ensure that all required information is present in the master performance data base and that maximum advantage is taken of new tools and technologies in the field of performance monitoring.

9) Set a date for renegotiation of the agreement - Under normal circumstances, Service Level Agreements should be renegotiated at least annually as a part of the budget cycle. As a practical matter, they may need to be reviewed more often due to SLA's being significantly out of bounds for several months in a row or if significant changes have occurred in the host environment.

What Do SLA's Require?

In other words, some of what SLA's require are:

- A signed contract with each user
- On-line response times by application
- Job turnaround by class
- System availability by application
- Accuracy limits
- Load of each application by transactions per hour and jobs by day
- A plan for reporting SLA results
- Priorities if service cannot be delivered
- Penalties if the user exceeds the load
- Penalties if IT does not provide service as agreed
- A schedule for follow-up meetings and interface

Getting Started

In developing a plan to start SLA's, the following steps are recommended:

1) Develop Service Level Objectives (SLO's)
2) Select critical, high visibility, well understood applications
3) Target only things that can be measured
4) Examine historical trends and operating requirements
5) Obtain executive commitment
6) Educate User on SLA's
7) Observe User work environment
8) Develop negotiation strategy
9) Negotiate a win/win contract (SLA)
10) Adjust SLO's
11) Obtain executive signatures:

    IT executive commits to:

    - Monitoring Service Levels
    - Educating staff to take corrective actions
    - Reviewing potential impact of tuning decisions/operating changes
    - Reporting performance clearly and accurately
    - Taking actions to avoid service degradation

    Customer executive commits to:

    - Accurately forecasting workload
    - Accurately representing service requirements
    - Changing operating procedures to observe restrictions
    - Providing feedback on reports
    - Providing hardware upgrade justification

12) Monitor and report