SLA Management Handbook

Volume 4

Enterprise Perspective

TeleManagement FORUM™

THE Open GROUP
Making standards work™

October 2004

SLA Management Handbook – Volume 4: Enterprise Perspective

ISBN: 1-931624-51-8

Document Number: G045

This document is equivalent to TMF document reference GB917, Version 2.0, Volume 4.

Published by The Open Group, October 2004.

Comments relating to the material contained in this document may be submitted to:

The Open Group
Thames Tower
37-45 Station Road
Reading
Berkshire, RG1 1LX
United Kingdom

or by electronic mail to:

ogspecs@opengroup.org

Foreword

The use of Service Level Agreements (SLAs) between telecommunications service providers and their customers has been common for some time for services such as Frame Relay and ATM. TeleManagement Forum recognized this trend and developed the SLA Handbook (GB 917) in response.

The Open Group Quality of Service (QoS) Task Force also recognized that there was an increasing need and move towards providing SLAs for higher-level services, such as IP and security. The QoS Task Force membership is comprised of vendors, enterprise end users, as well as service providers, and clearly showed that SLAs are being considered with both internal and external service providers. The trend towards outsourcing also promotes SLAs as a means of specifying and verifying the agreed level of performance of the outsourced service.

The Open Group commissioned a survey of enterprises with Sage Research to further understand their perceived SLA issues. Over 150 enterprises in the USA were polled and follow-on, more detailed interviews followed to further clarify the results. A high percentage (60%) the respondents indicated they have internal SLAs in place, and a further 20% would be adding them. The most common reasons for adopting SLAs by the enterprise were: to allow better planning (69%), help to reduce or control costs (61%), to allow the enterprise to offer SLAs to internal customers (55%), and, perhaps as importantly, to convince management to try new services and applications (52%).

As expected, the percentage of respondents either had SLAs, or planned to add SLAs for traditional telco services like Frame Relay (81%) and ATM (70%). However, these were similar to those employing or about to deploy SLAs for enterprise technologies like Voice over IP (VoIP) (77%) and outsourced facilities like video conferencing (63%), although the penetration was smaller.

The follow-up interviews suggested that there is a wide gulf between what the standards bodies are thinking about in terms of technologies to ensure quality of service and what is known and used by many IT managers responsible for running networks at the enterprise level. Perhaps the most interesting example is the marking of traffic. Technologies under developed or being deployed – like ATM or DiffServ – allow the marking of traffic on a per-application or traffic type; yet without exception, the survey respondents in the follow-up interviews used marking as a means of identifying critical traffic and putting them on dedicated servers or networks.

With increasing reliance on e-Commerce, it is perhaps not surprising that the most likely entity for which an enterprise may provide an SLA is its customers (63% presently, with a further 24% planned).

The survey clearly identified that there is a pressing need to include further discussion on the needs of the enterprise in the development and administration of SLAs. Both internal and external customers need to be considered and services other than communications networks which must include higher-level services (such as applications) and support functions (such as help desk).

In 2002, TeleManagement Forum began revising its SLA Handbook and wanted to consider true end-to-end SLA. The earlier documents focused on delivery of service to the customer's premises and, as such, were 'edge-to-edge' rather than end-to-end delivery. It was clear that in consideration of a true end-to-end SLA, The Open Group, with its wide enterprise focus, and TeleManagement Forum, with its service provider focus, could and should collaborate. This document represents a joint initiative between The Open Group QoS Task Force and TM Forum's SLA Management Team. The QoS Task Force is responsible for writing this document, with feedback from TM Forum. This document will become Volume 4 of TM Forum's SLA Handbook and is focused on SLA requirements as driven by enterprise application requirements.

In May 2003, The Open Group initiated the Application Quality and Resource Management (AQRM) initiative to investigate application quality metrics and resource management standards and interfaces for distributed and stand-alone systems. This work will have direct bearing on this volume.

Contents

Tables

Equations

Figures

Preface

About The Open Group

The Open Group is a vendor-neutral and technology-neutral consortium, whose vision of Boundaryless Information Flow will enable access to integrated information within and between enterprises based on open standards and global interoperability. The Open Group works with customers, suppliers, consortia, and other standards bodies. Its role is to capture, understand, and address current and emerging requirements, establish policies, and share best practices; to facilitate interoperability, develop consensus, and evolve and integrate specifications and Open Source technologies; to offer a comprehensive set of services to enhance the operational efficiency of consortia; and to operate the industry's premier certification service, including UNIX certification.

Further information on The Open Group is available at www.opengroup.org.

With its proven certification methodology and conformance testing expertise, The Open Group is the international guarantor of the interoperability that single economic entities require to achieve independence.

The flexible structure of membership of The Open Group allows for almost any size of organization to join and influence the future of the IT world, and the introduction of membership for individuals is currently being considered. However, members include some of the largest and most influential organizations in the world and buy-side members have combined budgets of over $50 billion per annum.

About TeleManagement Forum

TeleManagement Forum is an international consortium of communications service providers and their suppliers. Its mission is to help service providers and network operators automate their business processes in a cost and time-effective way. Specifically, the work of the TM Forum includes:

- Establishing operational guidance on the shape of business processes
- Agreeing on information that needs to flow from one process activity to another
- Identifying a realistic systems environment to support the interconnection of operational support systems
- Enabling the development of a market and real products for integrating and automating telecom operations processes

The members of TM Forum include service providers, network operators, and suppliers of equipment and software to the communications industry. With that combination of buyers and suppliers of operational support systems, TM Forum is able to achieve results in a pragmatic way that leads to product offerings (from member companies) as well as paper specifications.

About this Document

This document is Volume 4 (Enterprise Perspective) of the SLA Management Handbook. It has been developed and approved by The Open Group in association with TeleManagement Forum.

Relationship to Other Volumes of the SLA Management Handbook

Volume 1 of the SLA Handbook represents the Executive Summary for the whole four-volume set. Volume 2 outlines Service Provider SLA to its customers and other Service Providers. Examples of these services are discussed in Volume 3. Volume 4 (this volume) outlines the issues for enterprise SLAs, together with business applications that may use the Service Provider telecommunications services discussed in Volumes 2 and 3. Therefore, overlaying the structure of the volumes onto the figure provides:

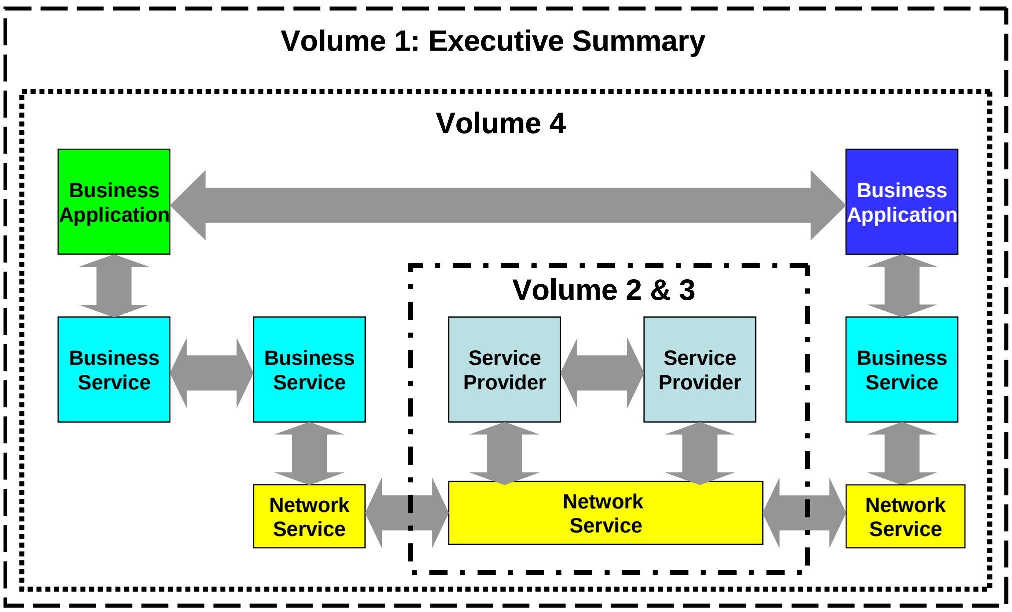


Figure 1: Relationship between Volumes 2, 3, and 4 of the SLA Management Handbook

The relevance and scope of services in the Service Provider network (and SAP) will be discussed within this volume, but more detailed discussion of the services and supporting services can be found in Volumes 2 and 3.

Assumptions

Enterprise Business Processes

Many enterprise organizations adhere or aspire to business processes for the furtherance of their business goals or objectives. Adopting best practices and learning from experience are important aspects of these processes to assure success. Common processes are:

- ISO 9000 2000
- ITIL
- Capability Maturity Model (CMM)
- eTOM/NGOSS (ICSP)

This volume will not make any attempt to pass judgment as to whether the existence of a process model or any particular model is superior or inferior to any other. The volume does, however, assume that such a model exists and proposes a possible process that can be incorporated into existing processes. As each model is different and tailored to an individual enterprise, a number of common assumptions are made of any such process. These assumptions are:

- The process is documented in electronic or paper format.
- The process covers all aspects of the business, including, but not restricted to, product or service lifecycles.
- A management structure for business decisions is in place with milestone decision points within the process.
- There is a requirement to document instances of the process for audit trail and to learn from experience.
- The process itself is subject to continual review to allow for changing business circumstances and past experience.

This Handbook addresses SLA parameters that support an informal or formal contract between two parties. There are numerous business parameters contained in a contract that are not covered by this Handbook. For example, while the rebate process is described in terms of the parameters that may drive it, parameters for payment and terms are not included.

Parameter metrics will be introduced, but the exact value of the parameters will not be included as they will depend on the environment and available service, and will also change with time as new techniques and technology improve service offerings.

Underlying systems and processes tend to be proprietary and will not be covered. The customer and SP(s) often choose to be free to develop their own internal processes for market competitiveness. Therefore, the Handbook provides the information to both parties that is necessary to develop and manage their internal systems in a more efficient manner.

This Handbook is designed to be a reference for both enterprises and SPs as they develop new services and to assist in customer discussions. It will inform and educate about the intricacies of SLAs, as well as provide a structure for defining agreements such that relevant SLA parameters and values can be agreed that are achievable and verifiable.

The Handbook is not intended to provide a cookbook of tables to be filled out without knowledge of the underlying requirements, capabilities, or limitations of the technology or service(s).

Executive Summary

This volume addresses the enterprise issues in the provision of end-to-end Service Level Agreements (SLAs), and is a collaboration between The Open Group, representing the enterprise, and TeleManagement Forum, addressing the service provider markets. This work was further inspired by survey data gathered on behalf of The Open Group by Sage Research which indicated great interest in SLAs in the enterprise, but a large gap between where enterprises are considering SLAs and where standards bodies, such as the IETF, are currently concentrating their efforts.

The scope of the market addressed by “enterprise” is very broad with diverse business practices. It was therefore necessary to generalize the applications used by an enterprise so that SLA metrics could be applied, measured, and reported in a contractual manner.

In the attempt to describe enterprises generically, an application is considered as a collection of applications that fulfil the enterprise objectives. Such objectives would include raising revenue (commercial), giving taxes (government), curing patients (healthcare) and defending a nation (operational defense). In support of these objectives the enterprise needs to deploy a number of business applications. These applications would include call center, trading, information delivery, manufacture, distribution, etc.

To enable the business applications, a number of business services – e.g., voice, video-conferencing, email, transaction processing, and telemetry and network services (IP, Ethernet, DSL, etc.) – need to be deployed to an agreed level (the service level), maintained, and reported. There are some cases where a hierarchical or peer-to-peer relationship between these services must be considered.

This work uses the concept of key quality and performance indicators (KQI/KPI) developed by TM Forum’s Wireless Services Measurement Handbook (GB 923). The importance of the KQI/KPI concept is that it allows the provider of the service to concentrate on the quality rather than the performance of a service, as in the past. The relationship between the KQIs could be identical for the simple case or complex. It may be derived empirically or inferred from a number of KPI and other data. The mapping between the KQI and KPI forms an important part of the SLA negotiation.

For each of the generic business services discussed, the KQIs are determined and then KPIs for the services and methods are tabulated.

The form of an SLA is discussed and special attention is paid to the form of an SLA between internal parties and external parties, especially in terms of penalties. A monitored and reporting process is then discussed allowing for real time, near-real time, and historical reporting of both asynchronous events and polled parameters.

A number of use cases are considered to validate the approach. The first is a common example where VoIP is used to connect remote sites of an enterprise to a corporate HQ. Data is also supported, but only on a best effort basis. The second scenario is from The Open Group Mobile Management Forum’s work on the “Executive on the Move”[1] where an executive is considered to have voice and data

1 “Executive on the Move”, Business Scenario, published by The Open Group, July 2002 (K021).

connectivity wherever they are – in the office, in transit (car, airplane), at home, or at a remote site. Voice, data, and video (conferencing) are also supported for the executive. The final scenario is a naval application where voice, data, and video applications are supported in highly distributed and arduous environments. The VoIP and naval scenarios envisage a common IP infrastructure and use a differentiated services (DiffServ) marking scheme to separate the different service domains for prioritization.

Trademarks

Boundaryless Information Flow™ is a trademark and UNIX® and The Open Group® are registered trademarks of The Open Group in the United States and other countries.

ITIL® is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.

The Open Group acknowledges that there may be other brand, company, and product names used in this document that may be covered by trademark protection and advises the reader to verify them independently.

Acknowledgements

The Open Group and TeleManagement Forum would like to thank the following individuals for contributing their time and expertise to the production of the SLA Management Handbook:

- Gregory Bain. SIAC
- Jay Dia, Box Concepts
- Neil Davies, U4EA Technologies
- Jean Hammond, JPH Associates (The Open Group QoS Chair)
- Sally Long, (The Open Group QoS Director)
- Jon Sapiera, JDS Consulting
- Malcolm J Sinton, Qinetiq
- Deborah L. Smith, Mitre Corporation
- Tobey Trygar, Telcordia
- Lightsey Wallace, Lightsey Enterprises (Team Leader)

with particular thanks to

- David Banes, TenTen Communications (Document Editor)

A number of members and contributors from both The Open Group and TeleManagement Forum have provided input and comments to the work of the team.

1 Introduction

Service Level Agreements (SLAs) have been a common product in support of services offered by telecommunications service providers for many years. SLAs are now being considered for non-communications (network) services and are being adopted both internally and externally to define the agreed performance and quality of the service or product and as an important part of a customer relationship management (CRM) program.

To achieve quality and performance targets for the products or services may require the enterprise to establish and manage a number of SLAs. The growing complexity of global services brings together a myriad of services, suppliers, and technologies, all with potentially different performance requirements.

For a business to perform its intended task, whether it be offering a service or product as a utility (e.g., electricity, water) or as a revenue generating company responsible to its shareholders, expectations of the level of service to be offered will be defined between the source provider (the business, enterprise, or department) and the ultimate end user (the customer). This simple relationship is shown schematically in Figure 2:

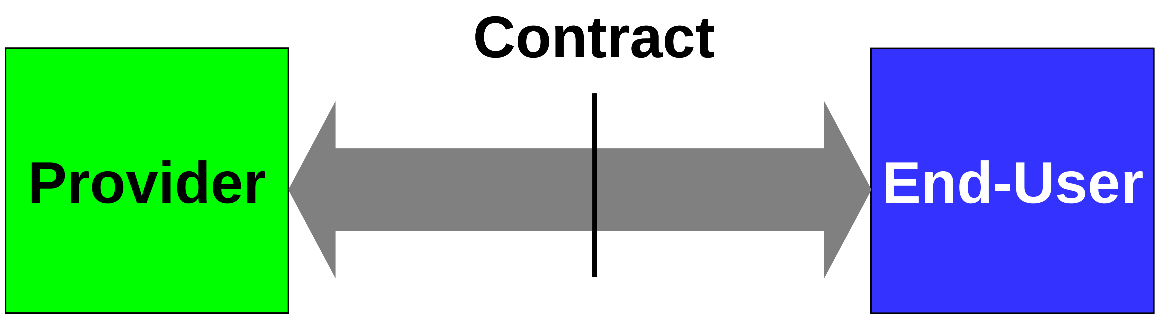


Figure 2: Schematic Representation of the Business Relationship between Provider and User Governed by a (Sales or Service) Contract

The contract between the provider and end-user services or goods may also constitute an SLA, although not necessarily so. The contract would be for the sale, supply, or support of either single or multiple services or products.

The goal of the enterprise SLA process is ultimately to improve the Quality of Experience (QoE) of the service or product to the enterprise clients, whether they are internal or external to the organization. QoE is a collective term to form a measure of the quality of a service or product and includes all aspects of service: its performance, level of customer satisfaction in the total experience, pre and post sales, and the delivery of its products and services. Determining the QoE provides a discriminator between various types of service or product that an enterprise provides, and leads to opportunities to balance the level of quality offered against price and customer expectation.

In support of enterprise or business applications (BA), business services (BS) are available to facilitate the application. For example, in a call center (the application), an obvious business service is voice communications. Business services in themselves do not raise revenue, but they are crucial in achieving business objectives and the effectiveness of a business application. One or more business services may be necessary to support a business application. To support a particular business service, a number of service resources will be required to fulfil the requirement.

In turn, business services are supported by network services (NS). Network services may be either supported internally or outsourced to external providers. It is likely that any enterprise will have to consider both internal and external network resources. External network services may be supplied by an information or communication service provider (ICSP) in their entirety or may provide transport methods (Frame Relay, Leased Line, ATM, SONET/SDH) for which an overlay (e.g., IP) may be administered by an internal enterprise. This hierarchy between enterprise business application, business service, and network service is shown in Figure 3:

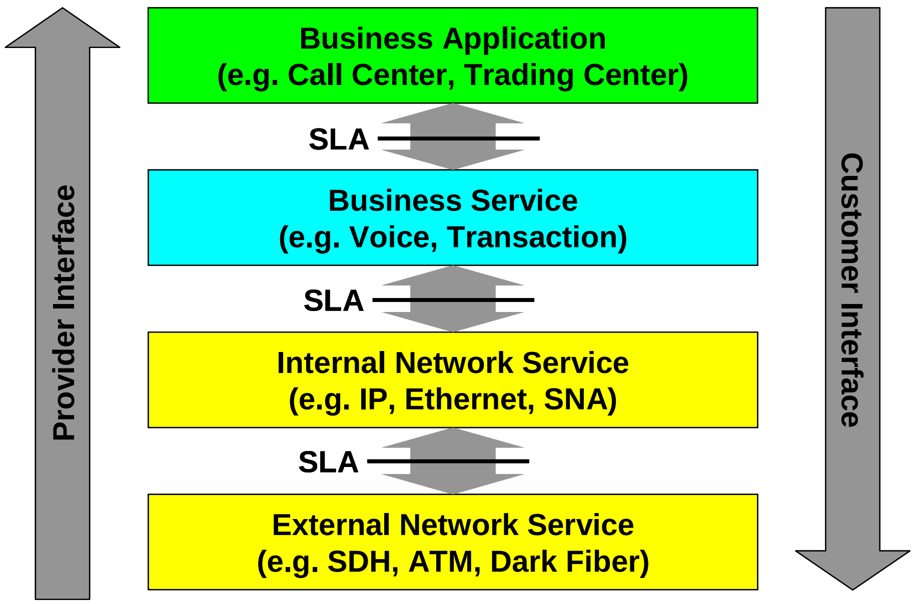


Figure 3: Hierachy of Business Application, Business Service, and Network Services in the Enterprise

Each layer in the hierarchy has scope for an SLA between the lower layer (the service) and the higher layer (the application). Each layer represents a different terminology, management, and even contractual domain. In some applications, some layers will be absent and the exact structure will depend on the size of enterprise and on other factors including budgets, manpower, skill, and training level of users and support staff.

For the contract to be achieved by the provider, a number of SLAs have to be established and monitored. These may be internal (business or end user) or external (e.g., service providers, partners, outsourcers, or suppliers). SLAs between businesses, end users (other business or residential users), and network service providers (commonly known as service providers) are the subject of the earlier volumes of this Handbook and, notwithstanding the existence and requirement for the other volumes, are better understood and synonymous with the term SLA. It should also be considered that some of these services may not involve a network at all (e.g., retrieving a part from a warehouse) and that some of the SLAs may require only internal services or only external services.

As a result, what may initially be regarded as a simple contract or understanding between a customer and supplier may be multi-faceted and complex. It is appropriate to make the SLA between a supplier and customer as simple as possible yet complete enough to remove ambiguity and to understand what happens when the level of service defined in the SLA falls or exceeds expectations. As with any contract, they are really written to define how you will determine whether your SLA is being met and also what will happen when it isn't.

In general there will be a tier of SLAs as the business function is satisfied with different service offerings – for example, the LAN and WAN – and in some instances there will be peer services where the service requires the co-operation of other services of the same type; for example, in a parts retrieval system, parts may be distributed across a number of warehouses and as such for any one warehouse to supply all parts to its customer, may require parts to be retrieved from other warehouses. To implement delivery of a service or to provide performance monitoring for a service, service resources are required, which may include personnel, hardware, software, and appropriate licenses.

For every tier or peer SLA, one party can be considered a supplier or provider and the other a customer. Simplicity and clarity are also key in these contracts, but as there is always a close business or technology relationship between the parties, the appropriate language and terminology can be used to provide the simplicity and clarity without any unnecessary explanation. It is not necessary for the SLA to have the same clarity and simplicity to other than adjacent tiers or peers. This complexity is shown in Figure 4 as an expanded view of Figure 2.

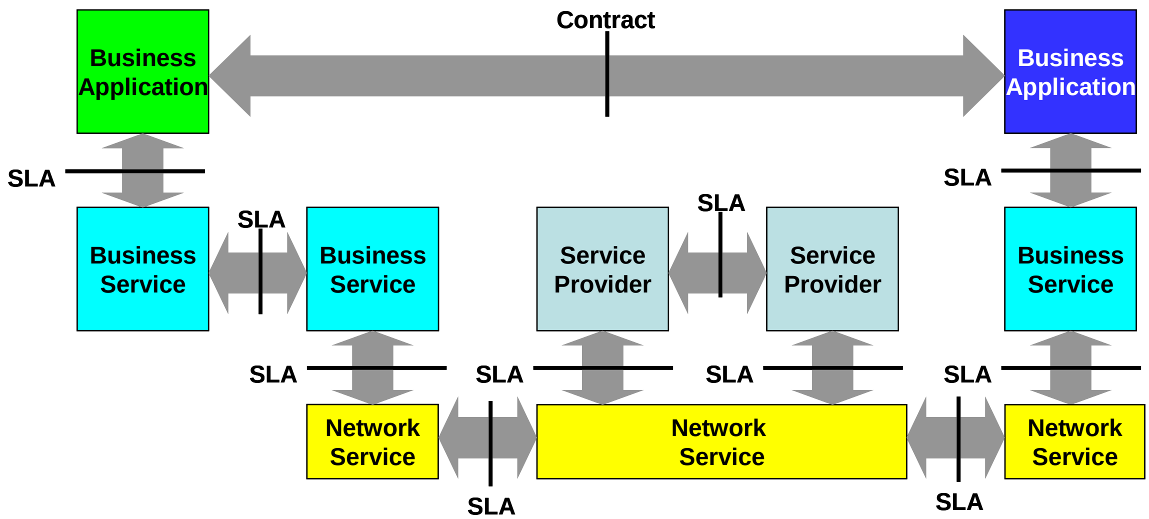


Figure 4: Expanded View of SLA in Achieving the End-to-End SLA

Important issues to note in Figure 4 are that business applications (e.g., call center, online stockbroker) are supported by a number of business services (e.g., voice, database, email) that in turn are supported by network services (e.g., IP, ATM). There may be internal network services (MIS) or external network services (service providers). Some business services – for example, shipping – do not require network services to fulfil their business application (parcel delivery) but increasingly rely upon them to provide value added services; for instance, online parcel tracking.

For an SLA to add value in providing the business application, the system must be instrumented adequately so that metrics can be determined to ensure conformance, prevent or warn of non-conformance, and measure non-conformance. An audit log may also be necessary for capacity planning, cost control, and dispute resolution.

An SLA is an understanding of expectation of service. For this to be reasonable, the system design and requirements that the business function mandates must be clearly understood. If the requirements of the system to fulfil the business function are unknown, it is unlikely that the SLA can be adequately monitored to ensure conformance against the business function. Furthermore, if the system design is not well understood or capable of fulfilling the SLA, the terms and conditions have no meaning.

In summary, for an enterprise SLA to have meaning and achieve its aims, the delivered service must be:

- Achievable. The requirements and design must be understood and documented adequately to ensure that the requirements of the service can be met.
- Defined. The SLA must be definable in terms appropriate to the customer and provider.
- Instrumented. The system must have the capability to be monitored to ensure conformance, warn and prevent non-conformance, and measure non-conformance.
- Enforceable. Penalties for non-conformance and rewards for high performance should be defined to act as an incentive.

The SLA should be an integral part of the business planning and execution process, without it the business objectives are unlikely to be achieved, ongoing performance cannot be determined, corrective action taken, or future plans made.

1.1 Enterprise Complexity

Enterprises as a collective are organizations that adapt business models to new opportunities and technologies as appropriate with product lifecycles typically measured in months or less. In contrast, service providers are in general highly regulated organizations – especially the traditional voice telecommunications providers, the national PTTs – and have sometimes been mandated to measure product lifecycles in decades. Telecommunications service providers are often required to make special considerations for harsh environmental conditions, redundancy, and resilience that are not considered mandatory within the enterprise.

As a result of these very different regulatory environments and the history associated with them, two distinct sets of protocols emerged, one developed for the service provider (for example, Frame Relay, ISDN, and SONET/SDH), and the other targeted at the enterprise (for example, Ethernet, IP, and SMTP). Enterprise protocols were developed opportunistically and exploited in the same manner. Those enjoying widespread adoption gained from economies of scale that lowered costs of equipment and services based on the technology, further increasing adoption. Service provider protocols and equipment, on the other hand, were meticulous, robust, and the equipment much more costly.

The net result is that the enterprise ‘enjoys’ an ever-changing selection of technology that is continually increasing in performance, robustness, and lowering cost. These opportunities and cost savings have not been lost on the service provider and some of the technology is now available to the service provider. Through deregulation, the service provider has also been able to move into the enterprise as a managed service provider (MSP) managing all the way to the desktop and even the application or as an outsource supplier. A protocol hierarchy to illustrate the resultant complexity is shown in Figure 5.

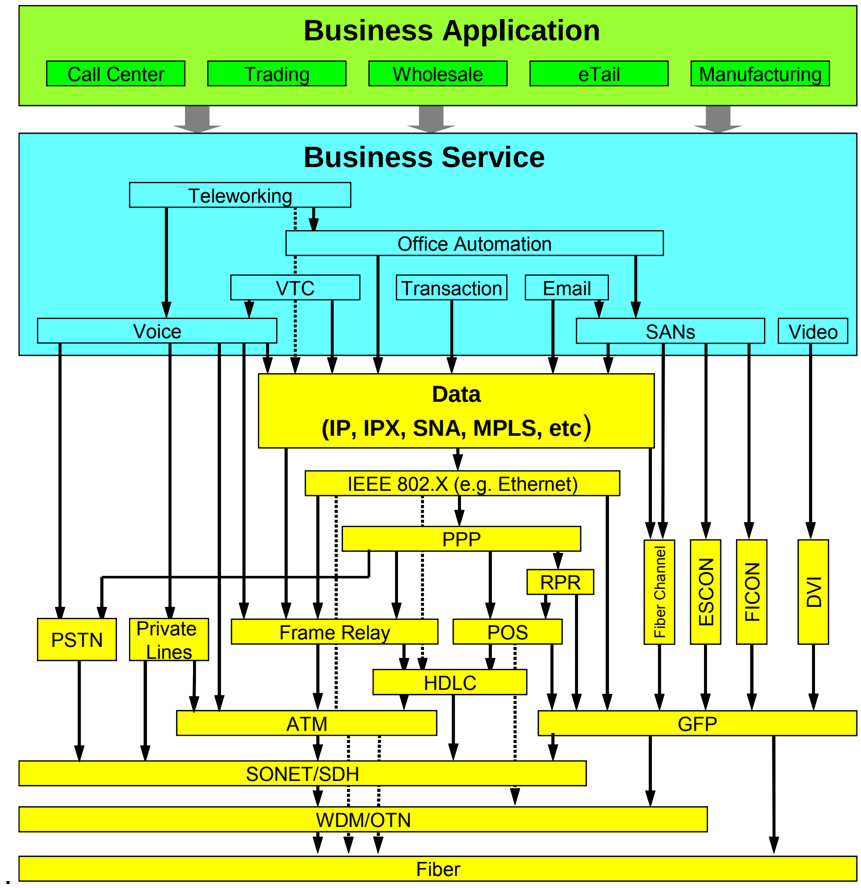


Figure 5: Enterprise Protocol Stack

This discussion illustrates that:

- There are a number of important business services required for the successful operation of an enterprise, perhaps the most important being voice.
- There are a number of protocols available which support legacy, future legacy, and current applications and services.
- There are relatively few, clear regulatory boundaries between the enterprise and Service Provider; demarcation is generally governed by economics.

1.2 Service Resources

To implement or monitor any service or products requires the application of service functions and resources in a relationship shown in Figure 6. Service functions allow the service to be physically implemented and can be decomposed into three main functional areas:

1. Primary Functions: The primary service; e.g., ATM, Email, etc.
2. Enabling Functions: Allows the primary function to be implemented; examples include: operating system, heating, ventilation, and air conditioning (HVAC)

3. Support Functions: In support of the primary and enabling functions; e.g., accounts, help desk, operations, administrations, and maintenance

To implement the service functions requires the assignment of service resources. Service resources may include:

- Hardware
- Software
- Personnel, including training
- Licenses and Intellectual Property

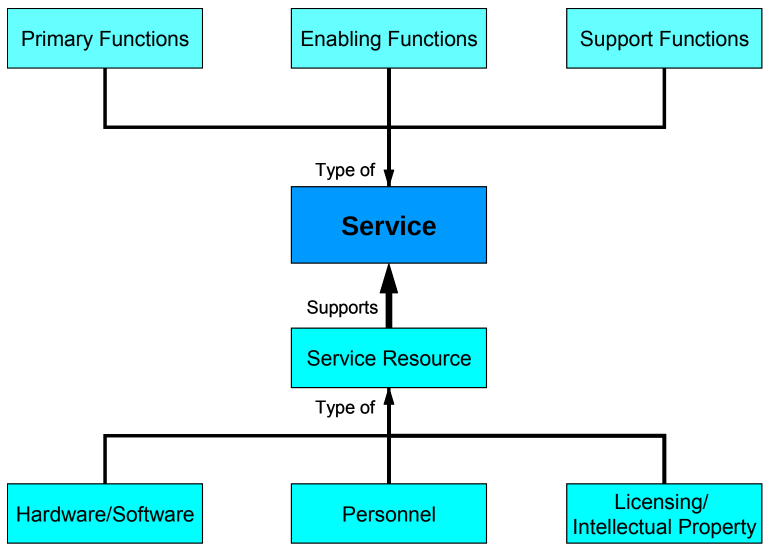


Figure 6: Service Functions and Resources Supporting a Service

Other issues may be considered as service resources including facilities (premises) and budget.

1.3 Key Quality and Performance Indicators

There is a difficulty in mapping service-specific parameters to technology-specific parameters that are more easily measured and reported. As a consequence, traditional SLAs have focused almost solely on the performance of the supporting service.

However, an accelerating trend towards service-focused management has led to the requirement of a new 'breed' of indicators that focus on service quality rather than network performance. These concepts were introduced to TeleManagement Forum by the Wireless Services Handbook (GB 923). These new indicators or Key Quality and Performance Indicators (KQI/KPI) provide a measurement of a specific aspect of the performance of the application or service. The KQI is derived from a number of sources, including performance metrics of the service or underlying support services as KPI. As a service or application is supported by a number of service elements, a number of different KPI may need to be determined to calculate a particular KQI. The mapping between the KPI and KQI may be simple or complex, empirical or formal.

1.4 Quality Concept and Mapping

The relationship between QoE and SLA is that QoE relates to the perception of the quality of the product or service, whereas SLA refers to the definition, measurement, and reporting of objective measures of the service or product. As such, QoE and the SLA are related in that if the perception of the service or product is poor yet the service parameters fall within the limits defined by the SLA, the SLA must be redressed. The key concept is to map the perceptive measures from the QoE into objective measures for the SLA. This mapping may be multi-dimensional, empirical, functional, or complex in nature.

Being subjective, KQI parameters can be hard to include as a contractual requirement; terms like “80% of respondents should respond Satisfied or Higher”. However, there will be a number of Key Quality Indicators (KQI) that relate to the QoE that should be included in the SLA. To achieve the KQI, a number of Key Performance Indicators (KPI) must also be defined, measured, and agreed in the SLA which may have a simple or complex mapping to determine one or more KQI in a relationship as shown in Figure 7.

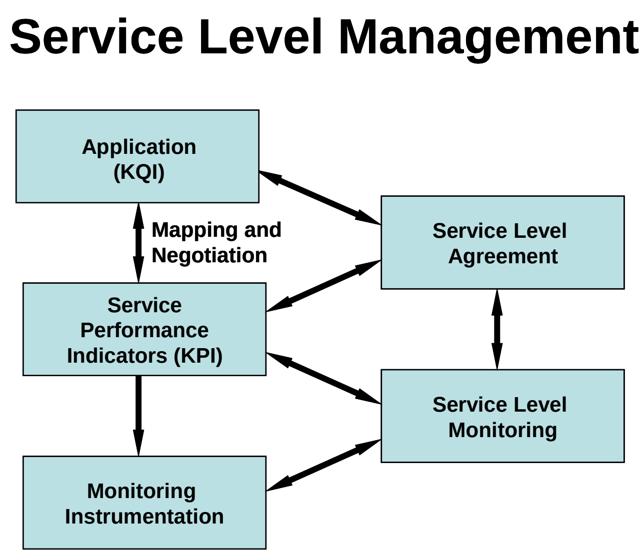


Figure 7: KQI , KPI, and SLA Relationship

The challenge with enterprise SLAs is translating QoE metrics for business applications into measurable parameters for business applications, business services, and network service providers (NSP) that can be defined and reported against an SLA and monitored under Service Level Management.

The principal concept is to agree and define Key Performance Indicators (KPI) for the service or product that collectively can be amalgamated to form a Key Quality Indicator (KQI) through a simple or complex relationship. The relationship between KPI, KQI, SLA, and the SAP is shown in Figure 8.

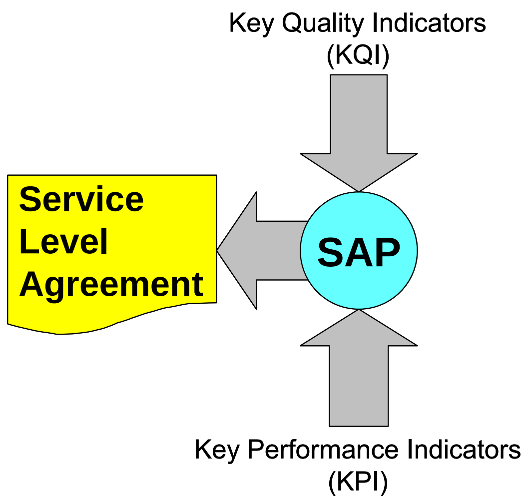


Figure 8: Relationship between SAP, KQI, KPI, and SLA

At each stage of the hierarchy, there will be a number of KPIs that create dependencies on lower-tier or higher-tier KPIs.

The KPIs can then be defined in the SLA and the KQI derived and monitored during the SLM process. It should be noted that 'Network Availability' is a KQI for the network service and as such requires KPIs from the supporting network service to allow the KQI for higher-level services (such as voice) to be determined.

Each KPI or KQI will have a lower and upper warning threshold and lower and upper error threshold as shown in Figure 9.

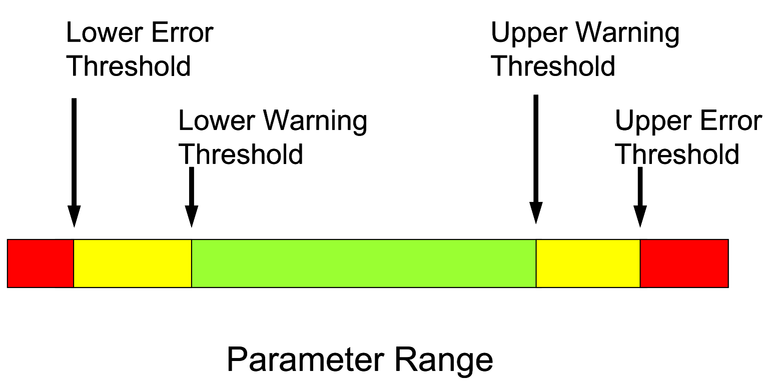


Figure 9: KPI/KQI Parameter Thresholds

The KPIs are then combined by some empirical or theoretical function to lead to a measure of KQI as illustrated in Figure 10.

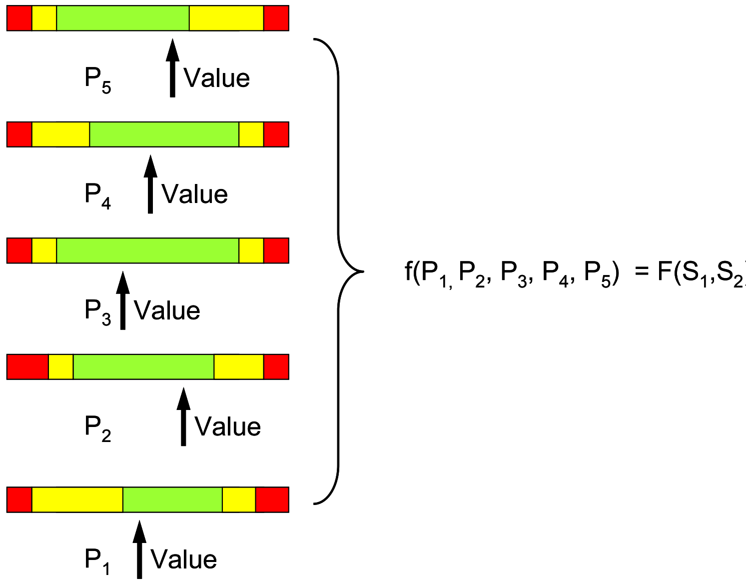


Figure 10: Combing KPIs to Determine a KQI

The importance of the warning thresholds from Figure 9 can be seen in Figure 10 as in some instances a single indicator in the warning zone may indicate that an SLA threshold may be violated for the particular KPI; a collection of such KPI in the same state may indicate a violation of a KQI threshold.

The exact form of the function linking KPI to KQI is an important concept for SLA negotiation. There has been considerable work in determining the functional relationship between KPI and KQI by standards bodies such as the International Telecommunications Union (ITU) (R-value) and commercial entities, such as NetForecast (www.netforecast.com). If no relationship can be determined, measurements can, in real or laboratory environments, determine the relationship as shown in Figure 11.

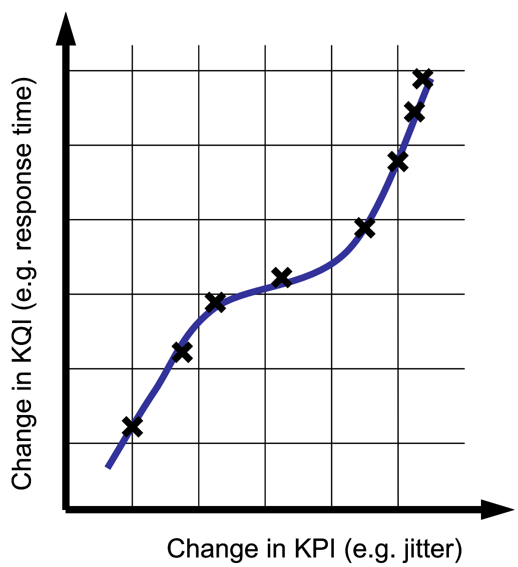


Figure 11: Relationship between KPI and KQI from Measurements

At each stage of the hierarchy shown in Figure 17, it is necessary to derive the KQI and KPI and map them to the KQI and KPI of the lower and upper tiers.

1.5 Consideration of QoS

QoS is used by an enterprise to measure performance both technically and in terms of customer satisfaction. The use of QoS agreements in SLAs provides providers with the opportunity to support new services and applications, build stronger long-term relationships and brand image, and maintain/grow market share. The growing complexity of global services brings together a myriad of services, suppliers, and technologies, all with potentially different QoS requirements.

In the past, data services were considered separately from other important business functions such as voice (PBX, etc.). This separation extended not only to procurement, but also to planning, wide area connections, administration, even wiring (buildings were originally flood wired with CAT4/5 for voice, not data). In this era it is only possible to determine QoS at the application layer (e.g., URL), with little reliance on the support layers in achieving an SLA.

Convergence yields many cost benefits, as fewer systems are easier and less costly to administer, procure, and support. One of the initial barriers to convergence was the considerable difference in network and business requirements for voice *versus* data networks. The reliability of voice systems, both internal and external, generally exceeds that of data systems, except those considered mission or life-critical. Wide area connectivity, generally SDH or SONET based, is a major component of such reliability.

As a consequence, voice and data network were provisioned separately (although perhaps sharing the same trunk) as shown in Figure 12, and any SLA in place would have been considered separately for each service.

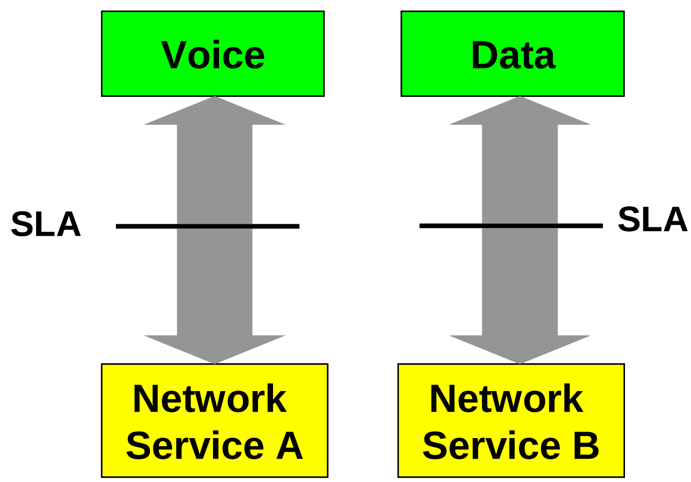


Figure 12: Voice and Data Services Contracted Separately

All data is not created equally. Some data could be mission-critical – such as telemetry – or covered by regularity bodies like the SEC (Federal Bank updates), while others – such as email (SMTP, X.400) or web traffic – may not be. The performance of an SNA network is greatly degraded if connections are not persistent. If voice is considered a data service – Voice over IP (VoIP), VoDSL – this may invoke the strictest criteria of quality as voice is very susceptible to delay, jitter, and loss.

Several possibilities exist: either traffic can be physically separated according to QoS (e.g., Figure 12), the system can be configured such that it can support the most rigorous requirement (e.g., over provision such that congestion does not need to be considered) as shown in Figure 13, or logically separated using Quality of Service (QoS). Combinations of these methods can also be employed.

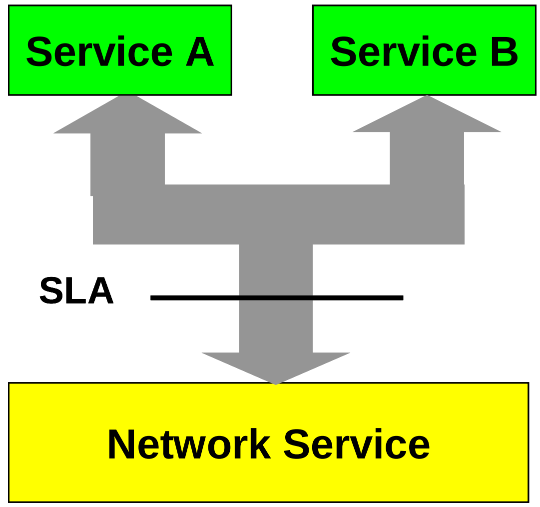


Figure 13: Two Services with an SLA Specifying the most Rigorous Requirement

Separation of the traffic by QoS requirement has been a common mechanism to achieve QoS, but is costly to deploy and support. A system that supports the most rigorous requirement will not in itself guarantee service availability (SA), and is costly to acquire as much of the system will be under-utilized. It is also subject to poor scalability as it will not adapt well to changing or growing usage unless they have been anticipated initially with even lower cost efficiencies. SLAs in these systems can also only be made on the aggregate system as a whole and not on its individual components; i.e., it does not accommodate priority such that if a system begins to fail, one type of traffic that is considered more important to the success of the business cannot be given preferential treatment in the system or in the SLA.

Of these options, QoS has the greatest versatility and, if properly implemented, can fulfil all requirements in the most cost-effective manner. In brief, QoS comprises three components:

1. Marking: Traffic is identified by either a logical or physical mechanism, or a combination of both. There must also be a common agreement throughout the system as to the meaning of the marking in terms of priority and QoS requirements.

2. Traffic Engineering: Mechanisms must exist in the system to manipulate traffic on the basis of its marking to ensure that it conforms with the requirements for the class (policing) and attempts to correct it (shaping).

3. Congestion Control: Mechanisms must exist in the system to be able to react to congestion within the system. Mechanisms would include signaling and selective dropping of traffic to slow down sources of congestion on a priority basis.

With the use of QoS, fewer network services are required with one SLA in place as an aggregate system. The SLA would define different parameters for each traffic type, also possibly with different reporting and support metrics. Each traffic type is marked in some way, as shown in Figure 14.

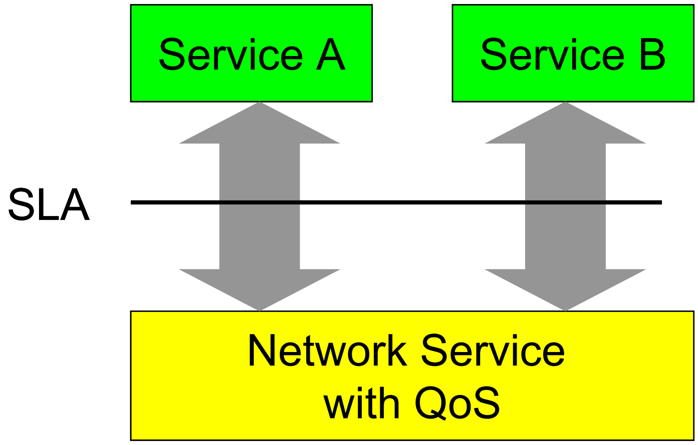


Figure 14: Network Service with QoS, Allowing one SLA for all Services

It is therefore important that if Network QoS is to be employed and enforced, that the SLA will define metrics for each QoS Class (Class of Service). This will require that marking mechanisms are employed by the enterprise. Several techniques for marking are available:

1. IEEE 802.1 Q: A portion of the IEEE 802.3 header is used to distinguish a unique virtual LAN (VLAN). Each VLAN could be considered as a separate class of service.
2. IEEE 802.1P: A portion of the IEEE 802.3 header is used to indicate a priority of the packet data.
3. TOS bits: A portion of the IP version 4.0 header is used to indicate a priority of the IP packet.
4. DSCP: DiffServ Code Point. A portion of the IP header is used to indicate the priority of the packet.
5. TCP/IP Header: Any combination of source, destination IP addresses, UDP/TCP port numbers, or any unique identifier of the packet is used to distinguish the priority of all packets that match that definition.
6. Circuit Identifier: For example, using VCI or VPI in an ATM circuit to define a Class of Service (CoS) or a label in an MPLS header.

For marking to be employed, it requires a common understanding of the service level marking in the end-to-end transmission of the packets.

It is also necessary to mark the packets at an appropriate point in the network. Choices are:

1. At the host (thick client) by the application
2. At the server by the application
3. By the Customer Switch or Router against known filters
4. By the CPE device supplied by the Service Provider
5. At the Access Edge Device in the Service Provider network
6. Between service provider networks

It is likely that an external service provider will remark the traffic against its own marking criteria to ensure correct marking, allow the possibility of a different marking scheme across the service provider network, and to apply traffic policing and monitoring of the traffic.

For SLA monitoring and reporting it will generally be necessary to monitor and report against each Class of Service (aggregate), all CoS (aggregate), and individual sessions (calls, etc.).

It will generally be necessary to employ traffic engineering and congestion control within the network to allow compliance.

1.6 Security

Security is often considered as a separate service and procured, monitored, and managed as such. An SLA is then applied to this service as any other. Although valid, security is an intrinsic characteristic of any service.

There are three basic security services to consider:

1. Integrity: Ensuring that data or service meets availability targets and has not been corrupted in anyway and only created, viewed, and modified by authorized personnel or processes as required. Denial of Service (DoS) attacks are attacks on the integrity of the enterprise information and services.
2. Confidentiality: Ensuring that data or service cannot be interpreted to its content or its presence by other than authorized personnel or services.
3. Non-repudiation: Data can be shown to have only come from the claimed source.

These are wide-reaching concepts that encompass many techniques. For example, integrity of data can be achieved by: physical security – denying access to the data so it cannot be changed, or the use of digital signatures.

In the formulation of a security business service, the threats to the enterprise and its information should be determined. Once determined, the risk of compromise of the enterprise and its information is calculated. It may be determined that the compromise of information is low risk with little or no impact on the enterprise or high risk where such compromise may impact or severely impact the performance of the enterprise in meeting its objectives. A determination of the level of risk allows the appropriate level of countermeasures to be deployed to mitigate the risk.

Security methods and countermeasures can be considered and applied in any of three environments:

1. GSE: Global Secure Environment (sites)
2. LSE: Local Secure Environment (rooms)
3. ESE: Electronic Security Environment (desktop)

With regard to SLAs, the KQI measures will be against the threat assessment. For example:

- No unauthorized access to data (confidentiality and non-repudiation)
- Systems will be available 99.9% of time (integrity)

The KPI against these KQI will be derived from measures of or with countermeasures; for example, firewalls, encryption devices, etc.

In these generic terms some of the KPI parameters only add further granularity to other parameters of a service or process that would be monitored and maintained in an SLA. The danger of having separate SLAs for security is that it adds to the complexity of the service and as such creates an opportunity for omission or error. Whether treated as a separate service with a separate SLA or not, any service should consider its security requirements in terms of integrity, confidentiality, and non-repudiation from the outset.

1.7 Training

It is important to understand that to implement and manage a service and its SLA, it may be necessary to provide more training for key personnel in both supplier and customer. For example, it may be necessary to provide training on the use of the Trouble Ticket application to achieve the reaction times necessary to achieve the SLA.

The correct level of training is determined by performing a Training Needs Analysis to all aspects of the service including:

- Service Resources: Do personnel have adequate training to configure, monitor, and maintain the resources?
- End Users: What training do the end users (the users of service) require to perform their tasks most efficiently?
- Systems Engineering: If real-time SLA monitoring is to be performed, what level of training is required to integrate the SLA interface into the existing or new systems?
- Support Staff: What training requirements are required for support personnel to maintain the service to the agreed SLA?
- Accounts: Any financial penalties or rewards as a result of the SLA process may require specific training in order to be calculated.

In each case, the medium (on-line, in-house, manual, etc.), time, and cost of the training should be assessed. This should be constantly reviewed so that new users, support staff, or most cost-effective service resources can be deployed.

1.8 Decommissioning

It is important to consider the costs associated with the de-commissioning of the service associated with the SLA, either as a single instance, or for the service as a whole. Issues and costs to consider:

- Removal of equipment to support the service or the SLA (e.g., routers, probes)
- Cost of storing archive SLA report information
- Retraining costs for alternative service for customers and support staff
- Compensation if early termination implies penalty charges

1.9 Methodology

There are many different enterprises, comprising a simple business model or more complex, multi-tiered, and peered business models. It is therefore impossible to define a comprehensive list of business services for which SLAs can be defined. For this reason, generalized business models will be discussed that, either singularly or in combination, will represent a wide business audience. Because they are generalizations, however, it will be necessary to define each business sector to understand the underlying rationale.

It is assumed that each enterprise will be using a number of distinct business applications in the furtherance of its objectives as an enterprise. These applications describe what the enterprise does: call center, healthcare, construction, distribution, etc. To support these applications, a number of business services are developed and supported within the enterprise. Examples of these business services would include: voice, transaction processing, office automation, and security. To fulfil the business services, a number of service resources are required to support such a service as well as other business services. Business services that support layers 4 and below in the OSI model are categorized as network services. Any business or network service may be provisioned internally or outsourced to an external service provider.

As no two enterprises are alike and the range of business applications and objectives are diverse and varied, it is necessary to produce a methodology that allows general principals to be applied to cover the largest number of enterprise scenarios. The methodology chosen is shown schematically in Figure 15.

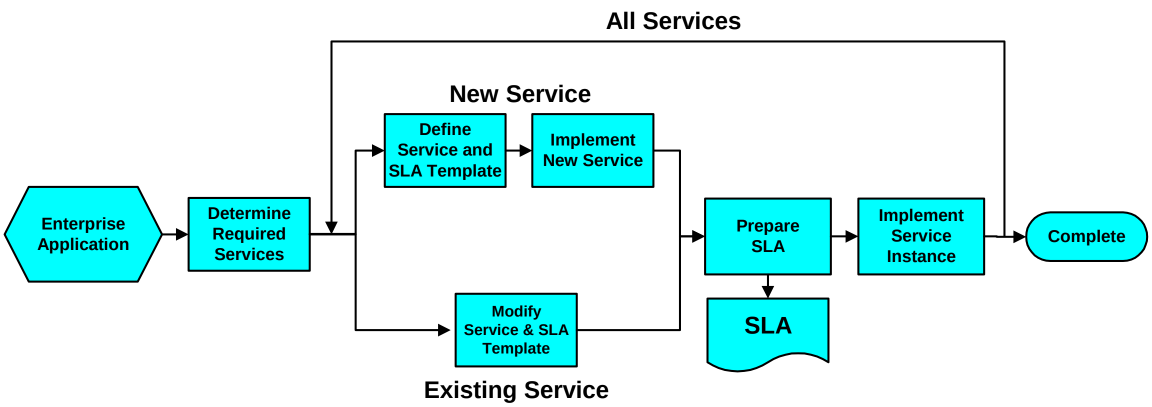


Figure 15: Enterprise SLA Methodology

The stages of the methodology are as follows:

1. Enterprises are generalized into a number of distinct sectors to cover as wide a scope as possible.
2. Business Applications are determined in a general fashion and applied to the target enterprises.
3. The Services and Service Resources for each of the Business Applications is determined.
4. The Key Quality Indicators (KQIs) of the services are determined. These KQIs may be objective or subjective.
5. The Key Performance Indicators (KPIs) of each of the services are determined for the objective measurement of the KQI.

6. The SLA is agreed.

7. Service Level Monitoring and reporting is performed.

It is doubtful that the enterprise is performing the task for the first time or that a particular service is not already in use by another business application.

2 Enterprise View

Enterprise refers to the overall business, corporation, or firm responsible for delivering products and services to its customers. An enterprise negotiates Service Level Agreements (SLAs) with internal and external service providers to satisfy the QoS requirements. Therefore, 'enterprise' is a general term to embrace all sectors including: commercial, utility, government, academia, and healthcare as well as information or communication service providers (ICSP). It is an assumption of Volumes 2 and 3 of this Handbook that enterprise refers only to ICSP. As a collective, no one single body provides regulation or standard business processes**.**

Many enterprises, however, do see the benefit of business processes and adopt frameworks such as ISO 9000 2000, ITIL, or similar to base business decisions and practices on best practice and experience. For ICSP, the business process framework is assumed to be eTOM/NGOSS.

This volume assumes such practices are in place and attempts to generalize enterprise sectors into common functions to describe the concepts of enterprise SLAs to address as wide an audience as possible. The techniques and discussions that are included should be assessed in the light of a particular enterprise and applied accordingly. Volumes 2 and 3 of this document address the services offered from a traditional service provider. It is recognized that through deregulation and competition between service providers, services that were once not associated with service provider offerings have become so, and could be addressed in this volume.

2.1 Enterprise Structure

It is assumed that the enterprise wishes to share or sell services or products to customers and partners in accordance with its defined business objectives. These business objectives therefore define the requirements for these services or products. We shall refer to the services and products as enterprise business applications (EBA) or simply business applications (BA). In general, an enterprise will comprise several business applications to achieve its business objectives. An enterprise will interface to its customers through the interface as appropriate for the application; for example, the PSTN would be the interface for a call center and the web would be the interface for an enterprise that sells through the Internet.

An enterprise will also have to consider restrictions on its business application, imposed by internal management doctrine or legislative issues such as environmental. These are considered business rules that may constrain the requirements for a business application. This structure is shown in Figure 16.

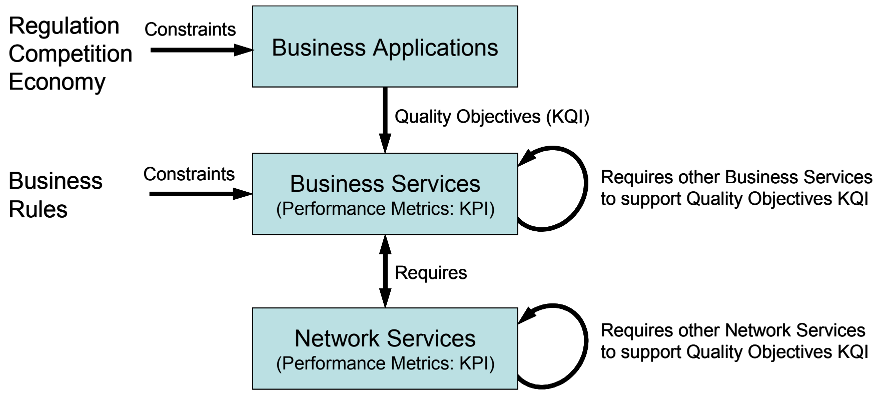


Figure 16: Enterprise Business Goals, Applications, and Services

In support of business applications (BA), business services (BS) are available to facilitate the application. For example, in a call center (the application), an obvious business service is voice communications. Business services in themselves do not raise revenue, but they are crucial in achieving business objectives and the effectiveness of a business application. One or more business services may be necessary to support a business application. A business service may be fulfilled internally or externally in an outsource arrangement. It therefore follows that an externally outsourced business service for an enterprise is a business application for the outsource supplier. To support a particular business service, a number of service resources will be required to fulfil the requirement.

In turn, business services are supported by a particular type of business service referred to as network services (NS). Network services may be either supported internally or outsourced to external providers. It is likely that any enterprise will have to consider both internal and external network services. External network services may be supplied by an ICSP in their entirety, or may provide transport services (Frame Relay, Leased Line, ATM, SONET/SDH) for which an overlay (e.g., IP) may be administered by an internal party within the enterprise. This hierarchy between enterprise business application, business, and network services is shown in Figure 17.

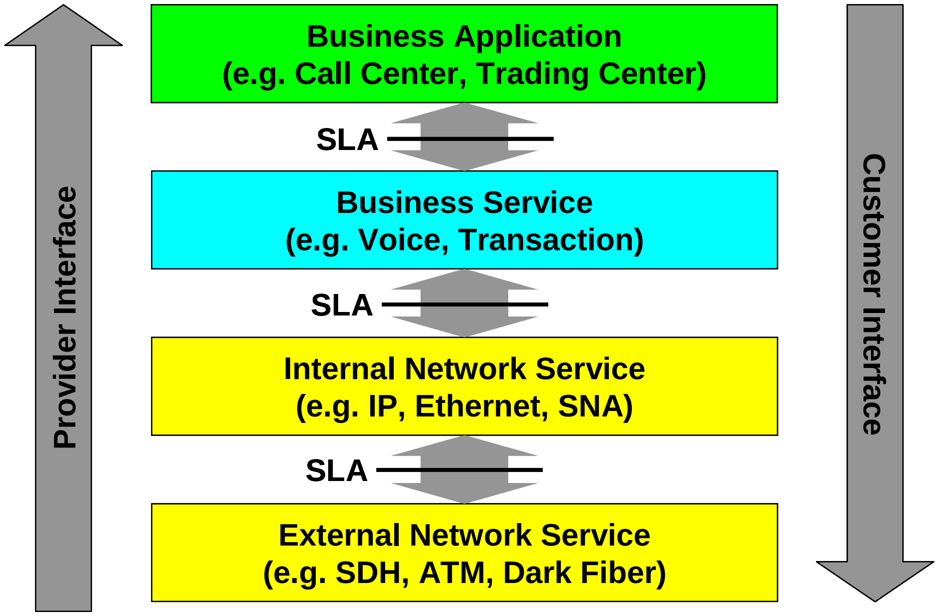


Figure 17: Hierachy of Application, Business Service, and Network Services in the Enterprise

Each layer in the hierarchy has scope for an SLA between each layer. Each layer represents a different terminology, management, and even contractual domain. In some applications, some layers will be absent and the exact structure will depend on the size of enterprise and other factors including budgets, manpower, skill, and training level of users and support staff.

A typical enterprise will comprise a number of business applications, supported by a number of business services that are in turn supported by internal or external network services as shown in Figure 18.

The exact combination for a particular enterprise will depend on:

- Size of the enterprise (people, revenue)
- Location (virtual office, branch offices, corporate HQ)
- Administrative structure
- Budgetary constraints
- Business constraints
- Political or regulatory issues

These same factors also influence whether some of the services are provided internally (e.g., PBX) or outsourced (e.g., Centrex).

Between each application and service (business or network service) there is a logical or physical interface where the service is delivered, known as the Service Access Point (SAP) or Service Delivery Point (SDP). The SAP is where a formal or informal SLA between the parties can be defined. Differences will be required between internal and external parties as to the structure of the SLA and to any penalties.

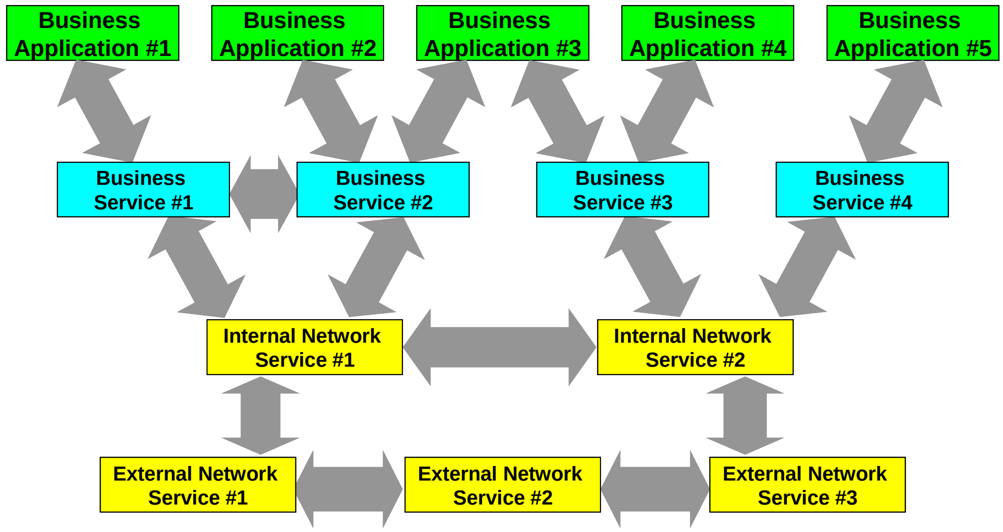


Figure 18: An Enterprise Offering Several Applications by Supporting Services

In general, therefore, to support an SLA for a particular business application or one of the supporting business or network services requires multiple SLAs to be developed in a parent-child relationship. These may be separate SLAs or annexes to the full SLA, depending on a number of factors including the administrative and logistic domains of the services and their management. The child SLAs may already exist in support of other applications or services or may need to be developed for the application. This relationship is illustrated in Figure 19. Although separate agreements, from Figure 18, it is clear that SLA performance data may need to be passed across the SLA interface to allow proactive management of the services. The availability and format of this interface may influence the choice of internal or external provider of the services.

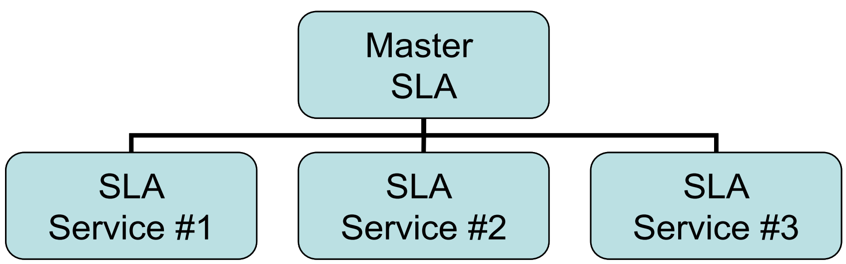


Figure 19: Multiple SLAs Supporting an Enterprise Application

The goal of establishing and managing end-to-end enterprise SLAs is only possible by mapping the requirements and objectives of the higher layers to measurable parameters at the lower layers.

Where external network services are offered by a service provider, fuller discussion of the interfaces and technologies will be found in Volumes 2 and 3 of this Handbook. Some aspects of the technology will be discussed for completeness and to avoid extensive cross-reference.

The preceding describes the generalized enterprise model for which enterprise SLAs must be applied and discussed. It is in this general model that all SLAs developed for a particular enterprise must be considered; i.e., many SLAs may be required and enforced. For simplicity, however, a particular instance of an SLA for negotiation will be considered as shown in Figure 20.

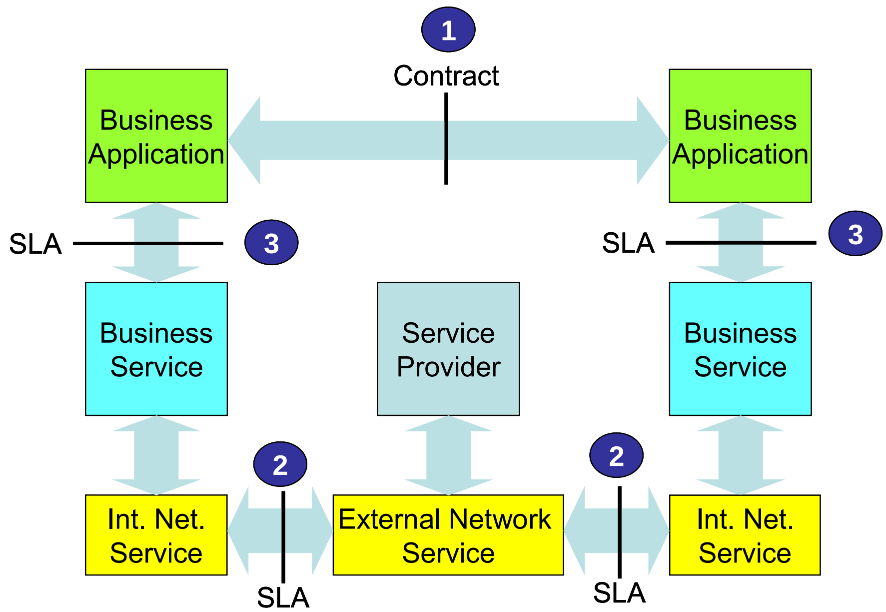


Figure 20: Generalized Enterprise SLA Environment

The principal components of the Figure 20 are:

1. A contract between two enterprise entities is required to deliver a service or product with better than best effort.

2. The product or service may use a network in one or more of the marketing, selling, delivery, or support of the product or service. The two enterprise entities are connected through a network provided by one (or more) service providers.

3. The responsibility for connecting the enterprises to the SP and the SP to the business application is the responsibility of a different enterprise group, possibly MIS or IT organizations.

Enterprise entities include:

- Divisions, branch offices, etc. of the same enterprise
- Enterprise and its partners
- Enterprise and its customers. Customers can include residential.

To allow compliance at SAPs 1, 2, and 3 in Figure 20 may require other SLAs to be defined, monitored, and enforced but would not be visible to the SLAs shown in Figure 20.

It is also important to note that to enable end-to-end SLAs to be enforced between enterprise entities, SLAs must be enforced between SAPs as well as across individual SAPs.

2.1.1 Business Applications

Generalized business applications will be discussed which represent a wide business audience, either singularly or in combination. Where examples are given, it still may be necessary to tailor the discussion to an individual enterprise. Because they are generalizations, it will be necessary to define each business application to understand the underlying rationale:

Call Center Handles CRM issues for an external audience.

Trading Deals with stock bond trading, financial transactions (money transfers), or other businesses such as air tickets and seating. Many such businesses may be involved in the transactions in either an external (online) or internal (stock market floor) manner. Automatic Teller Machines (ATMs) and Money Transfers for financial institutions would fall into this business application. Could be low or high value at high or low volume transactions.

Retail Deals in finished goods direct to the public. No major use of online services in the sale, except for POS, Credit Card validation. Characterized by low value (relatively), high volume transactions.

Etail Retail that engages its customers using an extranet (partners) or public Internet (B2C).

Electronic Information Delivery
Delivery of information in an electronic format. May be real time, near-real time, or best effort. Examples would include entertainment or news broadcasts, stock market information, sports highlights, e-books, etc. This information may be used for decision-making in real time.

Wholesale Resale of finished goods and other goods to other wholesale or retailer enterprises (B2B).

Manufacturing
Manufacturing of finished goods and equipment.

It is likely that any enterprise will encompass a number of these business applications. Examples are shown in Table 1:

	Business Application						
Enterprise	**Call Center**	**Trading**	**Retail**	**Etail**	**Info Delivery**	**Whole-sale**	**Manu-facturing**
Broadcaster					✓		
Healthcare					✓		
OEM		✓		✓		✓	✓
Distributor	✓	✓	✓	✓	✓		
Food Chain		✓	✓				
Outsourced Call Center	✓						
Stock Broker	✓	✓		✓	✓		
Transportation					✓		

Enterprise	Business Application						
	Call Center	Trading	Retail	Etail	Info Delivery	Whole-sale	Manu-facturing
Utility (Gas, Elec, Water)	✓				✓	✓	✓
Education					✓		
Service Provider	✓	✓	✓	✓	✓	✓	
Government							
Defense (procurement)	✓	✓		✓			
Defense (operational)	✓				✓	✓	
Tax Collection	✓	✓		✓			

Table 1: Illustrative Business Applications for Sample Enterprises

A broadcaster will perform Information Delivery to other broadcasters, subscribers, or to the public, either over a broadcast medium like satellite or cable or over the Internet. If the broadcaster offered pay per view or on-demand services, then the etail and the call center applications would be used, together with trading to process the charge or billing.

Healthcare is increasingly using telemedicine to make better use of resources and is a heavy user of telemetry equipment. If telemedicine is performed as an outsourced arrangement, it would be considered as an application; otherwise, it would be considered a business service. All systems that perform diagnosis would be considered an application, whereas a system distributing and handling patient records would be a business service. The Office application would also include patient records, drug records, etc.

An Original Equipment Manufacturer (OEM) will perform Trading and Etail to deal with suppliers and customers. In the example in Table 1, the OEM also performs Manufacturing, although this is increasingly performed as an outsourced facility. Research and development is considered under 'Office' although the requirements may differ from the normal administration tasks. As a result, R&D and administration may be separately managed and provisioned.

The illustrative Distributor deals with its customers through both outlets (retail), over the Internet (etail), and may rely on advertising over the network (information delivery). Money transactions are handled from point-of-sale (POS) or as part of e-commerce. For telesales, a call center application may be used.

A Food Chain increasingly processes money as a transaction at POS or supports in-house Automatic Teller Machines (ATM). Food chains are also moving to online ordering for their customers and use electronic means to distribute marketing material.

Call centers are a growing sector and are often outsourced, answering in the name of the client and having close links with the client's normal support infrastructure as an extranet connection.

Stock Brokers require up-to-the-minute business information and status from worldwide stock markets. Some brokerages are also the source of much financial information on which other traders and their customers will base their stock trade.

Shipping is an interesting enterprise in that it may not at first be perceived as being associated with networking technology and applications, but it increasingly uses telemetry to track vehicles. This information is passed on to its customers through information delivery systems.

Education (primary, secondary, tertiary) uses enterprise applications for administration in class, home, or distance teaching. Research also entails work in developing new applications and the requirements can be more extreme as a result.

In many countries, central and local government combined have the biggest budgets of all enterprises and as such are worthy of further consideration. Two important enterprises are considered in Table 1: that of Defense and Tax Collection. Defense in general requires information delivery to gain intelligence, develop plans, and communicate them to its forces and allies. Logistics is an important area of concern that effects military effectiveness, so the Defense sector can be considered as a wholesaler of equipment and supplies are on very large scale.

Tax collection is a crucial aspect of any government and is increasingly moving from paper systems (covered as 'Office') towards e-payment systems. In some instances, these systems are very advanced compared to other enterprises.

2.1.2 Business Services

Business applications (BA) require (enterprise) business services (EBS) to enable them. These services do not in themselves fulfil the business application, but enable their function. One of the most important enterprise business services is voice. Voice communications is an important aspect of almost all enterprises, but the telephone, CTI, or cellular phone does not in itself further the enterprise cause – that requires employees performing a task that use the business service (the application).

To fulfil a particular enterprise business service may require the cooperation of other business applications or enterprise business services. For example, an online reseller of equipment (business application) may require an enterprise business service for inventory control, order processing, and web services.

Again, in an attempt to generalize, services have been defined to adequately represent a wide business audience without being definition. These services and an outline of their characteristics are:

Voice — Normal voice services. These services are provided by either an ordinary phone or by a CTI device. Short duration. Full duplex.

VTC — (Video conferencing). Point-to-Point or Point Multi party video conferencing. These facilities are available on demand and are real time. They may be scheduled, fixed, or dial-up (or similar in packet world). Short duration. Full duplex.

Streaming — Streaming video, audio, or data services. These may be used on an *ad hoc* or permanent basis for news broadcasts, stock quotes, etc. Near-real time in context. Simplex.

Telemetry — Telemetry data for monitoring equipment. Half duplex. May need to be real time for control purposes.

Transaction — A short message for the transmission of time-critical and possibly valuable information. Rates may be low or high.

Email — SMTP or X.400 style messaging services. Would also include SMS and MIME.

Office Automation — Administration and business planning productivity tools. Examples would include word processing, spreadsheets, order processing, shared calendars, contact databases, etc.

Help Desk — A support service that provides assistance to users or customers via either email, phone, or in person (or any combination).

Broadcast — High quality video and audio transmission in real time. May be multiplexed with other feeds (e.g., outside broadcast of live events). One-to-many, simplex operation with possibly a slow return path (PSTN) for Video on Demand (VoD).

Tele-working
: Supporting a user to perform his/her normal business functions remotely. This may be on an *ad hoc* (hotel room) or permanent (SOHO) basis.

Intranet Web
: Web-based services for internal users that are not normally accessible to external users or by personnel outside the company. QoS and SLA parameters may be applied to Intranet.

Extranet Web
: Web-based services for external users with controlled access to business data. QoS and SLA parameters may be applied to an extranet. This is restricted access to *bona fide* users to differentiate from public access (no controls).

Public Internet
: Access to the Web and other Internet services such as SMTP, FTP, etc. on an unrestricted basis. SLA and QoS parameters can be applied within an ISP's network but not in general between any arbitrary points within the Internet.

Security — Supports the appropriate levels of availability, confidentiality, integrity, and non-repudiation for the enterprise and its information against understood threats to the enterprise and its information by deployment of countermeasures (service resources) appropriate to the level of risk associated with the threat posed by the compromise of the enterprise or its information.

In some instances business services can be accomplished without interaction with a network, but increasingly business services will communicate with other business services and business applications through network service providers (NSPs). These NSPs may be internal (LAN) or external (WAN) and may use internal resources (MIS), be outsourced, or provided by a service provider.

Not all business services are required to fulfil the business applications in the enterprise. Again, in an attempt to generalize, an illustrative perspective of the required business services for each of the applications discussed is shown in Table 2.

	Business Application						
Business Service	**Call Center**	**Trading**	**Retail**	**Etail**	**Info Delivery**	**Whole-sale**	**Manu-facturing**
Voice	✓	✓	✓	✓		✓	✓
VTC						✓	
Streaming		✓		✓	✓		
Telemetry					✓		✓

	Business Application						
Business Service	**Call Center**	**Trading**	**Retail**	**Etail**	**Info Delivery**	**Whole-sale**	**Manu-facturing**
Transaction		✓	✓		✓		
Email	✓	✓		✓	✓	✓	
Office Auto						✓	✓
Help Desk	✓	✓	✓	✓	✓	✓	✓
Tele-working							
Intranet	✓	✓	✓	✓	✓		✓
Extranet		✓		✓	✓		
Public Internet	✓	✓		✓	✓		

Table 2: Illustrative Enterprise Business Services by Business Application

In Chapter 4 the requirements for each of these business services will be further discussed and translated into quality and performance metrics. These can be used to define SLA agreements for the applications and also to establish the requirements for network service SLAs.

2.1.3 Network Services

Enterprises have significant choice in network services to use with their applications and services. For LAN connectivity, the IEEE 802.X protocols, especially variants of IEEE 802.3 (Ethernet), are perhaps the predominate today but significant legacy, revenue generating, or mission-critical applications are supported by other protocols such as SNA to which newer applications may need to interface.

For external communications, many of the transport services – such as SONET/SDH, ATM, Frame Relay, and ISDN – have traditionally been supplied by service providers, but deregulation and new market entrants have made new options – such as Ethernet and Dark Fiber – available to a wide enterprise audience, allowing cost-effective, high-speed connectivity.

Broadband connectivity to residential and enterprise users allows much lower-cost bandwidth to be deployed and utilized by applications. Aggregation links to the enterprise may also be low-volume, high-content as well as the traditional, high-volume, low-content sessions.

The pervasiveness of the Internet and the widespread adoption of the Internet Protocols have de-localized commerce over the Internet. Propagation delays therefore have to be considered when SLAs are to be established between entities separated widely in a geographical sense, especially if connection is made over the Public Internet.

Another difference that must be considered in enterprise applications is the asymmetry in both network requests and in network connections, especially Broadband. For example, ADSL provides a user with up to 8 Mbit/s downstream and up to 1 Mbit/s upstream (often much less). This provides opportunities for high-bandwidth content to the customer (e.g., video) with the upstream traffic used for control, signaling, and lower priority information such as emails. As a consequence, the SLA requirements for the downstream and upstream components of the network traffic may be very different.

2.1.4 Peer-to-Peer Networking

With the growth of applications like Napster, Morpheus and other peer-to-peer (P2P) applications, the importance of such techniques for the enterprise should be anticipated.

The difference between the traditional client server approach to applications and P2P is clear, although its impact on enterprise applications is uncertain. Initiatives such as grid computing and web services for highly distributed applications for resilience and scalability are likely to change the network environments of the future.

Client server architectures favor servers connected to fast networks, linking clients generally on lower-speed connections. Similarly, in terms of computing resources, a server architecture will provide the computational environment, with the client displaying the results.

In the P2P environment, clients establish connections with other clients via information held in servers in the form of a directory service or database. Once the connection is established, client-to-client communication only is important. The same technique can be used for both information (file) sharing and computing resource.

There are few enterprise applications that make full use of the P2P techniques popularized by the likes of Napster and Morpheus. However, their popularity and innovation means that such techniques will be more widely adopted in the enterprise in the future. This paradigm shift is likely to cause great impact in the enterprise and should be anticipated where possible in any SLA.

3 SLA Management and Reporting

3.1 Introduction

Conformance against an SLA is ensured by instrumenting the systems to provide the appropriate KPI and KQI measures at the required sample rate. It is important in the design process to ensure that the measurement process itself, especially in out-of-specification conditions, does not create or worsen the condition by adding further load to the system (processing power, management traffic overhead).

If a KQI is determined by correlating KPI or KQI data from another service, this information may be required in real time so that true measurements can be made for proactive management to allow fault prevention, rather than using aggregate or stored information for reactive management.

The SLA should be monitored continually at a rate appropriate to the requirement, to assure that corrective actions can be taken and also collated to form management reports.

Once SLAs are in place, conformance against the SLA is demonstrated by the production of reports.

As with the SLA itself, the display and interpretation of the report data should be clear and concise, it should be clear when the SLA is out of conformance, and not hidden within a myriad of conformance data.

Who generates the reports is an issue that is coming under scrutiny.[2] Options include the provider of the service, the customer, or a trusted third party. With each of the options there are different costs associated and a level of trust has to be developed to avoid conflicts. The choice of reporter will depend on the particular circumstances for the contract; for example, the level of penalties. It is also possible that if either the customer or provider generates the reports, this can be handled by groups other than those responsible for providing or using the service in question.

3.2 Service Monitoring

For each service, performance-related data is retrieved from the relevant instances of the service resource. These are collated and combined as discussed in Section 1.4 to form KQI for each resource, and further combined, again using techniques outlined in Section 1.4, to form the service and product KQI, as shown in Figure 21.

[2] Richard Sturm, Enterprise Management Associates, Network World Fusion, November 28, 2001.

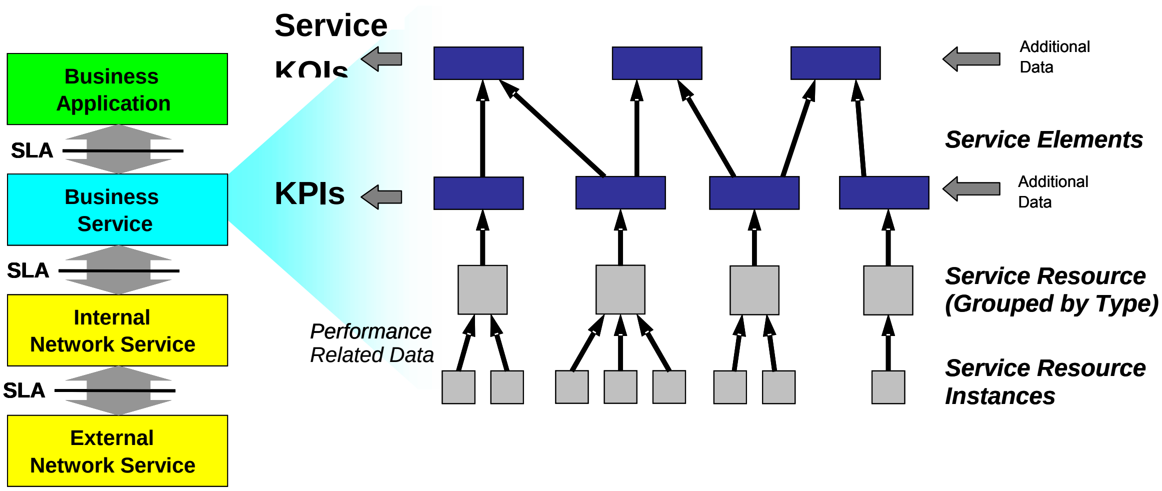


Figure 21: Relationship between Service Resources, KQI, and KPI

3.3 Process

The SLA Monitoring and Reporting process is illustrated in Figure 22.

Using instrumentation within the system to measure the KPI and KQI, service performance data is collected and collated into a form that can be manipulated to allow diagnosis and report generation. Instrumentation can include:

- User satisfaction surveys
- Test applications (including phantom callers)
- Client-based Monitor Agents
- Server-based Monitor Agents
- Network-based Monitor Agents

The period of collection should be defined in the SLA and should be such that it provides the granularity to ensure that the SLA meets its requirements. Instrumentation data may need to be passed across the SAP so that it can be used as KQI and KPI data to determine KQI or KQI for other services. For reactive management, this information may be required in:

- Real time, so that proactive management can be performed for fault prevention
- Near-real time or off-line, time-stamped format so that it can be correlated with other time-stamped data for improved performance information for reactive management
- Offline, aggregate information that can be correlated with other aggregate data to provide trend analysis for reactive management

Instrumentation data is collected from each of the relevant element instances as shown in Figure 22.

It is important to distinguish between performance “events” and performance “parameters”, as follows:

Events — Instant or near-instant phenomena that occur within the service or its environment that effect the KQI of the service. Examples include lost or misdirected packets, loss of signal, power failure (non-redundant), etc. Events may be signaled via mechanisms such as SNMP traps and interrupts, or may be inferred by catastrophic loss of service; for example, an earthquake.

Parameters — Derived by processing a series of measurements or events over a measurement period in a defined metric that can be reported. These may be time-related, ratios, or event rates. Examples are availability, throughput, utilization, average call response time, and ethernet collision rate per packet (collisions are a necessary and predicable consequence of the IEEE 802.3 design, so ratio with non-collision packets rather than collision rate is important).

The SLA Performance Data is collated to form an internal report that can be used to diagnose the performance of the system for internal diagnosis and to produce the customer report. Collation may require combining KQI or KPI from different services or products, covered by different SLAs, from potentially different providers. Collation may be performed directly by the collection tool or by using a middleware application, using a common interface technique such as CORBA, XML, or SQL. The sampling period may be real time, semi-real time, or historical. A real-time interface in a commonly used interface (SOAP, J2EE, XML, C/C++) may be required and necessary to maintain other SLA objectives.

The internal report may be in a different format from the external report to fit in with internal procedures and tools. Conformance thresholds may be set at more aggressive values than those defined in the SLA to ensure corrective action can be taken before non-conformance ensues. For this reason, internal reports are likely to be generated at more regular intervals than the external reports to allow remedial action to be taken and improve or enhance (cost reduce) the system.

In an enterprise application requiring a multi-tiered SLA, it may be necessary to demonstrate how the service has been performing recently (typically a few hours), as a result of a support call or in the diagnosis of non-conformance at a different tier.

The external report should be presented to the customer at the appropriate time interval and in the format agreed. This external report can be used by the enterprise to provide assurance of conformance and for trend analysis for future growth or new opportunities.

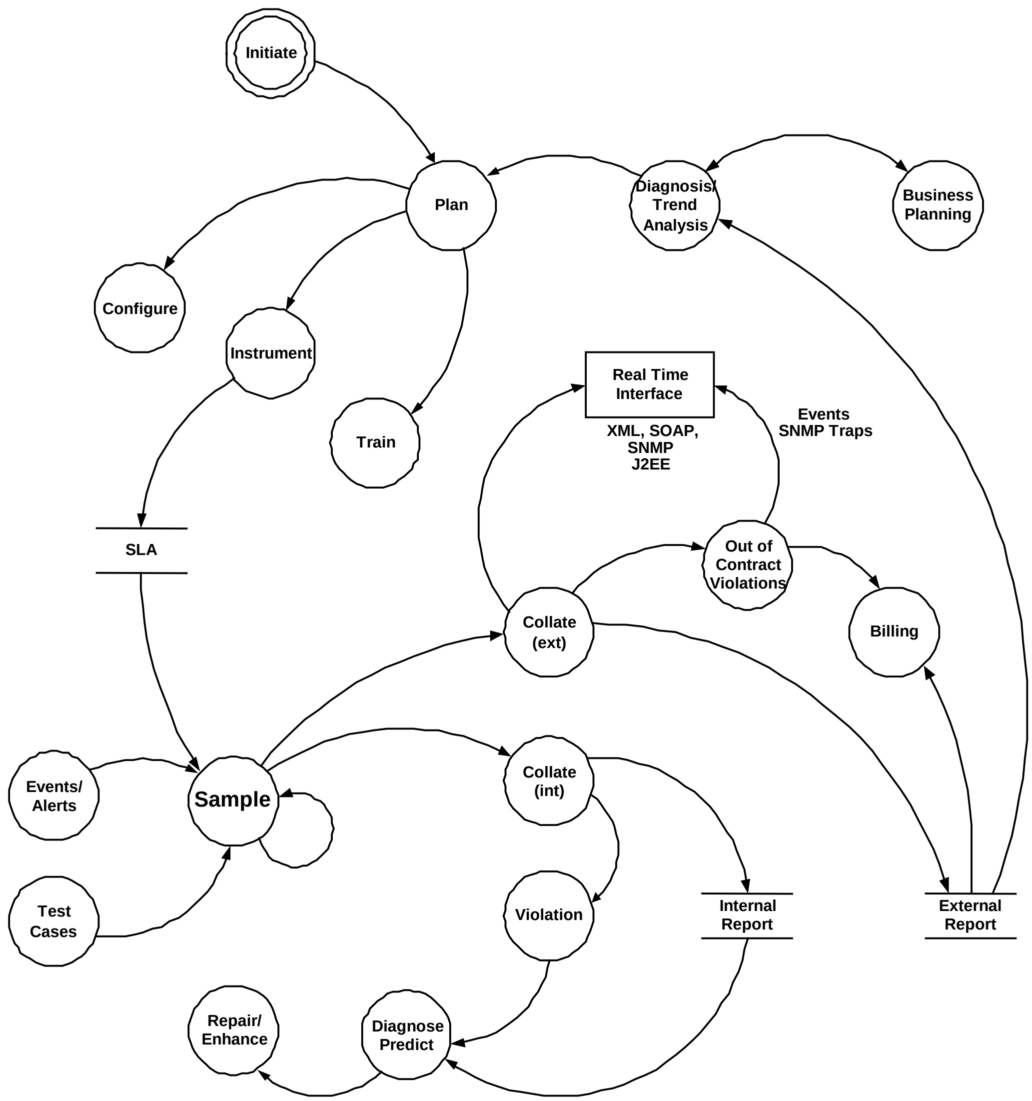


Figure 22: SLA Monitoring Process

It is likely that the SLA data and reports may have to be stored for audit purposes.

3.4 Reporting Mechanisms

It is likely that there will be a number of different functional groups who will wish to see the SLA Management Report. These may include:

- Senior Management, concerned with high-level achievement targets (e.g., Target 99% achieved)
- Finance, for billing and cross-charging
- Engineering, for diagnosis and planning

- End users, as part of CRM

The format, language, and style of each of these reports should be appropriate to the audience:

- Senior Management – presentation style, color, pie charts, etc.
- Finance – spreadsheet format (machine-readable)
- Engineering – trend graphs, raw data for developing own reports
- End user – web-based

The presentation of the SLA Reports is an important aspect of the customer's QoE, so care in the layout and format of the reports is important.

With the advent of new reporting tools, it is possible to allow (some) users to develop their own reports. Again, such facilities add to the customer's QoE as long as these tools are intuitive and add value.

4 Enterprise SLA Negotiation

4.1 SLA Development Process

At the highest level of the enterprise, it is assumed that the enterprise is working towards some higher-level objectives in all of its business decisions that an SLA or collection of SLAs will ultimately support. Examples of these would include:

- Become market leader in market sector
- Increase share price by 50% a year
- Improve patient waiting times

Any business process should be judged against these higher-level objectives and any conflicts highlighted. Conflicts may require modification to the application objectives or requirements or, less likely, changes to the enterprise objectives themselves. From the issues outlined in Chapter 3, a possible SLA development process is discussed and shown in Figure 23.

With reference to Figure 23, the process starts with the business decision (product manager, board level, as appropriate) to pursue an opportunity by providing an application (product or service) to its customers or partners. In the normal business process, the objectives of the application are defined, from which the requirements for the application are derived.

These application requirements are fulfilled by:

- Acquisition of product or service
- Development of a new product or service
- Outsource arrangement
- Enhancement of an existing product or service
- Integration of new or existing product services

For the purpose of this document, it is assumed that the application will require the use of business and network services. The relevant services are then determined. If the service is not already in use, then the service is required to be initiated and an instance created for the particular application. Where the service already exists, a new instance is created. SLA metrics can be applied to either or both the service as a whole and a particular instance. Further decomposition may also be required if real-time SLAs are required, but the principles expressed in this chapter remain the same.

For each service or instance required, the requirements of the application (KQI) are determined and mapped to the service requirements. This will allow definition of the Key Performance Indicators (KPIs) for the service instance that can be used in monitoring and performance reporting. Other secondary indicators may be derived and monitored for diagnostics, fault prevention, and resolution.

If either the service does not exist or requires modifications to existing services, the timescales, costs, and impact on other applications and services should be considered. If there is a conflict with the requirements for the service from other applications (as defined by the enterprise application), these must be escalated through management to resolve the conflict. The costs of commissioning, lifetime support, and decommissioning of each service instance and the service as a whole need to be considered in the decision-making process.

If no such conflicts exist, the level of training required must be assessed in terms of cost and time to ensure the appropriate personnel are trained adequately. Once again, the impact of any extra training must be assessed to ensure that the enterprise objectives are not compromised.

Monitoring should be considered to ensure that the KPI can be measured in the time period required and in a manner that does not impact on this or other services (e.g., management bandwidth overhead).

Once all these factors are considered for an individual service, an SLA can be developed (annex to a Master SLA or separate contract). This process is repeated for all relevant services until either a conflict is found that cannot be resolved or all relevant SLAs have been developed.

It is important to note that the KQI and KPI for each service should be the minimal set required for that service. If, however, other applications or services require underlying services with more stringent and rigorous KQI or KPI, it is those KQI or KPI that should be considered in the SLA. The SLA therefore reflects the most rigorous requirements for the KQI/KPI for all services supported by the SLA.

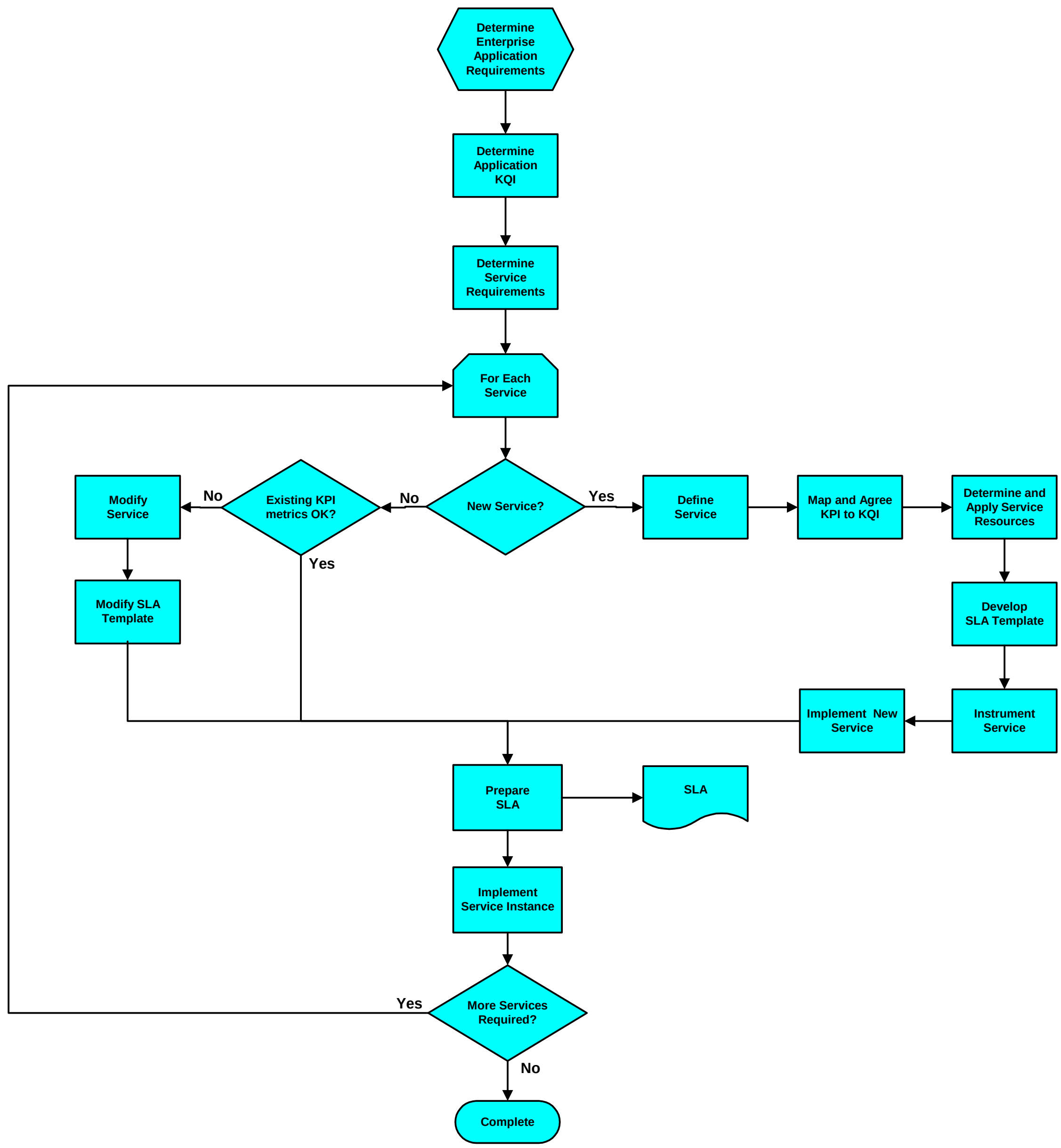


Figure 23: Enterprise SLA Development Process

4.2 Form of an Enterprise SLA

The exact form of an SLA will depend on the two entities that are entering into the agreement or contract. In particular, the form of the SLA will be different (especially in the area of penalties) whether the SLA is between an enterprise and an external party (e.g., service provider), between internal parties, or between the enterprise and its customers and partners. Each party may wish to consider the SLA as part of a long-term, strategic relationship, so should be written to foster such long-term relationships, as long or short-term tactical relationships, or to foster brand image for good customer relationships. The SLA is a mutual agreement between the two parties with expectations from both sides defined and agreed in the SLA as well as the course of action to be taken when deviations from these expectations occur.

An SLA is, in general, a part of (as an annex) or in its own right, a legal contract between the parties, especially for SLAs between an enterprise and external parties. It is therefore important to consider taking legal advice as to the exact form of the contract and the language used. If the SLA is to span international boundaries, legal advice that has an understanding of the differences in contract law, environmental, employment, and any relevant regulatory environment in the relevant countries should be considered. Even internal SLAs, where the SLA spans international boundaries, may have to take these issues into consideration.

The language and terminology used should be appropriate to the audience. A glossary may be necessary to explain common terms, but in principal the SLA should be written in a manner such that it can be read by someone versant in the particular service or technology in question. This also applies to any legal advice taken in the preparation and negotiation of an SLA. If the SLA is written between enterprises or between an enterprise and a service provider, it is likely that legal advice will have been sought by both parties so legal language and terms are not inappropriate. Between an enterprise and an end user (residential), this type of language may give a negative view of the enterprise in terms of QoE.

The relevant law should be stated and considered in the negotiation and preparation of the SLA. Once again, if the relationship is aimed to be long-term and strategic, then a mutually acceptable law should be considered; if tactical, then it would likely favor the party creating the initial draft.

The SLA should make it clear, in plain language, the aim of application that this service supports. Although unlikely to form part of the contract itself, it will help both parties in negotiation to better understand the requirement.

It should be considered whether any parts of the SLA (the existence of the relationship, the contract itself, reports, in total or part) will be considered confidential. There may be competitive advantage in either the service offered or the application supported. Those areas considered confidential should be clearly identified and the duration of the confidentiality stated.

When a provider defines a service for the first time, an SLA template will be defined that forms the basis of all instances of the SLA.

An outline of the main topics for inclusion in an SLA is discussed in the following sub-sections. The exact form of the SLA will depend on a number of factors, including whether the SLA is a separate contract in its own right or forms a part (annex) of a larger contract. If the latter, some of the sections that follows may not be required. It may ease further negotiations if annexes can be added to the SLA for new services without having to re-negotiate the main body. If this is likely, then the SLA should be written appropriately for the first service.

4.2.1 Introduction

This section should document the relevant parties that are entering in the SLA agreement. The introduction should also contain a brief overview of the need for the SLA and the application or services it serves. This information should include the KQI for the application to be included and how the KPIsof the service support the concept of the application KQI.

4.2.2 Customer Requirements

This section should document how the customer is to use the service such that it is clear what the service must support. For example, if the requirement is to support a round trip time of less than 1 sec. for a transaction, it will be necessary to understand the peak value and length of any bursts of transactions that are anticipated. It may be necessary to determine how the service should respond when (in this example) the transaction rate is exceeded.

4.2.3 Overview of Service

This section should describe the service including the location of the physical and logical interfaces between the two parties, who owns which part of the interface, number of locations, and any other information that describes the service or product adequately.

4.2.4 Term

This section should detail the period over which the SLA is valid, perhaps with a commencement date.

4.2.5 Responsibilities

This section should detail the responsibilities of both the customer (to the provider to ensure conformance) and those of the provider (to the customer). Expectations from both sides can be detailed in this section.

4.2.6 Details of Service

This section should describe the parameterization of the service in terms of the (KPI), as they will be reported to the customer. This will probably take the form of a table, similar to that shown in Annex A: KQI/KPI Parameterization. It should clearly show the levels of acceptable performance and non-conformance, and out-of-specification conditions.

4.2.7 Exceptions

It is likely that exceptions need to be included and clearly documented in the SLA. Downtime for upgrades or routine maintenance may be necessary but need to be described with notice periods, etc. in the SLA. In multi-site environments, care must be taken to ensure that the downtime is explicit. For example, if the SLA is between an enterprise and a service provider providing connectivity between a corporate HQ and its branch offices and the total maximum downtime is say 10 hours per month, it is likely that you may want to define the maximum downtime of the HQ and each branch (as in principal they may be different).

4.2.8 Sampling and Reporting

It will be necessary to define how often the SLA KPIs are measured as a measure of conformance and how often they are collated in the form of a report or to calculate the application KQI. The method of reporting (web, paper, etc.) may also be necessary. Reports for non-conformance (out-of-specification) may require a different frequency (instantly, within an hour, etc.) bypassing the normal collation process. The method of reporting non-conformance (e.g., pager) may also be different and should be documented. The reporting of asynchronous events – alarms, alerts, traps, etc. – may also be different and should be replied to. It may further be necessary to establish what frequency of asynchronous events can be tolerated by either the customer or the provider.

Sample reports should be agreed and included with the SLA document.

If the SLA performance data is required to determine performance metrics (KQI, KPI) in either real time or near-real time, then the format of this data, the interface (e.g., SQL, XML, CORBA, etc.) and the support, availability, integrity, and confidentiality of this interface needs to be defined.

It is possible that tiers of reports may be available in an online (web) and offline form. For example, a user may be able to view the reports for the SLA for his/her own use. Access control would have to be agreed and verified. How long these reports are stored, either online or as an archive, should be defined.

4.2.9 Penalties

The penalties for non-conformance should be detailed, but emphasis should be made on motivation rather than to antagonize the situation. These may range from notification of:

- Lost fees
- Repayment of fees
- Compensation for lost earnings
- Termination
- Combination of the above

Invoking penalty clauses does not necessarily gain great benefit, in that the damage is already done and any monetary penalty is unlikely to compensate even partially for business lost or damage to brand image. Terminating and moving to another provider may not necessarily improve matters; such action will lose the goodwill (if any) and cumulative knowledge gained between the enterprise and provider. For internal parties, penalties will not necessarily act as an incentive to remedy the condition once failed. Incentives to rectify the problem as quickly as possible should be considered to offset any penalties, to encourage co-operation at a likely stressful time and bonuses for over-performance.

4.2.10 Dispute Resolution and Escalation

This section should document how differences of opinion on the SLA in either the contract, its reports, or performance are resolved. It may be necessary to provide contact details for these instances and also to document how the situation will be escalated to senior management if the situation cannot be resolved. For SLAs between external parties, arbitration may be a quicker and more cost-effective remedy in many cases. For internal parties, this section is likely to be absent and resolved within the normal management process.

4.2.11 Change Requests

This section should detail the procedure for how change requests to the SLA will be made and handled, with any expense detailed. Frequency of change requests should be detailed, with terms like ‘not to exceed’ such that an understanding of the total cost of such requests can be determined. Notice periods for change requests should be documented. Performance of these change requests may be subject to the penalty clauses.

4.2.12 Termination

It is likely that either side may wish to terminate the SLA for a particular reason. These reasons may need to be documented and notice period for termination and any costs associated detailed. The notice period may differ for supplier-to-customer and customer-to-supplier.

It should also be considered what would happen in the event that one of the parties is acquired by another party or acquires another enterprise such that the service requirements may be very different. Consideration should be given to whether the SLA should terminate, continue as-is, or be renegotiated.

4.2.13 Relevant Law

This section should detail which is the relevant law to be considered for the SLA and under which jurisdiction any breach of contract will be detailed.

It is likely that this section may be missing between internal parties unless the two parties are international and there are significant differences in relevant law pertinent to the operation and performance of the service between the different countries.

4.2.14 Confidentiality

This section should detail and highlight any aspects of the SLA – its existence, performance, reports, and report data – that are confidential.

4.2.15 Warranties

Areas that are covered by any warranty conditions should be highlighted. Where warranties already exist for some service resources, how these effect the SLA should be detailed.

4.2.16 Indemnities and Limitations of Liability

It is important to understand who is liable in the result of failure of the SLA, either the provider or the customer.

4.2.17 Signatories

The SLA should be dated and signed by relevant signatories from both organizations.

5 Detailed Service Definitions

5.1 Service Resources

As seen in Section 1.6, it requires resources to deploy a service or product. Resources can include:

- Equipment
- Personnel
- Licenses

Resources either provide services directly, either singularly or collectively (e.g., phone, PBX, and CAT4/5 structured wiring), support the services (e.g., buildings, operators, HVAC), or allow monitoring or repair of the service (e.g., monitoring probes, agents).

A selection of resources to support the enterprise business services (EBS) under discussion is shown in Table 3.

	Business Service											
Service Resource	**Voice**	**VTC**	**Streaming Video**	**Broadcast Video**	**Tele-metry**	**Transaction**	**Office**	**Help Desk**	**Tele-working**	**Intranet**	**Extranet**	**Public Internet**
User Terminals												
Class Phone	✓	✓							✓			
PDA			✓		✓				✓		✓	✓
Cellular Phones	✓	✓	✓		✓				✓		✓	✓
Laptops			✓		✓		✓	✓	✓	✓	✓	✓
Thick Clients			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Thin Clients							✓	✓	✓	✓	✓	✓
Dumb Terminals					✓	✓	✓	✓				
Application Resources												
Directory Services						✓	✓	✓				
Office Applications							✓					
Custom Applications					✓	✓	✓	✓				
Database					✓	✓	✓	✓				
Email							✓	✓				

	Business Service											
Service Resource	Voice	VTC	Streaming Video	Broadcast Video	Tele-metry	Transaction	Office	Help Desk	Tele-working	Intranet	Extranet	Public Internet
DHCP Server	✓		✓				✓					
DNS	✓		✓			✓	✓					
Access Control							✓	✓	✓	✓	✓	✓
Host IDS										✓	✓	✓
Load Balancers						✓						
Mail Servers							✓					
Application Servers (all types)					✓	✓	✓					
File Servers							✓					
Print Servers							✓					
Web Servers							✓					
License Servers							✓			✓	✓	✓
Conference Bridge	✓	✓										
Gatekeeper/Call Agent	✓	✓										
Printers						✓	✓		✓			
Network Storage			✓	✓	✓	✓	✓		✓	✓	✓	✓
Backup Services					✓	✓	✓	✓	✓			
Internal Network Resources												
IP	✓						✓	✓	✓	✓	✓	✓
IP VPN							✓	✓	✓		✓	
SNA					✓	✓	✓					
Netware							✓		✓			
AppleTalk							✓		✓			
PABX	✓	✓						✓	✓			
Hubs			✓			✓	✓					
Wireless Access Points			✓			✓	✓					
Switches			✓			✓	✓					
Routers			✓			✓	✓					
Remote Access Servers									✓			
Terminal Servers					✓		✓					
Media Gateway	✓		✓									
Firewalls							✓		✓			✓
Network IDS							✓		✓			✓

	Business Service											
Service Resource	**Voice**	**VTC**	**Streaming Video**	**Broadcast Video**	**Tele-metry**	**Transaction**	**Office**	**Help Desk**	**Tele-working**	**Intranet**	**Extranet**	**Public Internet**
Internal Transport Resources												
10/100/1000/10000 Ethernet	✓			✓	✓	✓	✓		▪			
Wireless Ethernet					✓	✓	✓		✓			
Bluetooth	✓				✓							
Fiber Channel			✓				✓					
DECT	✓						✓					
Token Ring						✓	✓					
FDDI					✓	✓	✓					
ATM				✓	✓	✓	✓					
RS232/Centronics/Asynch					✓							
External Network Resources												
DNS									✓			
Mail Servers									✓			
Web Servers									✓			
Network Storage									✓			
Public Internet									✓			✓
IP VPN									✓			✓
Transparent LAN									✓		✓	
PSTN	✓								✓			
Conference Bridge	✓								✓			
Voice VPN	✓								✓			
Gateway VOIP	✓								✓			
Certificate Authority						✓	✓	✓	✓			
Centrex	✓								✓			
External Transport Services												
Ethernet											✓	
Wireless Ethernet									✓			
X.25					✓							
ISDN	✓				✓	✓			✓		✓	✓
Cable	✓								✓			✓
WLL (LMDS)									✓			✓
Leased Lines					✓	✓			✓		✓	✓

Service Resource	Business Service: Voice	VTC	Streaming Video	Broadcast Video	Tele-metry	Transaction	Office	Help Desk	Tele-working	Intranet	Extranet	Public Internet
xDSL	✓				✓	✓			✓			✓
Frame Relay					✓	✓			✓		✓	✓
ATM	✓			✓	✓	✓			✓		✓	✓
SONET/SDH				✓							✓	✓
PON				✓								
WDM											✓	✓
Dark Fiber											✓	✓
Facilities												
Physical Security							✓			✓	✓	
Power Services							✓					
HVAC							✓					
Co-location Facilities											✓	

Table 3: Service Resources to Support Business Services under Consideration

5.2 Key Quality Indicators for Business Applications

In determining enterprise SLAs it is important to determine the Key Quality Indicators (KQI) for the applications or services and then map them to measurable Key Performance Indicators (KPI) that can be used to measure them. Therefore, for each of the business services discussed in Section 2.1.2 a KQI is assigned. The KPI for each of the services will be discussed in the following sections.

It is important to note that the KQI and KPI for each service are the minimal set required for that service. If other applications or services require underlying services with more stringent and rigorous KQI or KPI, it is those KQI or KPI that should be considered in the SLA. The SLA therefore reflects the most rigorous requirements for the KQI/KPI for all services supported by the SLA.

The KQI can be considered generically and then applied to the particular application. The generic KQI are:

Availability — The service is available for use at the time required. As a KQI, this includes all aspects of the service, physical terminal availability, network, etc.

Speech/Video Quality — The speech or video has sufficient quality such that, within the context of the application, the information can be conveyed and interpreted in an audio or visual form. Information may include inclination, expression, body language, as well as content.

Response Time
How quickly the service responds to an internal or external stimulus.

Round Trip Delay
The time taken between making a request and seeing the response. This will include, network round trip delay, client and server processing delay, and any manual intervention in the system (like servicing a work order).

Delay
The one-way delay in the system. Delay may be different in one direction from the other.

Jitter
Variation in delay over time. The period over which the delay may vary should be understood and defined.

Locking
Information is locked for read or write to ensure integrity of the data and to inform others that the information may change.

Transaction Rate
Rate that the system or service can service requests. It is important to understand how the system or service reacts when presented with transaction rates higher than the value required. Burst rates and sustained rates and their periods should be defined.

Goodput (Carried)
The amount of valid information that is processed. For voice systems, this represents the total amount of voice traffic serviced (total calls minus blocked calls). For data applications, it is the total data, minus errored data, minus lost data, minus retransmitted data.

Throughput (Offered)
The total amount of information that is offered to the system. Throughput includes all processed information, including retries and replications .For voice and data systems, this represents the total amount of traffic presented to the system, but not necessarily serviced (includes lost calls, retransmitted, and errored information).

Idle Time
The amount of time that the system or service is idle, not performing a service or request. May only be relevant at peak times and may be required to be non-zero to account for bursts of activity and future growth.

Authorization
The system is only available to authorized resources (information and personnel) at times allowed.

Confidentiality
Information can only be seen by those intended to see the information.

Integrity
Information is available as required and has not been changed from the original. Integrity includes loss of service due to Denial of Service (DOS) or system failure. Loss of integrity implies either loss of information or a change of information.

Non-repudiation
The information has come from the source shown in the information and is a valid (authorized) source.

Disk Space
Disk space is available to service requests for the duration of the service request.

Help Desk A Help Desk is available to handle information requests. Requests may include information about the service, support, or other information.

Training Training is sufficient to perform the required task, including the use of services, responsibilities of users and providers, and any relevant tools.

Interoperability
The service or product will inter-work with all systems and services required.

Pick-up Time
How long it takes a human to respond to a request, normally by voice or video phone.

Time to Close
How long it takes to close, to the user's satisfaction, a support or information request.

Hold Time The time a support or information request is held in a queue without being processed.

Connect Time
How long a service takes to start a service.

Graceful Degradation
When a system or service fails or is overloaded, it fails in a controlled and gradual manner, servicing the highest priority tasks for as long as possible.

Revocation Termination or recall of authorization to use a service or product.

The particular KQI for an enterprise will depend to some extent on its objectives and business rules. However, in a further attempt to generalize the KQI for the business services discussed in Section 2.1.2 (as shown in Table 4) it is important to understand that the only KQIs under consideration are those pertinent to the business service under discussion. For example, in tele-commuting, only those KQIs that represent a positive QoE in gaining access to the enterprise systems are considered as it cannot be generalized as to how or what applications the tele-commuter wishes to use. If, for example, the end user wishes to use email in a telecommuting environment, the full service would contain the KQI for both telecommuting and email as in Figure 18.

	Business Service												
Key Quality Indicators (KQI)	Voice	VTC	Streaming Video	Broadcast Video	Telemetry	Transaction	Email	Office Automation	Help Desk	Tele-working	Intranet Web Services	Extranet Web Services	Public Internet
Availability	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Speech/Visual Quality	✓	✓	✓	✓			▪	▪	✓		▪	▪	
Mobility/Coverage	✓	✓	✓	✓			▪	▪	▪	✓	▪	▪	▪
Response Time	✓	✓	✓		✓	✓	▪	✓	✓	✓	✓	✓	✓
Round Trip Delay	✓	✓			✓	✓	▪	▪			▪	▪	▪
Delay	✓	✓		▪			✓						

	Business Service												
Key Quality Indicators (KQI)	Voice	VTC	Streaming Video	Broadcast Video	Telemetry	Transaction	Email	Office Automation	Help Desk	Tele-working	Intranet Web Services	Extranet Web Services	Public Internet
Jitter				✓	✓	✓							
Locking						✓							
Transaction Rate						✓	✓	▪	✓		▪	▪	
Goodput					✓	✓		▪	✓	✓	✓	✓	✓
Idle Time						✓			✓				
Confidentiality	✓	✓			✓	✓	✓	✓		✓	✓	✓	
Integrity					✓	✓	✓	✓					
Non-repudiation	✓	✓		✓	✓	✓	✓			✓		✓	
Disk Space							✓	✓					
Help Desk		✓	✓				▪	✓		✓	✓	✓	✓
Training							▪	✓		✓			
Interoperability	✓	✓					▪	✓			✓	✓	
Time to Close									✓				
Call Hold Time									✓				
Connect Time	✓	✓	✓						✓				
Graceful Degradation	✓					✓	✓	▪	✓				
Revocation						✓	✓	▪		✓	✓	✓	

Table 4: Illustrative Key Quality Indicators for the Enterprise Business Services under Discusussion

5.3 Availability

For each enterprise application or service under consideration, it is clear that all services and applications serve a common requirement for availability. It is therefore necessary to adopt a common understanding of the term.

Although many terms exist to describe network and element availability, the term has widely different application at the service level. This leads to misunderstandings and dissatisfaction of the customer. For example, if the customer believes that "availability" means their application is running without a problem, and the service provider uses the same term to mean that the service is working (even if impaired), then a mismatch in expectations is inevitable. Similarly, if a service provider relies on multiple sub-contracting providers for pieces of a service, the lack of common performance terms makes it virtually impossible to construct a picture of end-to-end performance.

5.3.1 Basic Service Availability Calculation Formula

As outlined earlier in this document, the telecommunications industry is moving toward solutions that allow customers to use service packages to support their communication needs. These packages are built from service elements, ranging from the lower layers up to the application layers, combined in a vertical (e.g., IP over Frame Relay over SDH) and/or horizontal way (e.g., IP and Voice over ATM).

Consequently there is a need to define a methodology to measure all the features provided through a service package, in order to evaluate the service availability (SA). The service availability is expressed as a percentage (SA%) to indicate the time during which the contracted service (e.g., SVCs, PVCs, end-to-end circuits including protocols, applications, etc.) is operational at the respective Service Access Points (SAPs). “Operational” in this context means that the customer has the ability to use the service as specified in the SLA.

An unavailability event affecting the service at the SAP can be defined as an outage. The duration of this specific event is the outage interval. Ordinarily this concept is used for the service unavailability percentage (SUA%) and service availability percentage (SA%) calculation as follows:

$$SA\% = 100\% - SUA\%$$

Equation 1: Basic Service Availability Calculation Formula

Where service unavailability (SUA%) is defined as:

$$SUA\% = \frac{\sum OutAgeInterval}{ActivityTime} \times 100\%$$

Equation 2: Basic Service Unavailability Calculation Formula

5.3.2 Service Degradation Factor (SDF)

A major issue is to determine whether an event affecting the service at the SAP is causing a complete service outage (service completely unavailable) or a partial service outage (service degraded, but still available).

$$SUA\% = \frac{\sum \left(OutAgeInterval \times SDF\right)}{ActivityTime} \times 100\%$$

Where: $0 \leq SDF \leq 1$

Equation 3: Use of the Service Degradation Factor

In order to address this issue, a service degradation factor (SDF) can be used in the calculation as follows:

A list of SDF values with the corresponding event type can be defined in the SLA. This definition procedure characterizes the service in the SLA and can be stated according to the customer’s business need.

It is very important to state an agreement on the time stamp information source (e.g., UTC) related to the event that causes an outage interval.

A list of SDFs is provided as an example in Table 5:

Service Degradation Factor Agreed Value	Event Type	Duration Source (time stamped information)
1	Service fully unavailable	System 1 or System 2
0.8	Outage Type A	System 3
0.6	Outage Type B	System 4 or other
0.5	Outage Type C	System 5
-------	-------	-------
0	Service considered available	Customer happy with it

Table 5: SDF Examples

Table 5 presents the relationship between Billing, Availability, and Degraded Performance. SDF enables the definition and orderly application of measures for degraded performance.

<table>
<tr><td rowspan="2">Billing</td><td>Full Billing</td><td>Acceptable Performance</td><td rowspan="2">Available</td></tr>
<tr><td>Conditional Billing</td><td>Degraded Performance</td></tr>
<tr><td rowspan="2">No Billing</td><td rowspan="2">No Billing</td><td>Unacceptable Performance</td><td rowspan="2">Unavailable</td></tr>
<tr><td>No Performance</td></tr>
</table>

Table 6: Relationship of SDF (shaded section), Billing, and Availability

5.4 Quality Equations and Measurement

There has been much research in the development of standard equations that provide quality measurements from performance-related data. These equations can be used to model a network before it's deployed, assign values for an SLA contract, and to perform analysis of data to predict the performance enhancement or degradation due to changes in the service (e.g., change CODEC, addition of route controller, move from narrowband to broadband connections).

These equations can be used to determine thresholds and sensitivity analysis of PKI parameters for SLA monitoring and reporting.

These equations are considered empirical and are presented and discussed for illustrative purposes only. A particular enterprise should validate these models against their own environments before adopting these equations or their derivatives.

5.4.1 Application Resource Measurement

The Open Group has developed the Application Resource Measurement (ARM) model to allow applications to be instrumented to allow the performance and availability of single-system and distributed applications. These may be visible to the users of the business application and those within the IT infrastructure, such as client/server requests to a data server.

ARM establishes transactions that are meaningful within the application. Typical examples are transactions initiated by a user and transaction with servers. As shown in Figure 24, applications on clients and/or servers call ARM when transactions start and/or stop. The agent in turn communicates with management applications, as shown in Figure 25, which provides analysis and reporting of the data.

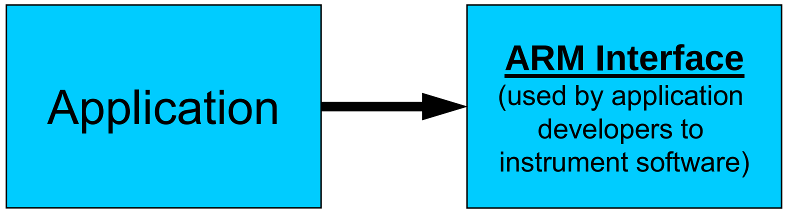


Figure 24: Application – ARM Interface

The management agent collects the status and response tie, and optionally other measurements associated with the transaction. The business application, in conjunction with the agent, may also provide information to correlate parent and child transactions.

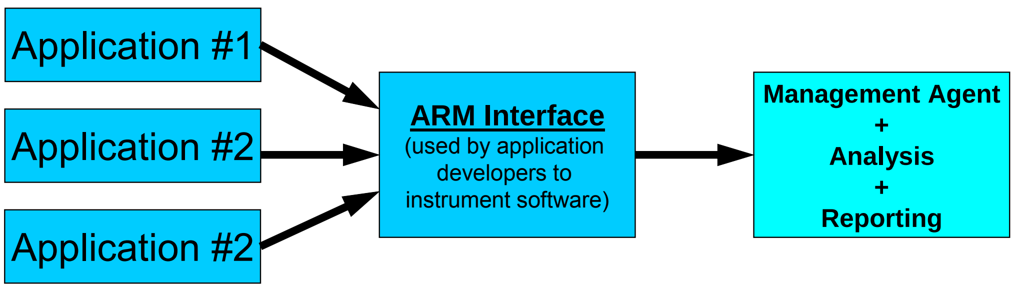


Figure 25: Application – ARM – Management System Interaction

Several versions of ARM have been developed to allow application program interfaces (APIs) in C (V1.0/V2.0) and Java (V3.0). Version 4.0 combines both the C and Java API and is currently under review.

5.4.2 E-Model

The standard measurement for user perception of voice call quality is the Mean Opinion Score (MOS). While MOS is useful and certainly valid, it is subjective and not easy to measure.

In an effort to remedy this, the International Telecommunications Union (ITU) has developed the E-Model standard (the ITU G.107 and G.108 standards) as a means of objectively measuring call quality. The output of an E-Model is called an R-value, with a value of 0-100. This has been shown to reliably map to an estimated MOS.

The E-Model was developed by ETSI to take account of all the impairments that led to speech degradation, and including impairments that typically occur in packet-based networks. In the E-Model, impairment values are assigned to a number of independent parameters, which are then combined to give a transmission rating factor R as follows

$$R = R_o - I_s - I_d - I_e + A$$

R_o Represents in principle the basic signal-to-noise ratio, including noise sources such as circuit noise and room noise.

I_s Combination of impairments which occur more or less simultaneously with the voice signal. This includes loudness, sidetone, and quantizing distortion from analogue to digital conversion. This will include impairment by packet loss

I_d Represents the impairments caused by delay, which includes Talker and Listener echo, and end-to end delay

I_e Impairments caused by low bit rate CODEC. This includes effects from packet jitter and loss

A Advantage factor to compensate for impairment factors when there are other advantages, such as mobility

Equation 4: R-Value Calculation

To ensure high speech quality, the following KPI measurements should be met:

1. Delay – end-to-end packet delivery delay
2. Jitter – variations on packet delivery times
3. Packet Loss – percentage of packets dropped during transmission
4. CODEC Selection
5. Trans-coding stages

For packet-based technologies, the MOS (R-value) can therefore be determined from the E-Model by measuring the KPI detailed in Table 7.

KPI	**IP (VoIP)**	**ATM (VoATM)**
Jitter	IPDV	CDVT
Delay	IPTD	CD
Loss	IPLR	CLR

Table 7: KPI for Voice Speech Quality for Packet-Based Technology

5.4.3 NetForecast Response Model

NetForecast (www.netforecast.com) have developed a response model for Internet Protocols that can be used to both predict and measure performance. Verification with measurement data provides good agreement to +/- 10%. This model is shown in Equation 5.

$$R = 2(D + L + C) + (D + C/2)((T - 2)/M) + D\,\mathbf{ln}((T - 2)/M+1) + \mathbf{max}(8P(1 + OHD)/b, DP/W)/(1- \mathbf{sqrt}(L))$$

Where:

B = Minimum line speed in path (bits per second)

$C = C_C + C_S$

C_C = Client processing time (seconds)

C_S = Server processing time (seconds)

D = Round trip delay (seconds)

L = Packet loss (fraction)

M = Multiplexing factor

OHD = Overhead (fraction)

P = Payload (bytes)

R = Response time (seconds)

T = Application Turns

W = Effective Window size (bytes)

Equation 5: Response Time for an Application due to NetForecast

From Equation 5, it is clear that the NetForecast model depends on a number of factors:

- Application design (number of turns between client and server)
- Client processing time (depending on processor load, processor speed, and client application design)
- Server processing time (depending on processor load, processor speed, and server application design)
- Payload (dependent on application design and request made)
- Effective windows side (network, client, and server configuration)
- Packet Loss, Round Trip Delay, Line Speed (dependent on network characteristics and performance)

Therefore, to determine the response time of an application will require KPIs from the network services to be combined with KPIs for the servers and clients. ARM is a possible candidate to measure client and server response time. Under certain circumstances, it is possible to assume that some characteristics (e.g., client processing time) remains constant so need not be measured.

Equation 5 can be used to:

- Predict performance of networked applications
- Determine KQI for networked applications for the SLA monitoring

- Allow 'what-if' analysis for network changes; e.g., Content Distribution Networks or Route Controllers[3]

5.5 TCP-Based Application Models

There has been extensive work on developing empirical models for the TCP protocol; for example, Padhye et al.[4]

$$B(p) \approx \min\left(\frac{W_{\max}}{RTT}, \frac{1}{RTT\sqrt{\frac{2bp}{3}} + T_o \min\left(1,3\sqrt{\frac{3bp}{8}}\right) p\left(1+32p^2\right)} \right)$$

$B(p)$: TCP throughput depending on packet loss probability (packet/sec)

$W_{\max}$: Maximum buffer size of receiver (packet)

RTT: Round Trip Delay (2xOne Way Delay for symmetric networks)

b: Round Trip Delay (2xOne Way Delay for symmetric networks)

p: Packet Loss Probability (One Way Packet Loss)

T_0: Initial retransmit timeout period

Equation 6: Empirical TCP Reno-Based Application Throughput

Again it is likely that some of these parameters will be constant for a particular application or service and can be eliminated. Care must be taken in the measurement of these parameters to ensure that the act of measure:

- Represents the true measure of the application and service rather than a standard test packet sequence
- The act of measuring does not impact on the performance of the service itself, especially under error conditions

Equation 6 estimates the likely effective throughput (sometimes called goodput) that an eager transmitter will obtain when sending an arbitrary long file to an eager receiver. It is a measure of the impact of network quality on the expected throughput for long-lived bulk data transfers. The time taken to reach this effective throughput also depends on the network characteristics and maximum throughput cannot be obtained for small transfers (e.g., < 10-20kbytes over transcontinental distances), the majority of web transfers are less than this size.

Bulk transfer is by no means the sole use of TCP; TCP is often used in network applications to provide a reliable stream transfer for:

- Control channels

[3] Improving User Experience with Route Control, Sevcik and Bartlett, NetForecast Report 5062, July 2002.
[4] Modeling TCP Throughput: A Simple Model and its Empirical Validation, Padhye et al, Sigcomm 1998.

- Terminal sessions
- Remote desktop
- Remote procedure calls
- Database access
- File sharing

In these systems the effect of delay and loss is both more difficult to quantify and at the same time potentially less damaging to the user's QoE. For example, in the case of web-based RPC where the data transferred is small (about one packet each way) then up to 50% of the packets (about 10% of the load) are 'redundant' in the sense that their reception is not necessary to complete the transaction (from the user point of view). These packets are transmitted purely to establish and open the congestion window used by the TCP protocol in preparation for future bulk data transfer. This suggests that there are potential gains to be made by treating distinct elements of the TCP protocol differentially.

5.6 Erlang Equations for Voice Systems

The sizing of networks to support voice (speech) calls has been the subject of much theoretical work and well understood with over a century of experience. Much of this work is due to Agner Karup Erland who developed the equations in the early part of the 1900s to predict the number of trunks needed to support the requirements of the plain old telephone (POTS) network. As with the E-Model and the Response Model, the Erlang equations and their derivatives can be used to predict performance characteristics of the network and to also measure performance from measurements such as Busy Hour Call Attempts (BHCA) and Blocking Probability.

A voice system is in one of four states:

- Idle: No calls present, no trunk or packet capacity used.
- Queued: Acall has been made but it is held in a queue. The call is using trunk or packet capacity that is not available for other calls.
- In Service: A call has been made and is being serviced by a user (talking!).
- Blocked: All capacity is used such that it can't be serviced.

The equations that govern the statistical behavior of voice systems are shown in Table 8 and depend on the number of sources (callers) to the system. Large is generally regarded as >10. For a normal office system, it is assumed that all calls are either completed or lost, there are no queues, and they are governed by the Erlang B or Engset. These use the average call rate and call duration to determine the traffic load (in Erlangs) and then either predict the number of servers required for a given blocking probability or the blocking probability for a given number of servers. If silence suppression techniques are used, these equations will overestimate the required bandwidth depending on the amount of silence of the conversation. This can depend on both the nature of the call (e.g., person-to-person or conference call) as well as the particular language and culture of the participants. The use of silence suppression can therefore drastically change the bandwidth requirements. A measure of the bandwidth estimate that includes silence suppression is as follows:

$$B_S = (1 - S) \times B_0$$

Equation 7: Bandwidth Modification in Consideration of Silence

Where B_S represents the modified bandwidth and S represents the fraction of silence in a conversation. Silence suppression may be catered for directly by the CODEC employed (e.g., G.729a) or at a higher layer where this is not supported by the CODEC (e.g., G.711, G.721).

In a similar manner, the protocol overhead of the voice transport must be considered. These may be insignificant for TDM services or significant for packet-based technologies like Voice over IP (VoIP) or VTOa unless header compression techniques – Compressed Real-Time Protocol (CRTP) or Robust Header Compression (ROHC) – can be employed.

In a call center there are more servers (trunks) than clients, so it is possible that a call can be held in a queue or delayed. The equations governing these requirements are covered by the Erlang C, Delay model, or Poisson or Binomial distribution models as shown in Table 8 and Equation 8.

	Source Population	
Block Traffic Behavior	Infinite (Large)	Finite (Small)
Lost	Erlang B	Engset
Held	Poisson	Binomial
Delayed	Erlang C	Delay

Table 8: Voice Traffic Characteristic Equations

Erlang B Model

$$P(n) = \frac{\frac{A^n}{n!}}{\sum_{x=0}^{N} \frac{A^x}{x!}}$$

Erlang C Model

$$P(n) = \frac{\frac{A^n}{n!}\left(\frac{n}{n-A}\right)}{\sum_{x=0}^{n-1} \frac{A^x}{x!} + \frac{A^n}{n!}\left(\frac{n}{n-A}\right)}$$

Binomial Model

$$P(n) = \frac{A^n e^{-A}}{n!}$$

Engset Model

$$P(n) = \frac{\left(\frac{S!}{n!(S-n)!}\right)a^n}{\sum_{i=0}^{N}\left(\left(\frac{S!}{i!(S-i)!}\right)a^i\right)}$$

Poisson Model

$$P(n) = \frac{A^n e^{-A}}{n!}$$

$P(n)$	Blocking Probability for Server n
S	Sources
N	Servers
A	Offered Traffic
a	Idle Offered Traffic

Equation 8: Traffic Queuing Models

5.7 Enterprise Business Services Key Quality Indicators

For each of the business services introduced in Section 2.1, the KQI will be further discussed and then the supporting KPI determined. Both the KQI and KPI may be included in the SLA. The intent is not to specify actual values since these may be considered commercially sensitive, subject to negotiation between the provider and the customer, and also dependent on the application supported.

5.7.1 Voice Services

Voice services are an extremely important component in enterprise systems today. Voice systems fall into thee distinct categories:

1. PABX or residential type services for normal voice communications, internal and external to the organization
2. Call Center type services where it is assumed that a queue of incoming calls will be maintained
3. Fast Response or Emergency Call Centers where it is assumed that a queue is undesirable

The main difference between the first and the other two services in that for optimum performance the ‘normal’ operation implies no blocking on average, that every call can be completed, whereas a call center should always have blocking on average, with calls held in a queue and where the call queues are never quite empty for optimal operation. The parameterization of these two services, dependent on voice, is sufficiently different to warrant separate consideration. It should also be considered that for either of the two types, other support services – e.g., web services – may be required to fulfil the enterprise requirements.

The network services to support voice also fall into two categories: circuit-based and packet-based. Circuit-based services rely on a TDM-style service consuming the maximum required bandwidth for the duration of a call. The number of concurrent calls in therefore dependent on the number of available circuits. Packet-based services are generally supported as IP packets or ATM Cells (AAL1 or AAL2 encapsulation) using ITU H.323 or IETF SIP standards. AAL2 and IP encapsulation have the benefits that they allow statistical gain from silence suppression (only sending when voice present), a choice of CODECs for improved bandwidth utilization, and, in the case of AAL2, statistical multiplexing by combining several packets from different calls into the same ATM cell. A representative choice of CODECs is shown in Table 9.

The efficiencies that packet-based voice services, together with the benefits of converged voice and data services, provide revenue generation and cost savings opportunities for the enterprise and as such should be parameterized, monitored, and reported in the SLA.

Standard	Coding	Bit rate (kbit/s	MOS	MIPS/ Channel	Fax Rate (kbit/s)	Packing/ Processing Delay (ms)	Tandem HOPS
G.711	PCM	64	4.2	0.1	56	0.125	14
G.721	ADPCM	32	4	10	4.8	0.125	4
G.726	ADPCM	16	4	12	14.4	0.125	4
G.728	LD-CELP	16	4	33	14.4	0.625	3
G.729	CSA-CELP	8	4	22	7.2	15	2
G.729a	CSA-CELP	8	4	13	4.8	15	2
G.723.1	A_CELP	6.3	3.8	16	0	37.5	2
	MP-MLQ	5.3	3.7	18	0	37.5	1
GSM-EFR	LD-LTP	13	3.7	5	7.2	20	2?
DoD FS1015	LPC-10	2.4	Synthetic	10	???	22.5	1?
DoD FS1016	CELP	4.8	Synthetic	21		15	1?
DoD FIPS 137	MELP	2.4	Synthetic	32		45.5	2
STANAG 4209	CSVD	16				20	4?

Table 9: CODEC Characteristics

Packet-based technologies, especially VoIP, are typically connectionless and as such subject to end-to-end packet loss. In the case of IP, this packet loss without traffic engineering could be considerable, perhaps as much as 10% or more at peak times in the public Internet. Typically this is of the order of a few % and can be managed with traffic engineering and congestion control techniques that will prioritize voice packets (that consume relatively little bandwidth) over more bandwidth-intensive, but less sensitive to delay and retransmission applications like email.

Packet loss can vary during the duration of the call and as packet loss is a KPI for voice quality (PQI) as measured by the MOS score, voice quality can also vary considerably during the call. Although not totally immune, circuit-based calls produce a more constant MOS during a call. This is illustrated in Figure 26.

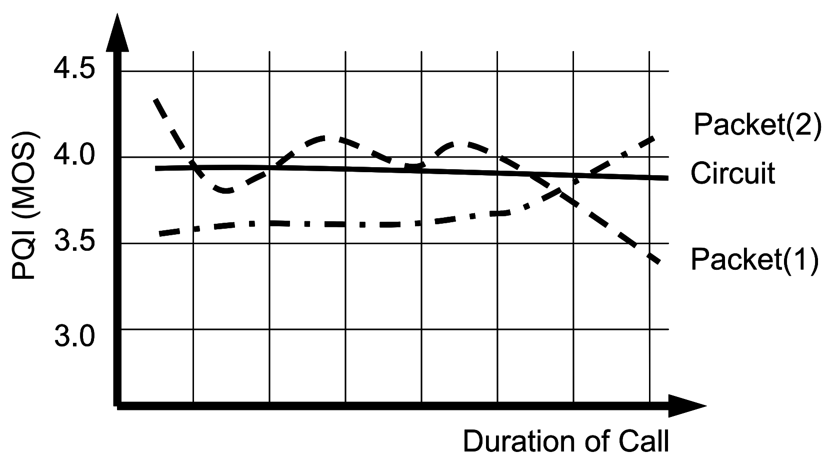


Figure 26: Variation of MOS During a Voice Call

With reference to Figure 26, the line shown as Packet (1) has an average MOS during the call compared to the circuit-based call example, but ends on a low value of MOS. Because of the physiological effect of human perception, the quality at the end of the call will have more effect on the QoE than the MOS at the beginning of the call. For this reason, the variation shown in Packet (2) provides on average a lower MOS, but ends the call with improved quality. As a consequence, it is likely that the perceived QoE for a voice call is likely to be considered higher for Packet (2) than Packet (1). Variation of MOS during a call is therefore as important for packet-based technologies as the absolute value of MOS. Lower, but less variable MOS may provide better user experience than higher voice quality with greater variation.

If the services are to be used for Fax or Modem (dial-up) access, then some CODECs are unsuitable. For the SLA it will therefore be important to determine whether Fax or Modems are to be used and what the average duration of each call is likely to be. The duration of the Fax or Modem call is important in that packet voice technologies gain benefit from silence suppression (not sending packets when no speech) in statistical gain, which isn't possible with a Fax or Modem call. These capabilities can be built into the design and the data converted to packets at an earlier stage.

5.7.1.1 Voice Service Key Quality Indicators

From Table 4, the KQIs for voice services are:

Availability
: The voice system should be available sufficiently such that non-availability does not impact the enterprise objectives. This may vary across the system; for example, call center phones probably need to be more available than an internal call only phone. Where silence suppression is in use, comfort noise may need to be inserted to provide assurance that the call is still in progress.

Speech Quality
: Speech quality should be sufficient to perform the task at-hand. High noise environments – such as manufacturing and military environments – require intelligibility of information rather than an understanding of expression of information. Call Centers require high quality in equipment and also a reassuring tone, expression, and accent from the operator. In some instances speech quality can be traded for another attribute, such as mobility.

Response Time
: When the voice system is used, the system should respond in a time sufficient that it does not impact on the performance of the user in the execution of his/her objectives.

Round Trip Delay
: In a conversation, if the round trip delay is too long, conversations can overlap, reducing intelligibility.

Confidentiality
: Voice communications are generally between known and valid individuals or groups (conference calls). The information relayed has a sensitivity that may be personnel, commercially, politically, or militarily sensitive. Confidentiality means must be applied as appropriate to the information to be protected.

Non-repudiation
: It is an implicit requirement that the corresponder and correspondee in a voice communication are who they pertain to be. Proof of identity will depend on the sensitivity of the information relayed in the conversation.

Interoperability
: It is an implicit requirement that phone calls can be made to (with exceptions) phone systems worldwide. The interoperability should not affect the availability, quality, confidentiality, and non-repudiation services.

Connect Time

Connect time includes: picking up handset (or CTI equivalent), receiving dial tone, dialing number, end user phone ringing, and end user picking up. If any of these components become too long, the QoE of the caller will be negatively impacted.

Graceful Degradation

Voice is an important system and it is therefore important to understand who it will degrade under load or distress (failure). For example, most PBX systems must pass emergency calls even under power failure and military communications must pass the highest priority messages for the longest time possible.

From the KQI, the KPI for voice services must be determined so that they can be documented and agreed in the SLA and monitored in operation for reporting, error prevention, and rectification.

Speech Quality is determined for voice phone systems using MOS or the R-Value. These can be determined by transmitting standard phone messages through the system periodically, but this has the disadvantage of not having a relation with real calls.

These should be measured over the duration of the call to ensure the MOS resides within the permitted range.

5.7.2 Video Teleconferencing

Video teleconferencing (VTC) has become a popular enterprise service, reducing travel costs and adding value over normal telephone conference facilities.

Most VTC facilities today employ the ITU H.320 series of standards as they allow variable bandwidth for improved video quality and can be supported on either circuit (ISDN) or packet (IP or ATM) technology. It is possible to allow inter-working between the different underlying technologies and to allow conference video, video and audio, or audio only.

The ITU-T Video Quality Expert Group (VQEG) and ITU-T SG9 are focused on video quality issues. Recommendation N.64 defines five grade scales (1,2,3,4,5) that can be quality-based (bad, poor, fair, good, excellent) or impairment-based (very annoying, annoying, slightly annoying, perceptible but not annoying, imperceptible). These scales can be used for subjective assessment of video quality. Objective video quality measurement techniques have also been addressed in Recommendation J.143/J.144.

In some cases, separate transports are used for voice and video; for example, IP video and ISDN voice. This allows some participants to enter the conference as voice only, but adds a further complication that synchronization between the two services, over two different transports, is necessary.

5.7.2.1 VTC Key Quality Indicators

As VTC, in many respects, is seen as an evolution over voice systems, many of the KQI for voice systems, as documented in Section 5.7.1, are also valid for VTC. As VTC technology is not as pervasive or the equipment as portable (at present), it is assumed that graceful degradation is not a KQI. However, help desk is added as it is often required to establish connections between different vendors' equipment. These KQIs may mature to those of voice with time and further adoption.

For point-to-point calls, end systems can talk to one another without any intervening equipment other than the network. Availability is then based on:

- Availability of the VTC equipment
- Availability of the network
- Scheduling of the equipment at each end of the call

For multi-party conferencing, video and audio, or just audio, requires a conference bridge. This may be hosted by one of the enterprises, by a service provider, or by a third party.

Introducing a conference bridge to the system means that resource availability in the bridge will also affect availability. Traffic and blocking probabilities will be governed by the same equation as discussed in Section 5.7.1.

As with voice CODECs, video CODECs (H.261, H.263) suffer from poor picture quality due to:

- Jitter
- Delay
- Packet Loss

Audio and video are handled as separate sessions within the protocols (RTP/UDP). Packets from the audio and video sessions are necessary to allow speech and lips to be synchronized.

Picture quality can be traded for decreased frame rate and frame rate has a dependency on available bandwidth, choice of CODEC, and the complexity of the picture being sent.

For an adequate KQI of picture quality, adequate bandwidth must be available to allow sufficient frame rate for the complexity of the picture and CODEC available or chosen.

5.7.3 Streaming Video/Audio

Streaming services differ from traditional video services in that they are predominately packet (IP)-based, can be multicast, and (like broadcasts) are half duplex. For IP systems, the video is transported using RTP/UDP. A further protocol, RTSP, allows control of the video stream for start, pause, play, etc. RTP ensures the timely delivery of information by QoS guarantees or by relying on lower-layer services such as ATM VBR or CBR. A further protocol, RTP Control Protocol (RTCP), periodically transmits control packets in the multicast session to provide feedback on the quality of the data distribution.

Popular CODECs for streaming services are MPEG-2 or MPEG-4. Again, CODECs are susceptible to jitter, delay, and packet loss.

Frame rate should be sufficient to display video content adequately. This will depend on available bandwidth, CODEC efficiency, and picture size (QCIF, CIF). Most CODECs for streaming services have been designed to be run on software using general-purpose (desktop) computers. As such, CPU load on the client (and server) increases with the number of concurrent sessions and picture complexity.

Although control through RTSP does not require the bandwidth required for high-quality video, response time on the request will be an important KQI for control.

Multiple services can be serviced as either separate streams or as multicasts. The bandwidth requirements for the aggregate services would therefore depend on the method of supporting multiple services (unicast or multicast) and, in the case of multicast, the location of the clients.

It is important to recognize the difference between real time, with a near constant delay from the streaming source, and non-real time, where the streaming source can be accessed on an *ad hoc* basis. An example of the first instance is for broadcast services (e.g., www.bbc.co.uk) and the latter being video on demand.

Video quality can be defined using five point scale definitions from the ITU-T from both subjective or objective means as discussed in Section 5.7.2.

5.7.3.1 Streaming Video/Audio Key Quality Indicators

Streaming services are in their infancy; therefore, it is assumed that the KQIs for such services are less rigorous than those of voice; for example:

Availability — The streaming should be available sufficiently such that non-availability does not impact the enterprise objectives.

Video/Audio Quality
Video and audio quality should be sufficient to perform the task at-hand or to gain the required satisfaction. For example, streaming stock quotations should be readily intelligible, whereas a Video on Demand service must have sufficient quality to both command a premium price and potentially differentiate the offering from its competition. In some instances quality can be traded for another attribute, such as mobility.

Response Time
If the system is interactive, using the RTSP protocol, the response time for such requests should be such that repeat requests are not made (because the user thinks the request is lost) or achieves desired aim (rewind to portion required).

Help Desk — Being relatively new technology and not in widespread use, help desk support is important, especially where client software has to be installed and configured.

5.7.4 Broadcast Video

Digital Broadcast Video is an important enterprise application for the entertainment and information distribution industry. The formats – generally ITU-R 601 (TV) or ITU-R 709 (HDTV) – provide a high-quality (broadcast quality) picture for sending video feeds from outside broadcast facilities (sports and news events), between regional broadcast stations, and to transmitters for analog transmission or satellite uplinks. As shown in Table 10, the bandwidth requirements to support these raw formats are high.

Video quality can be defined using five point scale definitions from the ITU-T from both subjective and objective means, as discussed in Section 5.7.2.

5.7.4.1 Broadcast Video Key Quality Indicators

Availability — The system must be available for use at the time and place required and for the duration of the broadcast.

Video Quality
The video quality must be appropriate for the task at-hand. Sports broadcasts for public distribution require higher quality than home video ‘blooper’ sequences. Unmanned Aerial Vehicles (UAV) for surveillance or military applications may trade spatial resolution (coverage) for video quality as appropriate.

Jitter — For commercial broadcast systems, it is often required to multiplex video feeds from different sources so that local content can be inserted. It is therefore important that time of arrival of a video broadcast does not vary markedly (jitter).

The video formats can be transmitted over short (2km) distances over fiber, but for longer transmissions they are encoded using a transport protocol like SONET/SDH, ATM, PON, or, in metropolitan networks, Gigabit Ethernet. Being dedicated links, congestion is not at present an issue.

Using CODECs can reduce bandwidth requirements but reduces picture quality so is avoided except at customer distribution points (transmitters, uplinks).

Timing is important so that different video feeds can be multiplexed to allow them to be interleaved and enable the insertion of local content (news, adverts, etc.). In this instance, jitter is more important than delay unless the video feed is meant to be interactive (for example, at a sports event where the studio host holds a conversation with the outside broadcaster).

	QCIF	**CIF**	**ITU-R 601 (TV)**	**ITU-R 709 (HDTV)**
Pixel/row (Y)	176	352	720	1920
Number of Rows (Y)	144	288	576 (480)	1080
Pixel/Row (U,V)	88	176	360	960
Number of Rows (U,V)	72	144	576 (480)	1080
Aspect Ratio	4:3	4:3	4:3	16:9
Frame Rate	5-15	10-30	25(30)	25(30)
Bits/frame (Kbyte)	38.02	152.1	829.4 (691.2)	4424 (3686)
Bits/Image (Mbyte/s)	0.84-3.8	10.1-30.4	165.9	884.7

Table 10: Video Standard Characteristics

Broadcast video therefore requires high availability (at the time of transmission) networks with a low BER to produce a high-quality (KQI) image.

5.7.5 Telemetry

Telemetry is an important service for automation of monitoring services like utility (Automatic Monitoring Requests), remote control (manufacturing robot and extra-terrestrial satellite), and telemedicine where medical diagnostics can be routinely collected.

SCADA (supervisory control and data acquisition) is a system that allows the gathering of data in real time from remote locations in order to control equipment and conditions. SCADA is used in power plants as well as in oil and gas refining, telecommunications, transportation, and utilities.

Information may be either synchronous or asynchronous and may use dedicated connections, leased line services (including ATM Circuit Emulation), and Internet protocols.

5.7.5.1 Telemetry Key Quality Indicators

To be added in a later revision.

5.7.6 Transactions

Transactions are becoming increasingly used in many enterprise systems today. Examples include:

- Credit Card transaction processes
- Stock trading
- Automatic Teller Machine transactions.
- Electronic Funds Transfer
- Ticket Systems (airline, travel agents)
- Work Orders
- Online Auction Sites

There are varying loads and volumes for each of the different enterprise transaction applications but all must consider call volume, what to do in excessive load, and how to handle fairness to ensure that multiple requests are handled appropriately (depending on policy). As many transactions are a many-to-one relationship (e.g., travel agents claiming a particular plane seat), it is important to have a locking mechanism to ensure correct operation. How the system responds to lost or inappropriate locks is an important service requirement. In some instances it is important to store all transactions in a separate file store for verification purposes by regulator bodies (e.g., stock trades). Security, and ensuring integrity and authenticity of the data are important.

5.7.6.1 Transaction Key Quality Indicators

Availability — The transaction system should be available sufficiently such that non-availability does not impact the enterprise objectives.

Response Time — When the transaction is issued, the system should respond in a time sufficient that it does not impact on the performance of the user in the execution of his/her objectives.

Round Trip Delay — In a transaction, if the round trip delay is too long, transactions can overlap, creating confusion.

Jitter — Some transactions may be used in a competitive environment together with other transactions; e.g., stock trading. In these types of environments, it is commonly important that each transaction is treated fairly. Therefore, the absolute delay through the system is not as important (because it can be compensated for) as jitter, variation in delay.

Locking — For some transactions (e.g., booking seats on a flight) it is necessary to 'lock' the appropriate record while the transaction is processed for uniqueness.

Transaction Rate — The transaction rate should be sufficient to handle normal loads without affecting other KQIs.

Goodput The system should be applied to process all requests up to the maximum transaction rate.

Idle Time A measure of the efficiency of a system can be measured by the idle time of the system. This can be considered for providing extra transaction capacity or as a measure of optimizing the cost of the system.

Confidentiality
Voice communications are generally between known and valid individuals or groups (conference calls). The information relayed has a sensitivity that may be personnel, commercially, politically, or militarily-sensitive. Confidentiality means must be applied as appropriate to the information to be protected.

Integrity Many transactions have a high value, in both commercial and human terms. It is therefore important to ensure that the message in the transaction is not lost or has been altered from the original. In some instances, like money transfers, this may need to be verifiable.

Non-repudiation
It is an implicit requirement that the corresponder and correspondee in a transaction are who they pertain to be. Proof of identity will depend on the sensitivity of the information relayed in the transaction.

Graceful Degradation
It is important to understand how the transaction system will respond under load or distress (failure). For example, if the transaction rate is exceeded, it is likely that goodput should not be effected and an understanding of how fairness for the oversubscribed transaction rate achieved.

Revocation The ability to use the transaction system can be terminated or revoked. This may need to be verified.

5.7.7 Electronic Mail/Messaging

After voice, electronic mail (email) and instant messaging services like SMS and IRC have become the most important network applications in terms of network usage today.

Email standards include SMTP/POP3 and X.400 messaging. The principals are the same in both instances, but the SMTP/POP3 standards will be discussed further because of their wider adoption, in line with the uptake of Internet applications during the 1990s.

Instant message services like IRC are different from traditional email in that messages can only be sent if the recipient is ‘on-line’ (connected to the network service and visible to the sender). Email and SMS use a store and forward mechanism such that the sender sends the message on a client (User Agent) and sends it to the server (Message Transfer Agent). The recipient then retrieves the message using his/her UA.

As is normal in most environments, email clients are assumed to be part of an office suite like StarOffice or Microsoft Office. It is therefore assumed that some KQIs, such as response time, are covered under the Office application. Similarly, training, availability of help desk and interoperability are not considered KQIs because of the maturity of the technology.

5.7.7.1 Electronic Mail/Messaging Key Quality Indicators

Availability The messaging system, both server and client, should have sufficient availability such that non-availability does not impact the enterprise objectives.

Delay The delay sending or retrieving the message to or from the server. This should be sufficient so as not to impact enterprise objectives and message failure messages can be responded to in a time that does not impact on enterprise objectives.

Transaction Rate
The transaction rate of processing message requests (MTA) should be sufficient to handle normal loads without effecting other KQIs.

Confidentiality
Messages are generally between known and valid individuals or groups. The information relayed has a sensitivity that may be personnel, commercially, politically, or militarily sensitive. Confidentiality means must be applied as appropriate to the information to be protected.

Integrity In some countries, emails can carry a contractual obligation. It is therefore important to ensure that the message is not lost or has been altered from the original. This will include messages in transit (between client and server) and in storage on the client or server. In some instances, this may need to be verifiable.

Non-repudiation
It is an implicit requirement that the corresponder and correspondee in a message are who they pertain to be. Proof of identity will depend on the sensitivity of the information relayed in the message.

Disk Space The MTA and UA should have sufficient disk space to store all required messages to fulfil enterprise objectives, even in the presence of stored messages not required to fulfil enterprise objectives.

Graceful Degradation
It is important to understand how the message system will respond under load or distress (failure). For example, if the transaction rate is exceeded, it is likely that goodput should not be effected and an understanding of how fairness for the oversubscribed transaction rate achieved.

Revocation The ability to use the messaging system can be terminated or revoked. This may need to be verified.

5.7.8 Office Automation

Office automation tasks form a vital and routine part of any enterprise operation and include word processing, spreadsheets, databases, and presentation packages. Most departments will use such systems and may be integrated into their ERP or CRM systems. The facilities, usability, and the availability of such systems is therefore vital to the operation of the enterprise.

5.7.8.1 Office Automation Key Quality Indicators

Availability The system, both server and client, should have sufficient availability such that non-availability does not impact the enterprise objectives.

Response Time

The response time of the system should be such that it doesn't affect the effectiveness of the user. This will include response from the terminal (echoing key strokes, scrolling texts) and the server (displaying files, folders) and both (opening, closing, and saving files). This should be independent of the number of users of the system, up to the maximum number of users.

Confidentiality

Information developed and displayed by the Office Automation applications that have a sensitivity that may be personnel, commercially, politically, or militarily sensitive. Confidentiality means must be applied as appropriate to the information to be protected.

Integrity

In some countries, emails can carry a contractual obligation. It is therefore important to ensure that the message is not lost or has been altered from the original. This will include messages in transit (between client and server) and in storage on the client or server. In some instances, this may need to be verifiable.

Disk Space

The servers and clients should have sufficient disk space to store all required messages to fulfil enterprise objectives, even in the presence of stored information not required to fulfil enterprise objectives.

Help Desk

A help desk should be available for upgrades, problem resolution, etc.

Training

Training courses, material, and information should be sufficient to allow the enterprise users to operate at full efficiency in the execution of their tasks towards the enterprise objectives.

Interoperability

In the exchange of information, it may be necessary to exchange information between systems using different Office Automation tools and applications. It is important that such systems can interoperate without loss of information (e.g., formatting) or efficiency.

5.7.9 Help Desk

The customer care help desk is an integral part of the overall service provided by an internal or external provider. Since customers need to be well-informed about the nature of any fault, the impact of faults on their contracted service, the status of resolution, and anticipated time to restore the service to full operational status, the help desk plays a very important part in complying with an SLA contract.

An organizational service, such as the extended help desk support, enables a provider to differentiate its service offering for different types of customers. For example, a network bearer service, such as an Internet access to an Internet service provider via DSL, could be offered as a stand-alone service with a limited help desk support (e.g., 8 a.m. to 5 p.m.). A SOHO (Small Office Home Office) customer who uses the Internet connection for e-business purposes (possibly not only during usual office hours) could need, and therefore pay for, an added service providing extended help desk support.

Help desks may be driven through requests made by customers using a communication medium or may be generated asynchronously by the service or its resources in the form of events.

It is common that a priority is assigned to a help desk request as it is initiated depending on the severity of the condition or the contractual obligations of the SLA. A possible severity level is:

Critical — The priority of the request or event has a critical impact on the enterprise. Failure of backup up power supplies during a power outage would constitute a critical condition, for example.

Major — The impact of the request or event is major in that it may severely impact one or more services or products in the enterprise or may cause another service or product to enter a major or critical condition or may effect a significant population of the users. Failure of backup power supply would constitute a major failure, as redundancy for power is lost.

Minor — The impact of the request or event has minor impact on the performance of the service or enterprise. An alternative service or mechanism may be invoked to rectify the problem.

Informational
The request or event has no effect on the performance of the service, product, or user community. Enhancement requests fall into this category.

Priorities may be assigned by the help desk with or without the agreement of the user or system. The SLA for help desk will have different reaction parameters according to priority. It may also be necessary and prudent to quantify the number of events or requests at higher priority (critical, major) as significant numbers of these events or requests may be indicative or a poorly implemented or designed service.

5.7.9.1 Help Desk Key Quality Indicators

A set of time-of-day-based parameters could be negotiated between the customer and the service provider as part of the SLA for this type of service. For example, the following parameters could be part of the contract:

Availability — The voice system should be available sufficiently such that non-availability does not impact the enterprise objectives.

Speech Quality
Speech quality should be sufficient to perform the task at-hand. Help desks and call centers require high quality in equipment and also a reassuring tone, expression, and accent from the operator.

Response Time
The time to acknowledge a call, including any on-hold time.

Transaction Rate
The number of concurrent calls the Help Desk or Call Center can handle.

Goodput (Offered)
The help desk should be able to process all requests up to the maximum transaction rate.

Throughput (Actual)
The total amount of voice traffic presented to the system (including lost calls) represents a measure of the efficiency of the help desk or call center.

Idle Time — A measure of the efficiency of a system can be measured by the idle time of the system. This can be considered for providing extra capacity or as a measure of optimizing the cost of the system.

Time to Close
: The time taken to close a request. This should be such that it doesn't impact on the performance of the enterprise objectives.

Call Hold Time
: The time a call has been made but is held in a queue.

Connect Time
: The time taken for a call to be made and acknowledged by entering a queue (on hold) or being serviced.

Graceful Degradation
: It is important to understand how the system will respond under load or distress (failure). For example, if the transaction rate is exceeded, it is likely that goodput should not be effected and an understanding of how fairness for the oversubscribed transaction rate achieved.

It is important to consider as KQIs both metrics pertaining to the performance of the center – for example, call duration, pick-up time, and operator utilization – and those to the value of the information provided in the call. Metrics to measure this value will be application-specific, but one such metric would be the number of calls per call reference, implying that the number of calls that are resolved in the first call is a measure of both the efficiency of the operators and the value of the information provided. Care must be taken to balance call duration with minimizing repeat calls.

5.7.10 Tele-Working

Tele-working is becoming an important facility in the operation of many enterprises. The same facilities available within the enterprise should be made available to tele-working. As physical security measures cannot be employed to ensure that the user is not masquerading as a valid user, higher-level security authorization measures are generally required and employed to ensure a higher level of confidence of the authenticity of the user.

The tele-working KQIs only refer to the access to the enterprise system. If the applications run in the tele-working environment, then the KQIs of those applications must also be considered in the tele-working SLA.

5.7.10.1 Tele-Working Key Quality Indicators

Availability
: The access system, both server and client, should have sufficient availability such that non-availability does not impact the enterprise objectives.

Response Time
: The response time of the system should be such that it doesn't affect the effectiveness of the user. This will include response from the terminal (echoing key strokes, scrolling texts) and the server (displaying files, folders) and both (opening, closing, and saving files). This should be independent of the number of users of the system, up to the maximum number of users.

Goodput
: The system should be applied to provide sufficient goodput so as not to affect the efficiency of the user.

Confidentiality
: The information developed and displayed by the Office Automation applications that have a sensitivity that may be personnel, commercially, politically, or militarily

sensitive. Confidentiality means must be applied as appropriate to the information to be protected.

Non-repudiation
It is an implicit requirement that the corresponder and correspondee in a message are who they pertain to be. Proof of identity will depend on the sensitivity of the information relayed in the message.

Help Desk A help desk should be available for upgrades, problem resolution, etc.

Training Training courses, material, and information should be sufficient to allow the enterprise users to operate at full efficiency in the execution of their tasks towards the enterprise objectives.

Revocation The service can be terminated or revoked.

5.7.11 Intranet/Extranet and Public Internet Access

Intranet, Extranet, and Public Internet access KQIs are all very similar and generally only vary by whether the provider is internal or external. The actual parameter requirements for each of the services are also likely to be different in each case.

The Internet KQIs only refer to the access to the systems via Internet protocols. If the applications run in the Internet environment, in particular voice, then the KQIs of those applications must also be considered in the SLA.

5.7.11.1 Internet Key Quality Indicators

Availability The system, both server and client, should have sufficient availability such that non-availability does not impact the enterprise objectives.

Response Time
The response time of the system should be such that it doesn't affect the effectiveness of the user. This will include response from the terminal (echoing key strokes, scrolling texts) and the server (displaying files, folders) and both (opening, closing, and saving files). This should be independent of the number of users of the system, up to the maximum number of users.

Goodput The system should be applied to provide sufficient goodput so as not to affect the efficiency of the user.

Confidentiality
The information developed and shared across the extranet may have a sensitivity that may be personnel, commercially, politically, or militarily sensitive. Confidentiality means must be applied as appropriate to the information to be protected.

Help Desk A help desk should be available for upgrades, problem resolution, etc.

Interoperability
Systems across the extranet should be interoperable.

5.7.12 Voice Services Key Performance Indicators

	Voice Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	MTBF, MTBR	Event Logs	
	Loss of Service	Help Desk Logs	Loss of service marked by help desk call if event log event not shown. As this is a common area of dispute, the method of determining the time of loss of service should be clearly defined and agreed by both parties in the SLA.
	LostCalls	Event Logs, SNMP, XML	Call unintentionally terminated
Speech/Visual Quality	MOS	Customer Survey	Reactive
	Loss, Jitter, Delay	R-Value/E-Model	Proactive, per call
Response Time	ResponseTime	Probe	Assume CODEC, client fixed
Round Trip Delay	OWD, RTT	Probe, Agent	
Delay	OWD, RTT	Probe Agent	Assume CODEC, processing time fixed
Confidentiality	PhysicalAccessViolations	Event Log	For key security and taps
Non-repudiation	PhysicalAccessViolations	Event Log, RADIUS	For facility and terminals
Interoperability	InteroperabilityComplaints	Help Desk	
	RejectedCalls	PBX/IP-PBX	
Connect Time	ConnectTime	Probe or PBX/IP-PBX	
Graceful Degradation	CallsRejected	PBX/IPBX	By priority of call

5.7.13 Video Teleconference Key Performance Indicators

	Video Teleconference Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	MTBF, MTBR	Event Logs	
	Loss of Service	Help Desk Logs	Loss of service marked by help desk call if event log event not shown.
Speech/Visual Quality	MOS	Customer Survey	Reactive
	Loss, Jitter, Delay	R-Value/E-Model	Proactive, per call
	CustomerSatisfaction	Customer Survey	For video in particular, depend on video content, bandwidth, and CODEC.
Response Time	ResponseTime	Probe, Agent	
Round Trip Delay	OWD, RTT	Probe, Agent	Assume CODEC, processing time fixed
Delay	OWD, RTT	Probe, Agent	Assume CODEC, processing time fixed
Confidentiality	PhysicalAccessViolations	Event Log	For key security and taps
Non-repudiation	PhysicalAccessViolations	Event Log, RADIUS	For facility and terminals
Help Desk	See Help Desk		
Interoperability	InteroperabilityComplaints	Help Desk	
Connect Time	ConnectTime	Probe or PBX/IP-PBX	

5.7.14 Streaming Video Key Performance Indicators

	Streaming Video Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	MTBF, MTBR	Event Logs	
Speech/Visual Quality	MOS	Customer Survey	Reactive
	Loss, Jitter, Delay	R-Value/E-Model	Proactive, per call
	CustomerSatisfaction	Customer Survey	For video in particular, depend on video content, bandwidth, and CODEC.
Response Time	ResponseTime	Probe, Agent	
Help Desk	See Help Desk		
Connect Time	ConnectTime	Probe, Server, or Client	

5.7.15 Broadcast Video Key Performance Indicators

	Broadcast Video Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	MTBF, MTBR	Event Logs	
Speech/Visual Quality	CustomerSatisfaction	Survey	Depends on circumstances.
	QualityComplaints	Help Desk	
Jitter	IPDV, CDV	TDR, Probe, NTU	

5.7.16 Telemetry Key Performance Indicators

	Telemetry Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	MTBF, MTBR	Event Logs	
Response Time	ResponseTime	Probe, Agent	
Round Trip Delay	OWD, RTT	Probe, Agent	
Jitter	IPDV, CDV	Probe, Agent	
Goodput (Offered)	Totaltraffic-ErroredTraffic-Retries	Server, Client, Agent, or Probe	Per Application and Session.
Confidentiality	AccessControlViolations	Domain Controller	
	PhysicalAccessViolations	EventLog	
Integrity	MD5 Violations	Server, Cllient	
Non-repudiation	ValidAccess	Domain Controller	Look at logs after revocation for validated access.

5.7.17 Transaction Key Performance Indicators

	Transaction Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	MTBF, MTBR	Event Logs	
Response Time	ResponseTime	Probe, Agent	
	ClientServiceTime, ServerServiceTime, WindowSize, AvPacketSize, IPLD, IPDV, IPLR	Response Model	Per application (aggregate) and per session (end user). Requires correlation between server, client, and network.
Round Trip Delay	OWD, RTT	Probe, Agent, NTU	
Jitter	IPDV, CDV	Probe, Agent, NTU	
Locking	ValidLocks, InvalidLocks, FailedLocks	From Transaction Server (Database)	
Transaction Rate	TotalTransactionsmade, Time	From Transaction Server	
Goodput (Offered)	TotalTransactionsProcessed - TotalTransactionsMade, Time	From Transaction Server	
Idle Time	ProccessorIdleTime	From Transaction Server	
Confidentiality	PhysicalAccessViolations, AccessViolations	Event Logs, Domain Controllers	For Facilities Access and terminal access.
Integrity	MD5 Violations	Server, Client	
Non-repudiation	AccessControlViolations	Firewall, Domain Controllers	
Graceful Degradation			
Revocation	ValidAccess	Domain Controllers	Look at logs after revocation for validated access.

5.7.18 Electronic Mail and Messaging Key Performance Indicators

	Electronic Mail and Messaging Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	MTBF, MTTR	Event Logs, help desk	
Delay	ReceivedTimeStamp-SentTimeStamp		Clock synchronization can be a possible issue, perhaps use received receipts.
Transaction Rate	Read/Write Requests		
Confidentiality	AccessControlViolations	Firewall, RADIUS, PGP, Directory Services	
Integrity	MD5 Violations, Signature Violations	Server, Cllient	
Non-repudiation	AccessControlViolations	Firewall, RADIUS, PGP, RSA, Directory Services Digital Signatures	
Disk Space	DiskSpaceUsed, DiskSpaceAvailable, FailledWriteRequests, FailledReadRequests	From storage device	
Graceful Degradation	PrioritisedMessagesRejected	From Server	
Revocation	ValidAccess ReturnReceipt(access error)	Domain Controller	Look at logs after revocation for validated access.

5.7.19 Office Automation Key Performance Indicators

Key Quality Indicators	Office Automation Key Performance Indicators Parameters	Method	Comments
Availability	MTBF, MTTR	Event Logs, help desk	
Response Time	ClientServiceTime, ServerServiceTime, WindowSize, AvPacketSize, IPLD, IPDV, IPLR	Response Model	Per application (aggregate) and per session (end user). Requires correlation between server, client, and network.
	ResponseTime	Network Probe	Measure response per request.
	ICMP Ping	Ping utility	Doesn't include application, may not show slow server or client response.
Confidentiality	AccessControlViolations	Domain Controller	
Integrity	Availability		
	PhysicalAccessViolations		For premises and key material.
Disk Space	DiskSpaceUsed, DiskSpaceAvailable, FailledWriteRequests, FailledReadRequests	From storage device	
Help Desk	See Help Desk		
Training	Satisfaction	Customer Survey	Taken before and after training.
Interoperability	InteroperabilityComplaints	Help Desk	

5.7.20 Help Desk Key Performance Indicators

	Help Desk Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	MTBF, MTTR	Event Log	
Speech/Visual Quality	Audibility, friendliness, knowledge, accent, politeness	Customer Survey or Phantom Calls	Often record sessions for this (and other) purpose.
Response Time	CallPickupTime	PBX/IP-PBX	Average, Mean, Median, Standard Deviation Required
Transaction Rate	TotalCalls, Time	PBX/IP-PBX	
Goodput (Offered)	TotalClosedCalls	Tracking Systems	
Throughput (Actual)	OfferedCalls	PBX/IP-PBX	
Idle Time	IdleTime	PBX/IP-PBX	Manned desks only?
Time to Close	JobCloseTime, JobOpenTime	Difference	Per call, mean, standard deviation
Call Hold Time	CallHoldTime	PBX/IP-PBX	
Connect Time	ConnectTim	PBX/IP-PBX	
Graceful Degradation	PrioritisedCallsDropped	PBX/IP-PBX	

5.7.21 Tele-Working Key Performance Indicators

	Tele-working Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	NotAvailable	Help Desk, Customer Survey	
Response Time	ConnectTine, AccessBandwidth, FirewallBandwidth	SNMP	
	ResponseTime	Network Probe	Measure response per request.
Goodput (Offered)	Totaltraffic-ErroredTraffic-Retries	Server, Client, Agent, or Probe	Per Application and Session.
Confidentiality	AccessControlViolations	Firewall, RADIUS	
Non-repudiation	AccessControlViolations	Firewall, RADIUS	
Help Desk	See Help Desk		
Training	Satisfaction	Customer Survey	Taken before and after training.
Revocation	ValidAccess	RADIUS	Look at logs after revocation for validated access.

5.7.22 Intranet Web Services Key Performance Indicators

	Intranet Web Services Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	MTBF, MTTR	Event Logs, Help Desk	
Response Time	ClientServiceTime, ServerServiceTime, WindowSize, AvPacketSize, IPLD, IPDV, IPLR	Response Model	Per application (aggregate) and per session (end user). Requires correlation between server, client, and network.
	ResponseTime	Network Probe	Measure response per request
Goodput (Offered)	Totaltraffic-ErroredTraffic-Retries	Server or Client Agents, Probes	
Confidentiality	AccessControlViolations	Firewall, RADIUS	
Help Desk	See Help Desk		
Interoperability	InteroperabilityComplaints	Help Desk	
Revocation	ValidAccess	RADIUS	Look at logs after revocation for validated access.

5.7.23 Extranet Web Services Key Performance Indicators

	Extranet Web Services Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	MTBF, MTTR	Event Logs, Help Desk	
Response Time	ClientServiceTime, ServerServiceTime, WindowSize, AvPacketSize, IPLD, IPDV, IPLR	Response Model	Per application (aggregate) and per session (end user). Requires correlation between server, client, and network.
Goodput (Offered)	Totaltraffic-ErroredTraffic-Retries	Server or Client Agents, Probes	
Confidentiality	AccessControlViolations	Firewall, RADIUS	
Non-repudiation	AccessControlViolations	Firewall, RADIUS	
Disk Space	DiskSpaceUsed, DiskSpaceAvailable, FailledWriteRequests, FailledReadRequests	From storage device	
Help Desk	See Help Desk		
Interoperability	InteroperabilityComplaints	Help Desk	
Revocation	ValidAccess	RADIUS	Look at logs after revocation for validated access.

5.7.24 Public Internet Key Performance Indicators

	Public Internet Key Performance Indicators		
Key Quality Indicators	**Parameters**	**Method**	**Comments**
Availability	NotAvailable	Help Desk, Customer Survey	
Response Time	ClientServiceTime, ServerServiceTime, WindowSize, AvPacketSize, IPLD, IPDV, IPLR	Response Model	Per application (aggregate) and per session (end user). Requires correlation between server, client, and network.
	ResponseTime	Network Probe	Measure response per request
Goodput (Offered)	Totaltraffic-ErroredTraffic-Retries	Server or Client Agents, Probes	
Help Desk	See Help Desk		

6 Use Scenarios

6.1 Introduction

This chapter focuses on user scenarios common to enterprise environments and discusses how the SLA principals outlined in this Handbook can be applied. Three cases are presented, in increasing levels of complexity:

- Remote Office Voice and Data Integration: Remote offices are integrated into corporate HQ using a common IP infrastructure for both voice and data to achieve efficiencies in management and bandwidth utilization.
- Executive on the Move: This scenario was developed as part of the Mobile Management Forum's ongoing activities in developing a Secure Mobile Architecture and Policy Management and outlines the processes to allow executives to maintain connectivity with their enterprise data in the office, at home, abroad, and in transit.
- Naval Application: In this scenario naval assets are interconnected with an IP infrastructure for the passage of voice and data applications with links that may be subject to loss because of loss of signal or because of jamming.

To reiterate the discussion of KQI and KQI in Section 5.2, the intent is not to specify actual values since these may be considered commercially sensitive, subject to negotiation between the provider and the customer, and also dependent on the application supported.

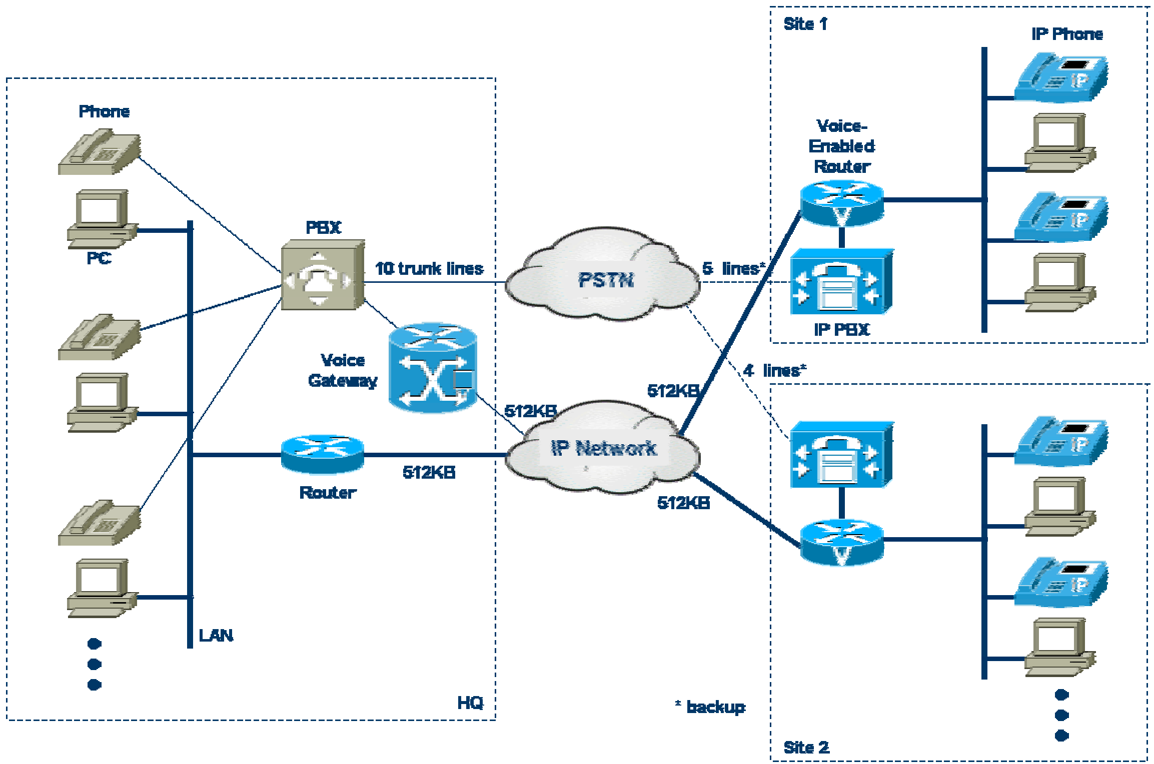
More scenarios and details of existing scenarios will be added in later revisions.

6.2 Remote Office Voice and Data Integration

The scenario depicted below describes the delivery of data and Voice over IP (VoIP) services within a multi-site corporation. This scenario is a composite of actual VoIP deployments and includes specific service measurements and best practices identified by VoIP vendors and industry experts as key to ensuring the highest order of service delivery.

As telephony service is an essential part of business operations, corporate VoIP service, when implemented, becomes the lifeblood of corporate communications. This is particularly true in today's business environment where "virtual teams" of individuals from different departments and physical locations often collaborate on corporate projects. Virtual teams cannot function effectively without telephone service.

Because voice communications occupies an important role in business communications and there is still some scepticism regarding VoIP services, the use of an SLA not only serves to allay the concerns of the users, but also show IT confidence in VoIP technology.



In this scenario, two new remote corporate offices are equipped with VoIP equipment, with connections to an existing traditional PBX-based telephone service at corporate headquarters. The HQ PBX handles all corporate traffic to and from the Public Switched Telephone Network (PSTN), and routes traffic from the remote sites destined for the PSTN. All intra-company voice traffic uses VoIP services over the existing IP network. Both remote sites have backup lines to the PSTN, which are only activated and used in case of IP network failures or congestion. Built-in Call Admission Control (CAC) and PSTN voice gateway technology within the VoIP PBXs work together to route calls to the PSTN upon detecting conditions where the IP network is unable to provide toll quality services over the IP network.

For corporations with geographically dispersed locations, telephone calls between facilities can amount to a significant monthly outlay. Even if a private network is deployed between the sites, the monthly cost of a private Frame Relay network can run into the thousands of dollars. With VoIP, a corporation can leverage the existing data network infrastructure and thus reduce the monthly inter-building networking costs.

Note that as the number of (and distance between) locations increases, the higher the potential cost savings using VoIP service, as both toll call and leased line costs increase with distance.

6.2.1 Underlying Service Characteristics

Successful VoIP service delivery relies on layers of services to provide a user experience that matches or even rivals those of traditional telephony service.

User Telephony Devices

Because the corporate voice network includes both VoIP and traditional PBX-based systems, different types of user telephony devices will be present. Spares of each type of telephony instrument used should be maintained to ensure quick turn-around types in case of device failure. Note that as the user telephony device is the user interface to the system, it has a significant effect on the user call experience.

Unified Messaging

The convergence of voice and data on the same infrastructure lends itself to a natural fit for unified messaging, where both voice and data messaging (i.e., voice mail and email) can be stored and delivered using the same server platform. As a business application, unified messaging promises gains in productivity simply by allowing users simultaneous access to voice and email messages using just one account. While not critical to VoIP operation (i.e., VoIP service should continue to run regardless of whether the unified messaging server is up or down), the absence of this service will affect how users perceive service availability and quality.

Alternate Path Routing

To provide a service with availability rates that match those of a PSTN-based voice network, alternative path routing using the PSTN as a backup path for voice traffic is required. The voice gateway provides the connection between the IP and TDM-based telephony worlds (PSTN and PBX) and must support a range of interfaces, as well as the call density, to handle large volumes of traffic. H.323 or Media Gateway Control Protocol (MGCP) support is the leading voice gateway protocol in use today. The use of alternative path routing is key to providing service availability that approaches a toll-quality experience.

Call Admission Control

If the IP network is down or there are insufficient resources to complete a call, Call Admission Control (CAC) technology instructs the voice gateway to pass the call to the PSTN. Dial Plans, which define reachability between all voice network entities and provide alternate path routing and policy restrictions, should be tightly coupled with the CAC to ensure service availability. Note that while the CAC and voice gateway functionality may exist in some IP PBX devices, they can also be separate devices. CAC is essential to ensuring that the VoIP system is able to prevent IP network overload and failure conditions from crippling the service.

IP Services Resources

Because VoIP systems are IP devices, they are dependent on the underlying network services in an IP network:

- Dynamic Host Configuration Protocol (DHCP) servers provide the IP addresses to the IP telephones in a VoIP network.
- Domain Name Service (DNS) servers provide name and IP address information for “directory” searches.
- Trivial File Transfer Protocol (TFTP) servers provide additional information about a connection (phone and user).

The availability of VoIP service is dependent on the availability of the underlying IP network services described above so, where possible, redundant IP services servers should be deployed.

IP Network Infrastructure

Because the VoIP service is implemented over the IP network infrastructure, it must be designed with the availability and performance requirements of VoIP in mind, including QoS and private networking capabilities.

6.2.2 Stakeholders

For this service, the service provider is the corporate IT department charged with planning, deploying, and operating the VoIP service. The customers of this service are the corporate employees.

6.2.3 Service Characteristics

The ubiquity of telephone service in corporate environments has resulted in a user base that is both experienced with and knowledgeable about the service. As a result, user expectations of telephone service are quite high as they relate to service availability and call quality.

As a corporate telephony service, VoIP is expected to have system availability and call quality characteristics that match or closely approximate those of a toll quality service.

6.2.4 Security

6.2.4.1 Confidentiality

This scenario envisages a normal enterprise environment. Information held by the enterprise is considered commercially sensitive and unauthorized exposure of information may provide advantage to competitors or adverse public scrutiny.

6.2.4.2 Integrity

As with many enterprises, the network and IT infrastructure is crucial to the success of the company. Availability is therefore an important threat to the enterprise. In a multi-national environment, such as in this scenario, there is a high reliance on external networks (PSTN, Internet), therefore there are both internal and external threats. The attack population (the number of people that could affect the network) is different in each case: internal employers, and partners with access to the internal network; external, general Internet population, external partners (and their employees), and the mobile workforce (including telecommuters). Particular threats would include:

- Component Failure (servers, terminals, network elements, applications, HVAC)
- Denial of Service Attacks
- Virus attacks
- Malicious or disgruntled employee
- Human error

6.2.4.3 Non-Repudiation

In this scenario, it is assumed that normal authentication procedures are in place – i.e., login with password – and that only external calls are sent over the PSTN. External repudiation is therefore handled by the PSTN in the normal way and the use of VoIP does not compromise the enterprise.

6.2.5 Service Measurement

Because it is telephony service, VoIP Service Level Agreements (SLAs) must necessarily track and measure the following metrics:

Availability

VoIP SLAs must track the frequency and duration of all aspects of VoIP service downtime to ensure that VoIP service continues to approach five nines availability. Simply put, downtime is defined when any one of the following three conditions happen on any given call attempt:

- No dial tone
- No ringing or busy signal
- Abnormal call termination

Each occurrence of the above conditions should be documented to pinpoint where the weakness in the entire system exists. To better understand the nature of the downtime, occurrences of downtime for all VoIP system components should also be tracked and documented:

- IP Telephony Equipment (IP PBX, H.323 Gatekeeper, IP Gateway, IP Phones, etc.)
- IP Network Services (DNS Server, DHCP Server, etc.)
- Network Infrastructure (PSTN Links, LAN Connections, Switches, IP Routers, etc.)

Finally, a Call Completion Percentage should be calculated for the system as a baseline for tracking performance and gauging system improvements.

Call Setup Performance

Because of fairly uniform experience with telephone systems, users will judge the quality of a service based on even mundane things such as the following:

- Dial Response Time, or the time to get a dial tone after picking up phone
- Call Setup Response Time, or the delay between dialing a number and getting a ring or busy signal

Call Quality

As discussed above, the standard measurement for user perception of voice call quality is the Mean Opinion Score (MOS). While MOS is useful and certainly valid, it is subjective and not easy to measure.

In an effort to remedy this, the International Telecommunications Union (ITU) has developed the E-Model standard (the ITU G.107 standard) as a means of objectively measuring call quality. The output of an E-Model is called an R-value that has been shown to reliably map to an estimated MOS.

The following constitute the components that contribute to the R-value output of an E-Model calculation: Delay, Jitter, Packet Loss, and CODEC selection together provide a reasonable means of estimating voice call MOS. To ensure toll-quality MOS ratings, the following service measurements should be met:

1. Delay: End-to-end packet delivery delay must not exceed 150 ms.
2. Jitter: Variations on packet delivery times must not exceed 50 ms.
3. Packet Loss: Percentage of packets dropped during transmission must not exceed 0.50%.
4. CODEC Selection: Only the following CODEC types should be considered:
    — ITU G.711u
    — ITU G.711a
    — ITU G.726-32
    — ITU G.729

The above CODEC types can all theoretically deliver a maximum MOS $\geq$ 4.0. Using devices based on other CODEC types may result in MOS < 4.0 (i.e., not toll quality).

6.2.6 Class of Service/Quality of Service Handling

In this scenario, VoIP traffic travels on the same infrastructure as data; it is therefore necessary to classify or mark VoIP traffic so that it is accorded a higher priority over data traffic and thus ensure the system's capacity to comply with the delay, jitter, and packet loss measurement criteria specified for the VoIP SLA. Traffic classification is performed as close to the edge of the network as possible, at the voice-enabled routers.

Using DiffServ Code Points (DSCP)

The VoIP QoS solution of choice is to use DiffServ Code Points to classify VoIP traffic, with distinction made between bearer traffic and call setup or control signaling traffic.

Alternative Solution: Using 802.1P and IP Precedence Bits

In deployments that require the use of legacy equipment that do not support DiffServ, an interim solution using IEEE 802.1P/Layer 2 Class of Service and IP Precedence/Type of Service (TOS) settings can serve as an effective alternative.

The following table specifies the COS/QoS values to be used for each type of traffic under the different classification schemes:

Packet Type	IP Precedence Value	DSCP (DiffServ)	IEEE 802.1P
Voice	5	EF	5
Control Plane Traffic	3	AF31	3
Data	0	0	0

Note that data continues to be treated and handled on a Best Effort basis under any scenario. Additionally, as the network is QoS-enabled, it is capable of classifying other types of traffic such that they are handled with different priorities based on their performance, delay, and jitter requirements for operation.

Once DSCP support is enabled on an IP network, it not only becomes capable of supporting VoIP services, but other QoS-based IP-based services, including streaming video broadcasts, IP videoconferencing, etc. These services have strict performance, delay, and jitter requirements that can only be reliably supported by a QoS-enabled IP infrastructure.

Because the underlying Layer 2 network supports IEEE 802.1P/Q, it is capable of delivering private networking via Virtual LAN (VLAN) technology. In fact, the recommended practice is to further separate voice and data traffic on a LAN by assigning them to different VLANs.

6.2.7 Monitoring and Reporting

Prior to placing the VoIP system in service, the system is tested and performance is measured to ensure that it meets the service measurement metrics discussed. The resulting measurements are then considered as the baseline. Each time a network addition or change is performed, system performance is measured and compared against the baseline figures.

Daily measurements are made (and reports generated) on network response time and jitter to monitor network health. Any measurement exceeding the target values necessitates further investigation, particularly if the measurements approach the maximum acceptable.

Note that the maximum acceptable response time and jitter monitoring thresholds are set at much lower levels than the performance metrics defined for acceptable VoIP service performance. The rationale for this is that as only data traffic measurement tools are being used, they may not detect problems in the voice service fast enough to initiate problem resolution procedures. By setting the thresholds low, personnel are alerted to pursue more aggressive monitoring of VoIP services way before a problem is experienced by the user.

Accounting and Billing

For planning and tracking purposes, IP PBX can collect aggregated call statistics as well as Call Detail Records (CDRs) for VoIP system calls. As this is an internal phone system, no formal billing is performed; VoIP operational costs are allocated among the different departments based on number of employees (users).

SLA Contract Compliance

Because voice communication occupies an important role in business communications and there is still some scepticism regarding VoIP services, the use of an SLA not only serves to allay the concerns of the users, but also show IT confidence in VoIP technology.

The SLA contract can be relatively simple, focusing on tangible user experiences such as service availability and call quality. Minutes of downtime or degraded performance are tracked starting from the time a problem call comes in until it is resolved. Semi-annual surveys of customer experiences with the service can also be useful in deriving an MOS figure that can be compared to an SLA-contracted MOS value.

For this scenario, the practical metrics to use in a VoIP SLA may be the following:

- Downtime minutes not to exceed X minutes for the preceding six-month period
- Dial and Call Setup response times not to exceed Y seconds for any call
- MOS not to fall below Z for the preceding six-month period

The above metrics are measurable, with specific known solutions for resolving problems should they occur. To minimize downtime, strict problem escalation and resolution procedures must be in place and adhered to by the IT staff.

Because it is an internal system, the penalties for SLA non-compliance are limited to personnel action, as in the case of negative performance reviews for staff charged with ensuring system performance to SLA-contracted levels.

6.3 Executive on the Move

The Executive on the Move scenario has been, and continues to be, a foundation document of the Mobile Management Forum (MMF). It was written in late 2000/early 2001 leading up to The Open Group Conference in Berlin 2001, on Mobility. The underlying scenario will continue to be used and built on in support of the MMF's ongoing activities in Secure Mobile Architecture, and Policy Management.

This scenario proposes a hypothetical SLA in support of the enterprise to support the 'Executive on the Move' and acknowledges the work and support of the MMF in the generation of this scenario. For a more detailed overview of the Executive on the Move scenario, refer to the MMF.

6.3.1 Overview

At the Berlin conference of The Open Group in April 2001, speakers from Motorola, Symbol Technologies, and Bertelsmann Media Systems described general use of mobile computing and communications by executives; while speakers from JP Morgan, Lufthansa, Brand Communications, St. Luke's Episcopal Hospital, Boeing, Critical Path, and Radicchio covered more specific topics. This scenario incorporates information from these presentations, especially those describing general use by executives.

The aim of the scenario is to increase efficiency of the enterprise and the executive by providing implicitly secure connectivity in a mobile environment (home, office, car, airport, plane) to:

- Improve productivity
- Improve decision-making by providing timely and critical information
- Improve corporate information

As a result of the conference, a Petroleum Refinery Contract Example was proposed. It is not real life, but puts together real-life situations in a fictional setting to illustrate some key points of computing and communications support for the Executive on the Move. SLAs would have been established to support this scenario.

6.3.2 Stakeholders

In this scenario it is assumed that the internal IT departments on the enterprise are the providers of the services to its customers (users). To accomplish this scenario, it is likely that secondary SLAs will have to be established with external service providers (e.g., mobile operators) that will be managed by the IT department for the benefit of its customers.

6.3.3 Business Services

The objective of the scenario is to allow the executive to perform informed decision-making while at home, in transit, and in the office. The following business services are envisaged:

Voice Communications
: Voice continues to be an important service for all enterprises. In this scenario it is assumed that voice connectivity is achieved through both wired and wireless connectivity throughout the world 24x7 hrs. Number mobility would be preferable as would reduced call charges, especially for roaming access.

Video Teleconferencing
: Face-to-face meetings with multi-media presentations over VTC facilities are important to cut travel costs and time, but involve all members of the decision process, irrespective of location and timezone.

Office Automation
: Business planning tools – spreadsheets, RDBMS, email, etc. – are important decision aids.

Information Delivery
: Alert services via Instant Messaging (IM) and Short Message Service (SMS) provide timely information on market and economic conditions, stock-market information, and news services for proactive and reactive decision-making.

Intranet Services
: Company-wide distribution of information or services with common access throughout the network of files, applications, and information.

Telecommuting Services
: To allow inclusion of personnel out of normal work hours and for home-working, telecommuting services are offered, with the same levels of information and security as if at home.

Internet Access
: Internet access provides access to the enterprise network, with appropriate restrictions for security for workers in public facilities, like hotels, airports, and Internet cafes and allows access to public domain market intelligence with access to search engines, etc.

Help Desk
: First-line support is offered by a help desk. Communication with the help desk would be through any appropriate medium: voice, data, or text services. Remote diagnosis and configuration of terminals may be appropriate.

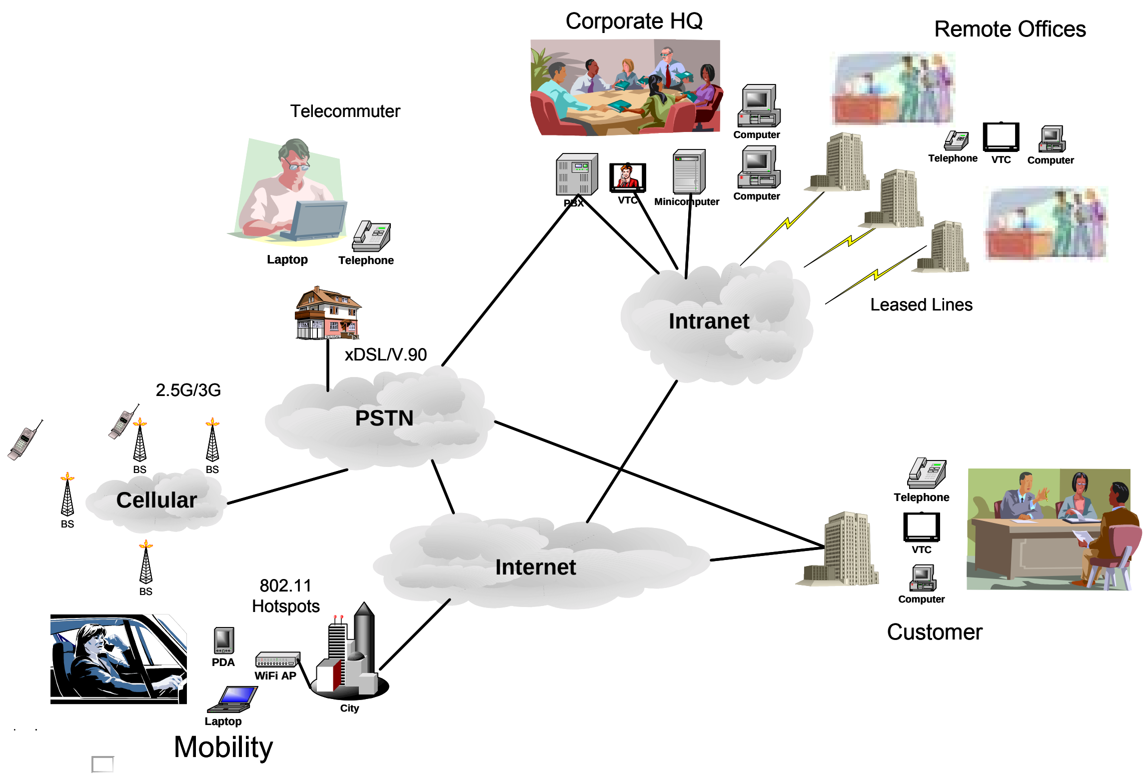


Figure 27: Schematic Representation of the Executive on the Move Use Case

6.3.4 Security

6.3.4.1 Confidentiality

This scenario envisages a normal enterprise environment. Information held by the enterprise is considered commercially sensitive and unauthorized exposure of information may provide advantage to competitors or adverse public scrutiny.

6.3.4.2 Integrity

As with many enterprises, the network and IT infrastructure is crucial to the success of the company. Availability is therefore an important threat to the enterprise. In a multi-national environment such as in this scenario, there is a high reliance on external networks (PSTN, Internet); therefore there are both internal and external threats. The attack population (the number of people that could affect the network) is different in each case: internal, employers and partners with access to the internal network; external, general Internet population, external partners (and their employees), and the mobile workforce (including telecommuters). Particular threats would include:

- Component Failure (servers, terminals, network elements, applications, HVAC)
- Denial of Service Attacks
- Virus attacks
- Malicious or disgruntled employee

- Human error

6.3.4.3 *Non-Repudiation*

With a mobile and distributed workforce with access over both private and public network access, access control and authentication is an important consideration to prevent masquerade and to ensure authenticated decision processes.

In the present scenario it is assumed that as the executive roams to different technology areas (office, car, airport, home), that the call is re-established; although there is the notion of session (layer 4) persistence (a great achievement of the MMF and the scenario).

6.3.5 Service Resources

In support of the business services, a number of service resources are required which include:

- Terminals: PC (Home and Office), Laptop, PDA, WAP/GPRS Cellular Phone (Laptop and PDA have 802.11 Wi-fi access.)
- Directory Services for the storage of profiles, handover, etc.
- IP VPN to allow secure transmission remotely across the Internet
- Servers for disk space, printing, storage, backup, audit
- Remote Access Servers to allow remote connectivity for home access
- Dynamic DNS/DHCP to allow information to be sent to the most relevant terminal
- IP-enabled PBX (VoIP would allow the executive to be cost-effectively included in conference calls while at home or remotely.)
- Security Resources: Firewalls, PKI, Virus protection software, IP VPN, Certificate Authorities

These service resources should allow the executive to maintain connectivity with corporate information and colleagues.

6.4 Naval Communications

Today, highly important time-sensitive warfighting applications are emerging on the network grid which require special handling at a ship's LAN/WAN interface to ensure proper timely delivery of information. Best-effort technology will no longer suffice. The ship must act as an Internet service provider to provide SLAs to users' applications.

There are two major challenges:

1. Varying transmission paths

   The changing available transmission paths present some significant challenges for QoS. The radio frequency (RF) media sometimes experience outages due to antenna blockage, pitch and roll, and look angle to the satellite (changing direction of the ship). Outages can often be detected and/or predicted, and possibly signaled to QoS appliances and other devices.

2. Multiple domains

The Navy's traffic traverses other network domains (e.g., DISN and NMCI) that may be considered analogous to external carrier networks with their own QoS/SLAs.

Data Types

The Navy's data types include commercial applications: VoIP, VTC, POTS, DNS, router updates, database replication (e.g., Lotus), WWW, email, ftp. They also include tactical applications data that have near-real time latency requirements and reliability requirements.

Data Classification and Marking

In the near-term, the Navy's traffic will be marked to indicate latency and priority. An example of a DSCP marking scheme is contained in Table 11 and Table 12 All TCP traffic is assumed to handle packet loss at Layer 4, so sensitivity to packet loss is not marked. This may be subject to change.

Traffic may also be distinguished by IP (Layer 3) header information, and possibly by higher layer headers as well.

6.4.1 Security

In the military environment, security is a key concern that often introduces great complexity into the design.

Integrity
: The network interconnects weapon platforms with sensors and command and control. It is therefore vital that the network be available for the duration of a mission and must gracefully degrade to allow the greatest fighting capabilities to be maintained for the greatest time. Mission-critical information capabilities should be preferred in preference to other systems; e.g., welfare. Voice communications represent an important command and control medium and as such often support multi-level priority so that high-priority orders are prioritized over less important communications. Today, voice communications take precedence over data communications, although this may change with time.

Confidentiality
: Confidentiality of information and communication is an important aspect of all military communications. The threat to the system can be from compromise of orders and also threats from unintentional disclosure of position and possible attack from emission. Military personnel will be assigned a clearance level appropriate to their task and other factors such as nationality (especially for coalition forces). Data should be further restricted to the appropriate security level and also to those of the appropriate need to know.

Non-repudiation
: Authentication is an important aspect to enforce security policies to ensure that only appropriate personnel and assets are interconnected.

As a result, most military networks can be considered as a number of separate security domains achieved cryptographically. Interconnection is achieved by boundary protection devices or within a trusted computer device.

For the purposes of this use case, therefore, it is possible to consider one of the security domains in isolation but assume in practice that there are a number of domains. It is likely that the SLA would be the same for each device.

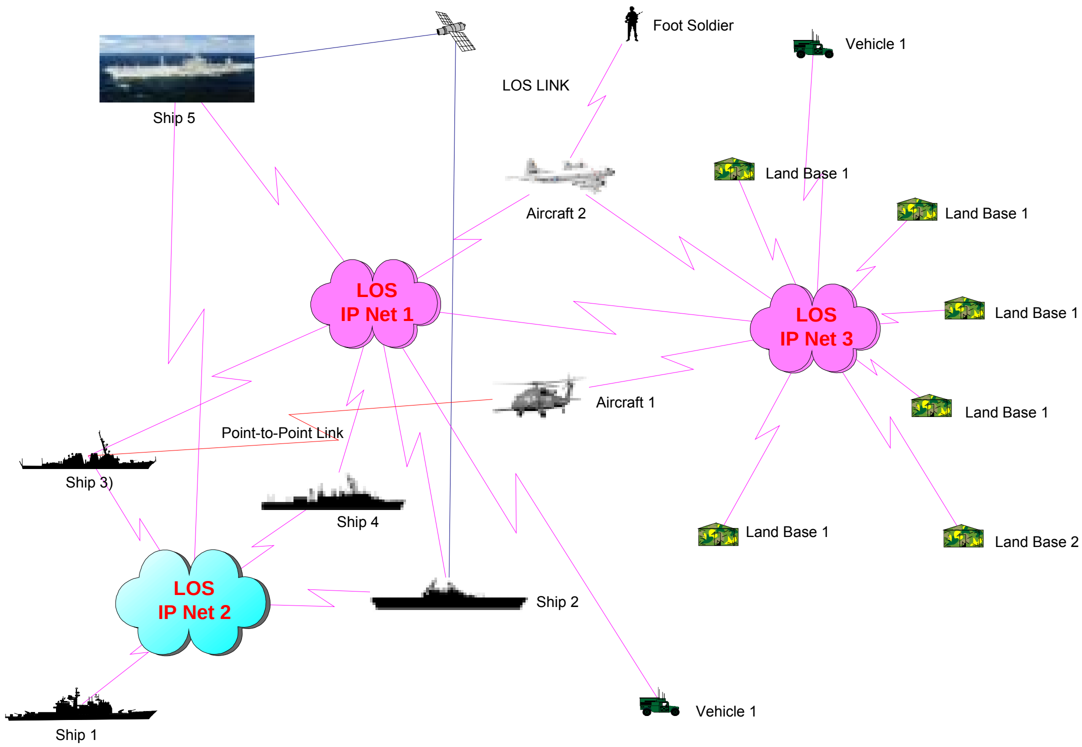


Figure 28: Navy and Joint Line of Sight (LOS)

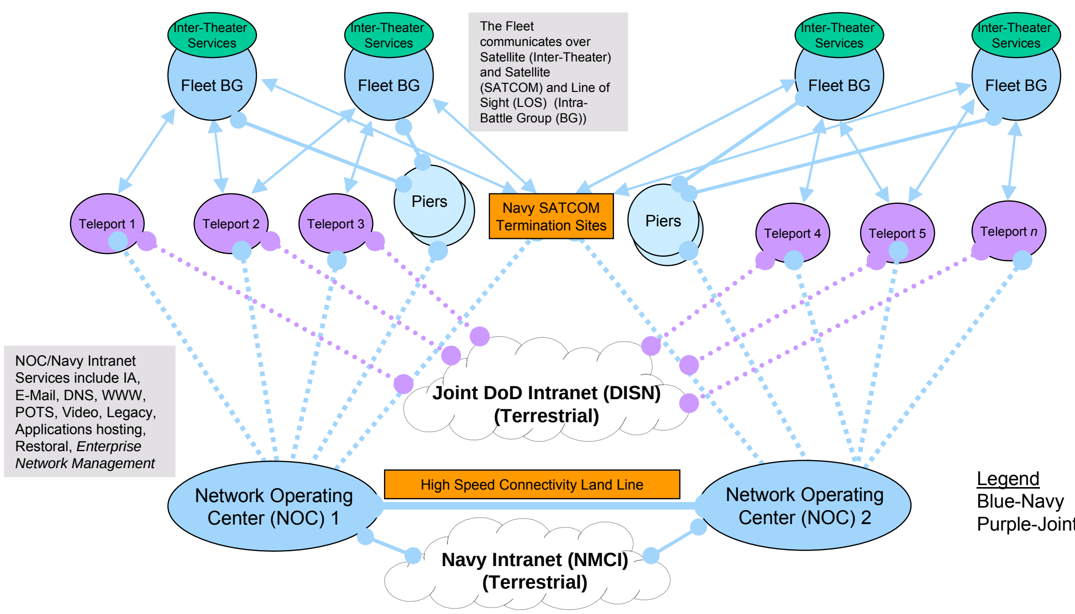


Figure 29: Navy and Joint Pt-Pt SATCOM and Terrestrial Network Clouds

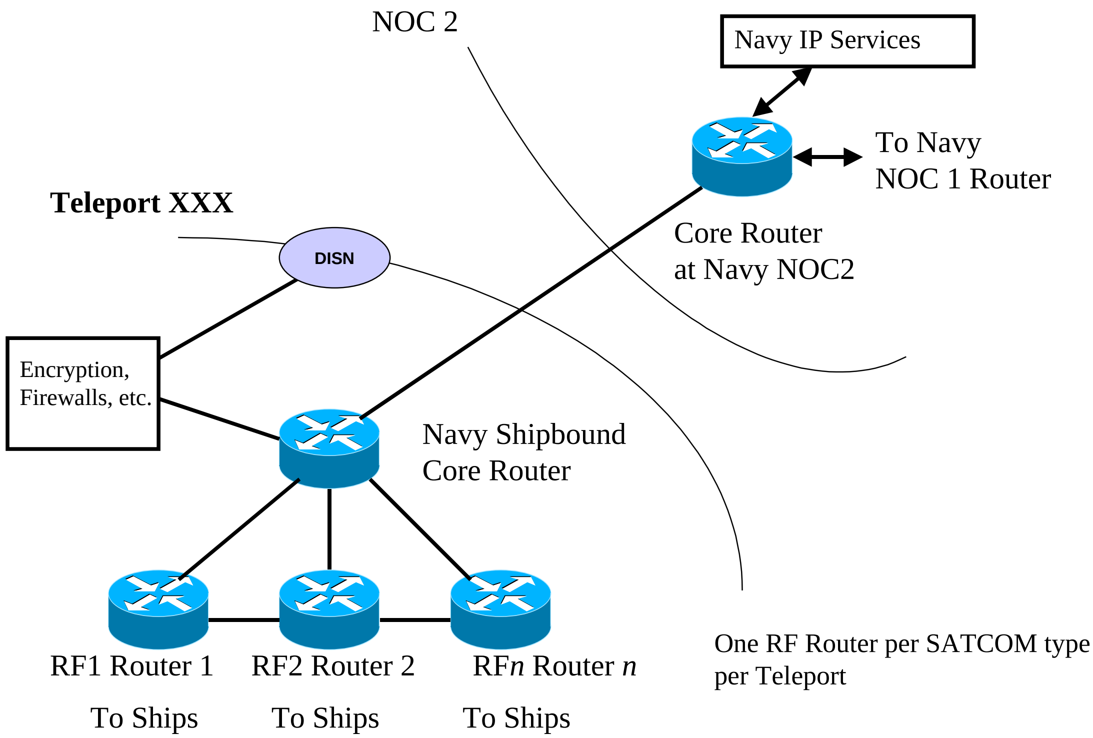


Figure 30: Navy Network Infrastructure at Teleports and NOCs

6.4.2 Navy SLA Requirements

Applications requirements for end-to-end QoS will include:

(Note that non-IP data (e.g., POTS) is not included in this document, to simplify/scope the problem.)

1. Minimum bandwidth guarantees for specified data types or sources
2. Latency and reliability guarantees for other data types
3. The usual SLA requirements that apply for VoIP and VTC
4. The Navy also requires the ability to do multi-level precedence pre-emption (MLPP) for voice calls where IP is the transport. (There are various degrees of importance regarding military phone calls: flash override, flash, immediate, priority, and routine.)
5. The Navy SLA for VoIP will only guarantee a fixed number of calls, over which the calls will be refused.
6. Best Effort (BE) for the rest
7. The Navy wants to maximize goodput over the satellite links (i.e., minimize retransmissions and backoff for TCP due to RF transmission error).

The Navy wants to utilize all available bandwidth over its RF links. The RF links vary in round trip delays, outage characteristics, and BER.

Navy Proposed SLA	DiffServ RFC Name	DSCP	TOS Bit Values
Delay-tolerant (hours)	Best Effort	0	000 000 00
		1	000 001 00
		2	000 010 00
		3	000 011 00
		4	000 100 00
		5	000 101 00
		6	000 110 00
		7	000 111 00
	Precedence 1	8	001 000 00
		9	001 001 00
Interactive (sec.) Low Pri	AF11	10	001 010 00
		11	001 011 00
Delay-sensitive (min) High Pri	AF12	12	001 100 00
		13	001 101 00
Delay-sensitive (min) Low Pri	AF13	14	001 110 00
		15	001 111 00
	Precedence 2	16	010 000 00
		17	010 001 00
Tactical (msec) High Pri	AF21	18	010 010 00
		19	010 011 00
Tactical (msec) Low Pri	AF22	20	010 100 00
		21	010 101 00
Interactive (sec.) High Pri	AF23	22	010 110 00
		23	010 111 00
	Precedence 3	24	011 000 00
Future Tactical Apps		25	011 001 00
	AF31	26	011 010 00
		27	011 011 00
	AF32	28	011 100 00
		29	011 101 00
	AF33	30	011 110 00
		31	011 111 00
	Precedence 4	32	100 000 00
		33	100 001 00
Video	AF41	34	100 010 00

Navy Proposed SLA	DiffServ RFC Name	DSCP	TOS Bit Values
		35	100 011 00
	AF42	36	100 100 00
		37	100 101 00
	AF43	38	100 110 00
		39	100 111 00
	Precedence 5	40	101 000 00
VoIP Control Reserved (B/W Guarantee)		41	101 001 00
		42	101 010 00
		43	101 011 00
		44	101 100 00
		45	101 101 00
VoIP	EF	46	101 110 00
		47	101 111 00

Table 11: Example DSCP Marking Scheme

Mapping of Applications	DSCP #	AF Class
All SMTP, best-effort (hours)	0	BE
(min.) FTP	14	AF13
(min.) DB repl.	12	AF12
(sec.) HTTP/HTTPS, DNS	10	AF11
(sec.) Telnet/Chat	22	AF23
(msec) Tactical App Low Pri	20	AF22
(msec) Tactical App High Pri	18	AF21
Future Tactical Apps	24-31	AF31-AF33
Video, all UDP	34	AF41
Reserved Apps (poss. future)	34	AF41
VoIP Control	41	EF
VoIP	41	EF

Table 12: Mapping of Applications

Annex A: KQI/KPI Parameterization

Parameter	Mandatory/ Optional	Type	Weighting Factor	Remarks
Unique Identifier	M	Alpha-numeric	N/A	
Measurement Parameter 1	M	KQI or KPI	$0 \leq x \leq 1$	The weighting factor for all measurements parameters should equal 1.
Measurement Parameter 1 Time Zone	O (see note 3)	String		The Co-ordinated Universal Time (UTC) string that enables the identification of the timezone the underlying KPI(s) is/are originated from.
Measurement Parameter 2	O	String	$0 \leq x \leq 1$	
Measurement Parameter 2 Time Zone	O (see note 3)	String		The Co-ordinated Universal Time (UTC) string that enables the identification of the timezone the underlying KPI(s) is/are originated from.
Measurement Parameter n	O	Ditto	$0 \leq x \leq 1$	
Measurement Parameter n Time Zone	O (see note 3)	String		The Co-ordinated Universal Time (UTC) string that enables the identification of the timezone the underlying KPI(s) is/are originated from.
Transformation Algorithm	M	N/A	N/A	The algorithm to be applied to calculate the KQI value from the combination of measurement parameters and their weighting factors. The simplest form of algorithm is "=" where a single KPI becomes a KQI.
Objective Value	O	N/A	N/A	The value that is used to determine whether the calculated KQI is within the target value. The value is used to determine whether a violation should be generated.

Parameter	Mandatory/ Optional	Type	Weighting Factor	Remarks
Violation Direction	O	+ve or –ve	N/A	The determination of whether a breach occurs if the calculated value is above or below the objective value; i.e., if this field is –ve a violation or warning will be generated if the calculated value drops below the objective value or warning threshold.
Warning Level	O	%	N/A	Optionally a % value may be specified to give a warning that a KQI objective is in danger of violating.
Abatement Threshold	O	%	N/A	Introduces a hysteresis factor to the warnings and alarms to prevent oscillation between an alarm and a clear condition.
Measurement Period	M	Seconds	N/A	The period over which the KQI performance is calculated (see note 1).
Grace Periods	O	#	N/A	How many times a value can remain un-updated without violation of the SLA in reference to the Measurement Period and/or KQI reporting period.
Active/Inactive Period	O	A UTC string I UTC string	N/A	A KQI measurement period may be controlled by specifying a start and stop time for the KQI; e.g., for a KQI that is only applicable between 9 a.m. and 5 p.m. this field would be set to A 09:00,17:00. If the field is preceded with the letter A, the time period has to be included in the measurement; if preceded with the letter I, the time period has to be excluded from the measurement.
KQI Timezone	O	UTC String		The Co-ordinated Universal Time (UTC) string which permits the identification of the timezone that the KQI is originated from. Can be derived from homogeneous KPI or timestamp if valid.
Active/Inactive Days	O	A Day list I Day list	N/A	List of days that the KQI measurement is active or inactive; e.g., for a KQI that is measured from Monday through Friday the list would be A 2,3,4,5,6. To exclude Saturday and Sunday's the list would be I 6,7.

Table 13: Key Performance Indicator Definitions

Note 1: As a KQI may be defined by a number of measurement parameters that have differing sample periods it must use a sliding window technique for calculating the current value. The current value will therefore be updated in real time as new data becomes available and aggregated over the time specified by the measurement period.

Note 2: The Weighting Factor is optional. When applied it must be included for all parameters. If omitted, the effective weighting factor for all parameters will be considered as equal.

Note 3: If the timezone for parameter one is not stated, then local time is assumed. If the timezone for the subsequent parameters is not stated, then the same time zone for parameter 1 is applied.

Annex B:
Six Parameter Category and KQI/KPI Model

B.1 Six Parameter Category Model (GB 917)

Seminal work on the six parameter category appears in earlier TM Forum work on Performance Reporting (TMF 701). This work was later extended and completed in GB 917 to distinguish between single-user (service) instance and aggregated (service) requirements as shown in Table B.1:

Service Perspective	**Parameter Category**		
	Technology-Specific	**Service/Application-Specific**	**Technology/Service-Independent**
Single Service Instance (User) (SAP-related)	E.g., physical interface details	E.g., service type or bundle	E.g., maximum down-time of one event
Aggregated Service Requirements	E.g., monthly reported parameters	E.g., billing method (usage or time-based)	E.g., aggregate availability over all users

Table 14: Six Paramter Framework for Parameter Categories

It is important to consider the parameters for each user instance of the service and overall service instances offered to the customer. Poor performance for a single instance must be assessed together with the average performance of the whole service.

Service and technology-independent parameters are those that are often specified in an SLA that don't pertain to the performance of the service. Examples would include availability, MTBO, MTBF, time to first yield, and average call response times. These are sometimes referred to as "operational performance criteria". Other examples would include billing period, alternate routing/redundancy, and backup and restore performance.

Technology-specific parameters refer to the technology that supports the service or application. Examples would include a web hosting application where relevant technology-specific parameters would include the performance of the web server software or the operating system. For a network service such as ATM, the supporting technology may be SONET/SDH or a PDH transport service. In some instances, the supporting technology may be a service in its own right, supported by an SLA.

Service-specific parameters are those typically related to the business service carried by the service. An example business service would be voice. Service-specific parameters would include: call quality (measured by MOS), call completion rate, and propagation delay.

The six parameter category framework maps to the service resources discussed in Section 1.2 and shown in Figure B.1:

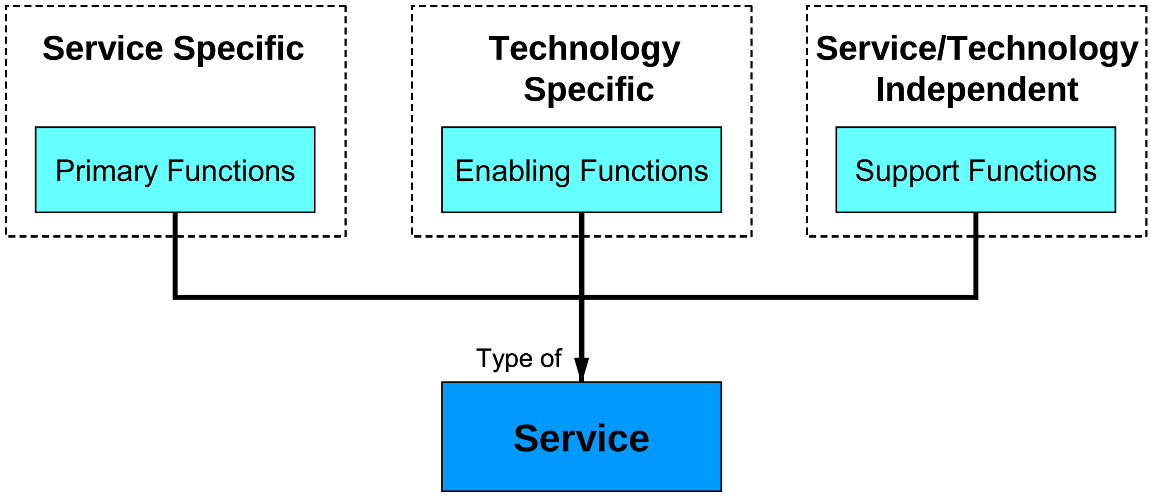


Figure 31: Relationship between the Six Parameter Model and Service Functions

B.2 Examples

Consider the example of Voice over IP (VoIP) supported by an IP network that is supported over an IEEE 802.3 (Ethernet) network. The service is provided and supported by the internal MIS department.

Mapping each function or service to the six parameter model is shown in the following table:

	Parameter Category		
Service Perspective	**Technology-Specific (IP, Ethernet)**	**Service/Application-Specific (Voice)**	**Technology/Service-Independent (Support)**
Single Service Instance (SAP-related)	IP Performance Metrics Ethernet Performance Metrics	Voice Quality Call Setup Delay	Help Desk Performance Availability MTBF, MTTR, etc.
Aggregated Service Requirements			

Table 15: VoIP Service in the Six Parameter Category Model

If the service is also the direct technology that enables the service, then the service and technology-specific parameters become identical. An example would be an intranet service for an enterprise supported by the MIS department thus:

<table>
<tr><th rowspan="2">Service Perspective</th><th colspan="3">Parameter Category</th></tr>
<tr><th>Technology-Specific (IP)</th><th>Service-Specific (IP)</th><th>Technology/Service-Independent (Support)</th></tr>
<tr><td>Single Service Instance (SAP-related)</td><td colspan="2" rowspan="2">IP Performance Metrics</td><td rowspan="2">Help Desk Performance
Availability
MTBF
MTTR, etc.</td></tr>
<tr><td>Aggregated Service Requirements</td></tr>
</table>

Table 16: IP Service in the Six Parameter Model

In this example it is likely that the IP service relies on, for example, underlying supporting services, Ethernet, or leased lines. These may be subject to separate SLAs or pertinent parameters may be included in the IP performance metrics. Whether subject to an SLA or not, such supporting services must be configured, monitored, and possibly reconfigured to maintain the SLA for the supported service.

B.3 Key Quality and Performance Indicators

There is a difficulty in mapping service-specific parameters to technology-specific parameters that are more easily measured and reported. As a consequence, traditional SLAs have focused almost solely on performance of the supporting service.

However, an accelerating trend towards service-focused management has led to the requirement of a new “breed” of indicators that focus on service quality rather than network performance. These new indicators, or Key Quality Indicators (KQIs), provide a measurement of a specific aspect of the performance of the application or service. The KQI are derived from a number of sources, including performance metrics of the service or underlying support services as KPI. As a service or application is supported by a number of service elements, a number of different KPI may need to be determined to calculate a particular KQI. The mapping between the KPI and KQI may be simple or complex, empirical or formal. The KQI/KPI concept adopted by this volume closely follows the discussion outlined in the Wireless Services Measurement Handbook (GB 923).

Using the example introduced earlier of VoIP, the first consideration must be the application that is supported. Assume the application is a call center. Using both the six parameter category model and the KQI/KPI model, the call center application SLA framework becomes:

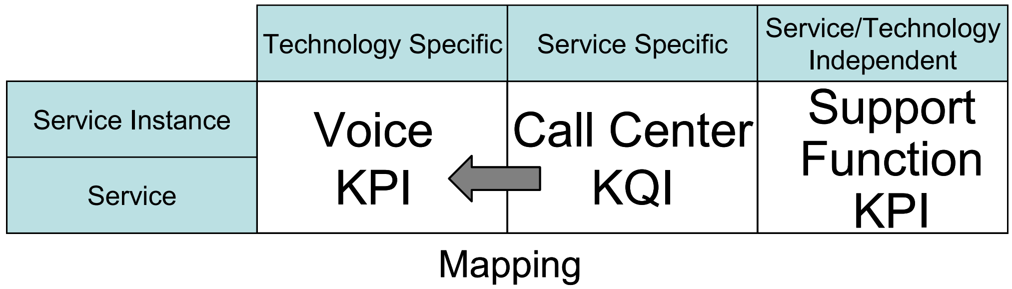


Figure 32: Mapping KQI/KPI to the Six Parameter Model for a Call Center Application

The mapping shown between the call center KQIs and the VoIP KPIs may be the subject of negotiation and may involve functional mapping of the voice KPIs to provide a measure of a KQI. There is likely to be a number of KQIs, each requiring technology-specific KPI to be monitored in the SLA. Such negotiation must result in an agreement that the KPI measurements represent a measure of the KQI required for the application.

The support functions – help desk, for example – may be a separate business service or may be included in the application or service KQIs.

In this example, one of the business services is voice, so the call senter KQIs must be mapped to (among others) KPI metrics on the voice service that provide measures of KQI. Such KPIs are discussed in Section 1.3 and would include MOS, blocking probability, and delay. These KPIs may be measured directly from the service (from the PABX or call agent, for example) or may be inferred by KPI metrics from supporting (IP) services. The KPIs for the voice service therefore map to the voice KQIs of the IP network service in a hierarchy as shown in Figure 33:

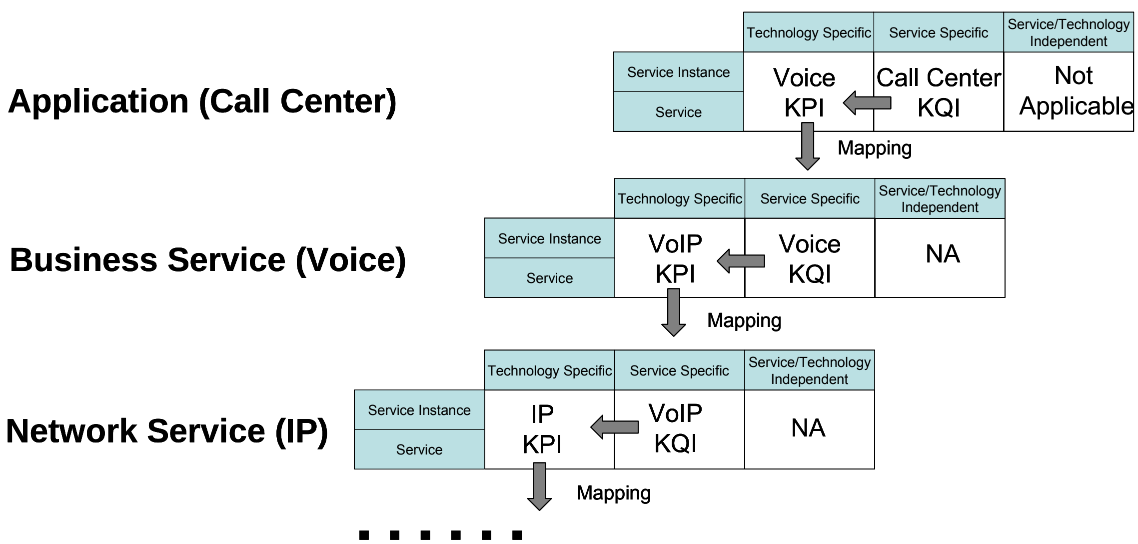


Figure 33: Hireachy of Service SLA KQI/KPI Parameters

This hierarchy could continue to, for example, the underlying support services, facilities, or power (or even regulatory authorities), or may terminate at a relevant level.

This hierarchy provides functional decomposition of a service or application in both technological and administration domains. Each level of the hierarchy may be subject to a separate SLA or administrative or cost domain. If the entire hierarchy represents one administrative domain, then it is reasonable to collapse the hierarchy into a single level, as shown in Figure 32.

B.4 Summary

The six parameter category model, as specified in GB 917, and the KQI/KPI model defined in GB 923 are both valid models for SLA parameterization. The KQI/KPI model, however, introduces the distinction between performance and quality service metrics and the concept of a mapping between the KQIs and KPIs that may be simple or complex. A simple mapping of the KQI/KPI model to the six parameter model is possible and a hierarchy determined that provides clarity in a complex environment, where many applications may be supported on a particular service or across different domains. This clarity should also provide completeness and a focus and understanding in SLA parameter negotiations.

Annex C: Qualify of Experience

Quality of Experience (QoE) is a term to allow for subjective as well as objective measures of QoS, performance, and all aspects of the interaction (experience) with the service or product. Covering all aspects of the customer or end-user experience, these include the harder to define aspects of a product or service:

- Politeness, knowledge, accent of call center in handling orders, problems, etc.
- Ergonomics of the product, ease-of-use, localization, quality of user documentation
- Assumed knowledge of customer or end user
- Responsiveness of system to failure
- Stock market performance of company
- Brand image
- Experience with partners, distributors, etc.

Many of these aspects fall outside the scope of this document, but QoE is an important consideration for the enterprise as non-conformance in one area can be balanced against better performance in another; e.g., the quality of problem resolution at a help desk or call center may counter, if only partially, the impact of the initial problem, brand image, etc. The different aspects of the customer experience should not be treated in isolation. Customer focus groups, user surveys, and registration questionnaires are all methods that can be employed to gauge opinion on subjective manners.

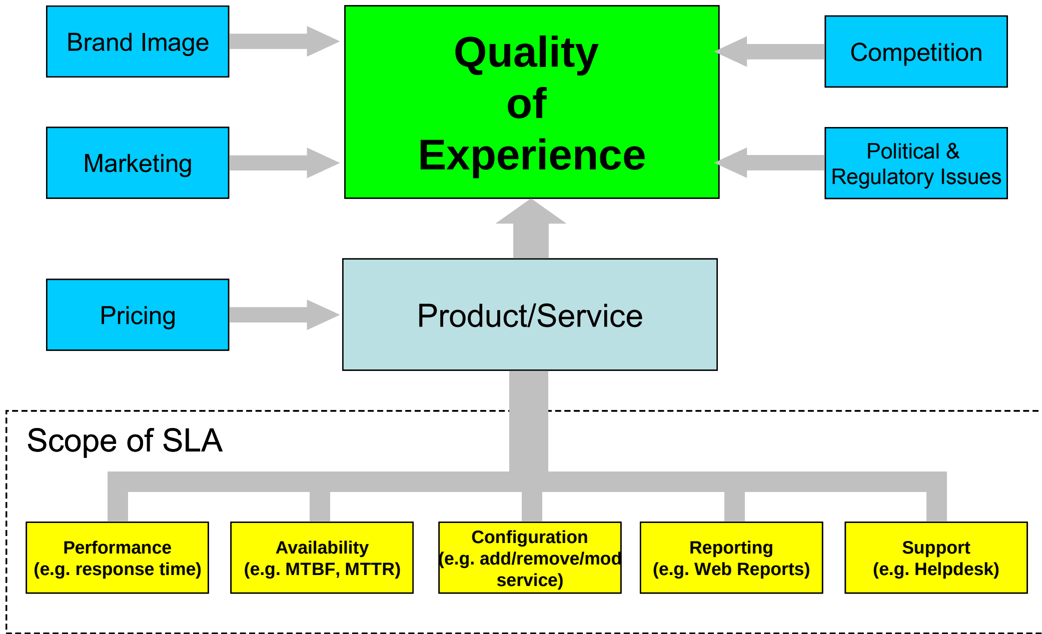


Figure 34: Role of the SLA in the Customer Quality of Experience (QoE)

The relationship between QoE and SLA is that QoE relates to the perception of the quality of the product or service, whereas SLA refers to the definition, measurement, and reporting of objective measures of the product or service. As such, QoE and SLA are related in that if the perception of the product or service is poor, yet the service parameters fall within the limits defined by the SLA, the SLA must be redressed. The key concept is to map the perceptive measures from the QoE into objective measures for the SLA. This mapping may be multi-dimensional, empirical, functional, or complex in nature.

Annex D: Terms and Definitions

Terms and definitions used in this Handbook are mainly based on TM Forum's Performance Reporting Concepts and Definitions Document (TMF 701), Wireless Service Measurement Handbook (GB923), and internationally agreed definitions published by the International Telecommunications Union (ITU). For example, ITU-T Recommendation M.60 (M.60) contains Maintenance Terminology and Definitions. In this chapter, only some key terms related to SLAs, QoS, and performance are defined and their source identified. For the purposes of this Handbook, some definitions have been modified and/or extended. Where multiple definitions are in use within the industry, only those appropriate to these volumes is given. Not all of the terms and definitions in this chapter are necessarily used in this version of the Handbook but are included as they are expected to be relevant in later versions[5]. For further information, consult TMF 701, ITU-T, and other references.

Active (condition) – Condition (e.g., alarm) not cleared (ITU-T Rec. M.2140).

Administration (task) – A broad group of functions that sustain services once they have been established.

Aggregate Interval – The time period over which raw data is summarized in a particular report. For example, raw 15-minute data may be aggregated into one-hour or 24-hour intervals for reporting purposes (TMF 701 modified).

Alarm – An alerting indication of a condition that may have immediate or potential negative impact on the state of service resources; e.g., network element, application, system, etc. (ITU-T Rec. M.3010 modified).

Anomaly – A discrepancy between the actual and desired characteristic of an item. The desired characteristic may be expressed in the form of a specification. An anomaly may or may not affect the ability of an item to perform a required function (ITU-T Rec. M.20).

Availability Performance – The ability of an item to be in the state to perform a required function at a given instant of time, or at any instant of time within a given time interval, assuming that the external resources, if required, are provided. Note that this ability depends on the combined aspects of the reliability performance, the maintainability performance, and the maintenance support performance of an item. In the definition of the item, the external resources required must be delineated. The term "availability" is used as an availability performance measure (ITU-T Rec. M.60).

Business Application (BA) – A process, function, product, or service or a combination that in isolation or together fulfils a business requirement (e.g., gain revenue). Examples would include call centers, distribution, and order processing. Most enterprises would have a combination of business applications to fulfil their business objectives.

Bearer Service – A type of telecommunication service that provides the capability for the transmission of signals between user-network interfaces (ITU-T Rec. I.112).

[5] Terms and definitions not yet used are marked with a *.

Business Service (BS) – A facility that provides a business function. Examples are voice, video conferencing, email, and Internet access. Business services don't fulfil a business requirement in their own right unless it is an outsourced service offering, but enable a business requirement to be accomplished. If business services require access to a network (e.g., call center), then the business services would be layered on network services.

Clear – The end of a fault; the termination of a standing condition (ITU-T Rec. M.2140).

Connection – An association of transmission channels, circuits or flows, switching, and other functional units set up to provide a means for a transfer of information between two or more points in a telecommunication network (ITU-T Rec. Q.9 modified).

Customer – The user of the service, services, or functions offered. The customer may be the ultimate user of the service, services, or functions or merely the procurer or administrator of the service, services, or functions who takes responsibility for the requirements and experience of the end user.

Customer Contact Point – A physical or conceptual point at which a service provider can interact with any customer of the offered service for the purpose of maintaining communication services (TMF 701 modified).

Customer Details – Information about the customer, sometimes called the Customer Profile.

Customer Premises Equipment (CPE) – Any network equipment sited at the customer's premises used to provide the contracted service. Note that the CPE may be owned and operated by the customer or the service provider (extracted from TMF 701).

Data Collection Interval – The period over which performance parameters are accumulated to compute each stored measurement and detect maintenance threshold crossings (ITU-T Rec. M.2140). The time interval when statistics are retrieved from the network. This interval does not have to be the same as the measurement interval because the network devices may buffer statistics so that a number of them may be collected later (TMF 701).

Defect – A defect is a limited interruption of the ability of an item to perform a required function. It may or may not lead to a maintenance action depending on the results of additional analysis (ITU-T Rec. M.20).

Degradation – The presence of anomalies or defects in the absence of a fault (ITU-T Rec. M.2140).

Degraded Service – The presence of anomalies or defects that cause a degradation in QoS, but do not result in total failure of the service.

End User – The ultimate user of the service, services, or functions offered by the provider.

Enterprise – A general term to embrace any commercial, utility, government, academia, and healthcare entity. As a service provider is either a government or commercial organization offering telecommunications services, service providers can also be considered as an enterprise in their commercial aspects and relationships with other service providers.

Event – An instantaneous occurrence that changes the global status of an object. This status change may be persistent or temporary, allowing for surveillance, monitoring, and performance measurement functionality, etc. Events may or may not generate reports; they may be spontaneous or planned; they may trigger other events or may be triggered by one or more other events (ITU-T Rec. X.700).

Failure – The termination of the ability of an item to perform a required function. Note that after a failure, the item has a fault (ITU-T Rec. M.20).

Fault – The inability of an item to perform a required function, excluding that inability due to preventive maintenance, lack of external resources, or planned actions. Note that a fault is often the result of a failure of the item itself, but may exist without prior failure (ITU-T Rec. M.20).

Grade of Service (GoS) – The designed service quality, also known as guaranteed QoS, used as a comparison with delivered (measured) QoS. A service provider commits a particular GoS to their customers and then the QoS is a measurement of what was actually delivered (TMF 701 modified). An alternative definition is offered below based on ITU-T Study Group 2 work. GoS is the minimum level of service quality designed into the network supporting the service and maintained by traffic planning and engineering management actions depending on traffic densities over the duration the service is offered or used. As such, GoS represents a guaranteed expected level of service quality for any connection in the same QoS class of a specified service at any instant, and may in fact be improved upon depending on traffic loading of the network.

Indication – An intermediate output of the event correlation process. A notification indicating a persistent network, application, or system-detected trouble condition. The three types of indication are fault indication, impairment indication, and trend indication (ITU-T Rec. M.2140 modified).

KPI – Key Performance Indicators provide a measurement of a specific aspect of the performance of a service resource (network or non-network) or group of service resources of the same type. A KPI is restricted to a specific resource type (TMF GB923).

KQI – Key Quality Indicators provide a measurement of a specific aspect of the performance of the product, product components (services), or service elements and draw their data from a number of sources including the KPIs (TMF GB923).

Mean Time Between Failures (MTBF) – The average time between failures of the service.

Mean Time Between Outages (MTBO) – The average time between outages of the service.

Mean Time to Provide Service (MTPS) – The average time to actually provide a specified service from the date of signing a contract to provide service. This may or may not be specified in the SLA.

Mean Time To Repair (MTTR) – The average time to repair service resources.

Mean Time to Restore Service (MTRS) – The average time to restore service after reporting a fault; this time including the time to sectionalize and locate the fault. This may or may not be specified in the SLA.

Measurement Interval – The time interval over which each performance parameter is measured. For example, a parameter may be the number of errored seconds that occurred over a 15-minute measurement interval (TMF 701 modified).

Network Operator (NO)/Provider – A company or organization which operates and provides telecommunication network paths and connections as a business. These may be direct to end customers (in which case the NO is also a service provider) or under contract to service providers who in turn provide services to end customers.

Network Performance (NP) – The ability of a network or network portion to provide the functions related to communications between users. Note that NP applies to the network provider's planning, development, operations, and maintenance and is the detailed technical part of QoS, excluding service support performance and human factors. NP is the main influence on serviceability performance. NP measures are meaningful to Network Providers and are quantifiable at the part of the network to which they apply. QoS measures are only quantifiable at a Service Access Point (SAP) (ITU-T Rec. E.800).

Network Service Provider (NSP) – An internal or external entity that provides network services and infrastructure. NSPs may be tiered to provide end-to-end network connectivity, and may require connectivity to several NSPs, with appropriate SLA contracts. The term service provider (SP) is reserved for the generic term of an external NSP, providing telecommunications services for enterprises and is subject to service provider regulation.

Operations – These include the operation of work centers, technical support centers, support systems, test equipment, methods, and procedures, as well as the personnel and training required to install and maintain all the elements that constitute the network capability underlying the relevant services (ITU-T Rec. M.3010).

Performance Management (PM) – A set of TMN management functions which enable the performance (i.e., the ability to reproduce the signal) of the network services to be measured and corrective actions to be taken (ITU-T Rec. M.3010).

Performance Report – A statement of performance achieved over a given measurement interval and made available to the customer by a variety of methods. These include printed or electronic reports provided on a regular basis, stored reports in a database accessible by the customer, or on-demand reports. Two basic types of report are normally available – QoS performance against SLA guaranteed values and traffic data/utilization reports (extracted from TMF 701).

Provider – The entity that provides a service, services, or functions to customers. These customers may be internal or external to the provider.

Quality of Experience (QoE) – A concept comprising all elements of a customer's or end user's perception of using a service or product. QoE is a subjective term that can be modified by changes in objective measures in a complex manner.

Quality of Service (QoS) – The collective effect of service performances that determine the degree of satisfaction of a user of the service. Note that the QoS is characterized by the combined aspects of service support performance, service operability performance, service integrity, and other factors specific to each service (ITU-T Rec. E.800).

Quality of Service Reports – Reports generated from the service quality and performance data to report the performance of the service as a whole.

Raw Performance Data – Raw performance data collected from various data sources including the network and service resources, such as network elements, network and element management systems, and network and application servers. In addition, data collected for the service provider's Operations Support Systems (OSS), such as trouble ticketing, order processes, maintenance and support, customer care, etc.

Reporting Period – The period at which performance reports are generated. It is defined independently for each SAP Group within the associated SLA (TMF 701).

Service – A service is a set of independent functions that are an integral part of one or more business processes. This functional set consists of the hardware and software components as well as the underlying communications medium. The customer sees all of these components as an amalgamated unit (TMF 701 modified).

Service Access Point (SAP) – A logical or physical element located on the interface between a customer's domain and the Provider's domain, representing the point at which a service is delivered. An SAP can be weighted in accordance with a business-critical factor that is defined in the SLA (TMF 701 modified).

Service Access Point Group – A group of SAPs against which service availability must be reported. Each SAP normally belongs to one (or more than one) SAP Group and each SAP Group contains at least one SAP. The association between SAP Groups and SAPs is defined within the associated SLA, according to the required computation and aggregation format (TMF 701).

Service Availability (SA) – A measure of the fraction of time during a defined period when the service provided is deemed to be better than a defined QoS threshold. SA is measured in the context of an SLA between the customer and the service provider concerned. It is expressed as a percentage (SA%) to indicate the time during which the contracted service (e.g., SVCs, PVCs, end-to-end circuits including protocols, applications, etc.) at the respective SAPs is operational. Operational means that the customer has the ability to use the service as specified in the SLA (TMF 701 modified).

Note that the calculation of the SA may include weighting of the SAPs as noted above. The detailed formula is contained in TMF 701 and ITU-T Rec. M.1539.

Service Degradation Factor (SDF) – A factor agreed between the customer and the provider used to weight the service availability (SA) calculation when the service is still available, but degraded from its contracted QoS (extracted from TMF 701).

Service Descriptions – Details of the service product catalog offered by the service provider.

Service Element – A service element comprises one or more network or service resources that are combined with other service elements to provide a service (TMF 701 modified).

Service Instance – A service that has been instantiated for a customer.

Service Level Agreement (SLA) – A formal negotiated agreement between two parties, sometimes called a Service Level Guarantee. It is a contract (or part of one) that exists between the service provider and the customer, designed to create a common understanding about services, priorities, responsibilities, etc. (TMF 701 modified). An SLA or Contract is a set of appropriate procedures and targets formally or informally agreed between network operators/service providers (NO/SP) or between network operators/service providers and customers, in order to achieve and maintain specified Quality of Service (QoS) in accordance with ITU (ITU-T and ITU-R) Recommendations. The SLA may be an integral part of the Contract. These procedures and targets are related to specific circuit/service availability (SA), error performance, Ready for Service Date (RFSD), Mean Time Between Failures (MTBF), Mean Time to Restore Service (MTRS), and Mean Time To Repair (MTTR) (ITU-T Rec. M.1340).

Service Level Agreement Contracts – Individual-specific SLA contracts between the service provider and the customer with the specific service details, agreed levels of service quality, reporting schedules, and details of actions to be taken when the service quality guarantees are not met.

Service Level Agreement Reports – Reports generated from the customer SLA quality and performance data to report the performance of the specific service instance for a specific customer against an SLA.

Service Level Agreement Templates – Definitions of the standard grades of service that can be offered with an SLA; for example, templates to describe gold and silver service characteristics.

Service Provider (SP) – A company or organization that provides telecommunication services as a business. SPs may operate networks, or they may simply integrate the services of other providers in order to deliver a total service to their customers. Providing a telecommunication service to any one end customer may involve multiple SPs, where one provider may "sub-contract" with other providers to fulfil the customer's needs (TMF 701). Note that the term service provider is now being used generically and may include telecom service providers (TSPs), Internet service providers (ISPs), application service providers (ASPs), and other organizations that provide services; e.g., internal IT organizations that need or have SLA capabilities or requirements.

Session – An end-to-end connection between two entities for the exchange of information.

Annex E: Acronyms

3GPP	Third Generation Partnership Project
ABR	Available Bit Rate
ABT	ATM Block Transfer
ADSL	Asymmetric Digital Subscriber Line
AIP	Application Infrastructure Provider
ANSI	American National Standards Institute
ANT	Access Network Transport
API	Application Programming Interface
ASP	Application Service Provider
ASPIC	Application Service Provider Industry Consortium
ASR	Answer/Seize Ratio
ATC	ATM Transfer Capability
ATIS	Alliance for Telecommunications Industry Solutions
ATM	Asynchronous Transfer Mode
BBE	Background Block Error
BBER	Background Block Error Ratio
BER	Bit Error Ratio
BICC	Bearer-Independent Call Control
B-ISDN	Broadband Integrated Services Digital Network
CATV	Cable Television (Community Antenna Television)
CBR	Constant Bit Rate
CCR	Call Completion Record
CD	Cell Delay
CDR	Call Data Record
CDV	Cell Delay Variation
CDVT	Cell Delay Variation Tolerance
CE	Cell Error
CER	Cell Error Ratio
CIF	Common Intermediate Format
CIM	Common Information Model
CIR	Committed Information Rate

CL	Cell Loss
CLI	Calling Line Identifier
CLR	Cell Loss Ratio
CM	Cell Misinsertion
CMR	Cell Misinsertion Ratio
CN	Core Network
CNM	Customer Network Management
CODEC	Compression Decompression
COPS	Common Open Policy Service
CPE	Customer Premises Equipment
CPoF	Common Point of Failure
CTD	Cell Transfer Delay
CUG	Closed User Group
DBR	Deterministic Bit Rate
DiffServ	Differentiated Services
DLCI	Data Link Connection Identifier
DMTF	Distributed Management Task Force
DNS	Domain Naming Service
DPL	Degraded Performance Limit
DS	Differentiated Services
DSCP	DiffServ Code Point
DSL	Digital Subscriber Line
DSS	Digital Subscriber SIgnalling System
DVI	Digital Video Interface
EBA	Enterprise Business Application
EBS	Enterprise Business Service
ECBP	End-to-end Connection Blocking Probability
EDC	Error Detection Code
EDGE	Enhanced Data rates for GSM Evolution
EFS	Error Free Seconds
EQoS	EURESCOM Quality of Service
ES	Errored Second
ESR	Errored Second Ratio
ETSI	European Telecommunications Standards Institute
EURESCOM	European Institute for Research and Strategic Studies in Telecommunications
FDD	Frequency Division Duplex
FEC	Forward Error Correction

FG	Finished Goods
FMCG	Fast Moving Consumer Goods
FR	Frame Relay
FRAD	Frame Relay Access Device
FSA	Framework Study Area
FTD	Frame Transfer Delay
G-CDR	Gateway GPRS Support Node – Call Detail Record
GERAN	GSM/EDGE Radio Access Network
GFR	Guaranteed Frame Rate
GFP	Generic Framing Protocol
GGSN	GPRS Gateway Support Node
GII	Global Information Infrastructure
GoS	Grade of Service
GPRS	General Packet Radio Service
GSM	Global System for Mobile communication
HDLC	High Bit Rate Digital Subscriber Line
HCPN	Hybrid Circuit-switched/Packet-based Network
HTTP	Hyper Text Transfer Protocol
IAB	Internet Architecture Board
IDS	Intrusion Detection System
IESG	Internet Engineering Steering Group
IETF	Internet Engineering Task Force
IMT	International Mobile Telecommunications
IN	Intelligent Network
INMD	In-service Non-intrusive Measuring Device
IntServ	Integrated Services
IOPS.ORG	Internet Operators Group
IP	Internet Protocol
IPDV	IP Packet Delay Variation
IPER	IP Packet Error Ratio
IPLR	IP Packet Loss Ratio
IPPM	IP Performance Metrics
IPTD	IP Packet Transfer Delay
IRTF	Internet Research Task Force
ISDN	Integrated Services Digital Network
ISM	In-Service Monitoring
ISO	International Organization for Standardization

ISP	Internet Service Provider
IST	Information Society Technologies
ISV	Independent Software Vendor
IT	Information Technology
ITAA	Information Technology Association of America
ITU-R	International Telecommunication Union – Radiocommunication Sector
ITU-T	International Telecommunication Union – Telecommunication Standardization Sector
LAN	Local Area Network
LDP	Label Distribution Protocol
LSR	Label Switching Router
MBS	Maximum Burst Size
MCR	Mean Cell Rate
MIB	Management Information Base
MOS	Mean Opinion Score
MPEG	Moving Picture Coding Experts Group
MPLS	Multi-Protocol Label Switching
MSP	Managed Service Provider
MTBF	Mean Time Between Failures
MTBO	Mean Time Between Outages
MTIE	Maximum Time Interval Error
MTPS	Mean Time to Provide Service
MTRS	Mean Time to Restore Service
MTTP	Mean Time to Provision
MTTR	Mean Time To Repair
NER	Network Effectiveness Ratio
NGN	Next Generation Network
NII	National Information Infrastructure
N-ISDN	Narrow-band Integrated Services Digital Network
NM	Network Management
NMDG	Network Measurement Development Group
NMF	Network Management Forum
NNI	Network-Node Interface
NO	Network Operator
NP	Network Performance
NP&D	Network Planning and Development
NPC	Network Parameter Control
NS	Network Service

NSP	Network Service Provider
NTE	Network Terminating Equipment
OAM	Operations, Administration, and Maintenance
OI	Outage Intensity
ONP	Open Network Provision
OS	Operations System
OSI	Open Systems Interconnection
OSS	Operations Support System
OTN	Optical Transport Network
OWD	One-Way Delay
OWPL	One-Way Packet Loss
P2P	Peer-to-Peer
PC	Personal Computer
PCR	Peak Cell Rate
PDH	Plesiochronous Digital Hierarchy
PDN	Public Data Network
PDP	Packet Data Protocol
PDU	Protocol Data Unit
PEI	Peak Emission Interval
PHB	Per Hop Behavior
PIB	Policy Information Base
PIR	Peak Information Rate
PLM	Product Line Management
PNO	Public Network Operator
POH	Path OverHead
PON	Passive Optical Network
POS	Packet Over Sonet
PRM	Performance Report Message
PPP	Point-to-Point Protocol
PS	Packet Switched
PSNR	Peak Signal to Noise Ratio
PSTN	Public Switched Telephone Network
PTT	Postal, Telegraph, and Telephone
PVC	Permanent Virtual Circuit
QCIF	Quarter Common Intermediate Format
QoS	Quality of Service
QOSF	Quality of Service Forum

QSDG	Quality of Service Development Group
RADIUS	Remote Access, Dial In User Service
RAN	Radio Access Network
RFC	Request for Comments
RFI	Request for Information
RFP	Request for Proposal
RFSD	Ready For Service Date
RPR	Resilient Packet Ring
RSVP	Resource Reservation Protocol
RTCP	Real-Time Control Protocol
RTSP	Real-Time Streaming Protocol
RTP	Real-Time Protocol
RTT	Round Trip Time
SA	Service Availability
SAP	Service Access Point
SBR	Statistical Bit Rate
S-CDR	Serving GPRS Support Node – Call Detail Record
SCR	Sustainable Cell Rate
SDH	Synchronous Digital Hierarchy
SE	Service Element
SECB	Severely Errored Cell Block
SECBR	Severely Errored Cell Block Ratio
SEP	Severely Errored Period
SEPI	Severely Errored Period Intensity
SES	Severely Errored Second
SESR	Severely Errored Second Ratio
SGSN	Serving GPRS Support Node
SIP	Session Initiation Protocol
SLA	Service Level Agreement
SLO	Service Level Objective
SLS	Service Level Specification
SLSU	Service Level Specification and Usage
SMI	Structure of Management Information
SNA	Synchronous Network Architecture
SOHO	Small Office Home Office
SONET	Synchronous Optical Network
SP	Service Provider

SP&D	Service Planning and Development
SQM	Service Quality Management
SVC	Switched Virtual Circuit
TCA	Traffic Conditioning Agreement
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TDM	Time Division Multiplexing
TEWG	Traffic Engineering Working Group
TFT	Traffic Flow Template
TIE	Time Interval Error
TIPHON	Telecommunications and Internet Protocol Harmonization over Networks
TMF	TeleManagement Forum
TMN	Telecommunications Management Network
TOM	Telecom Operations Map
TSAG	Telecommunication Standardization Advisory Group
TSP	Telecom Service Provider
TT	Trouble Ticket
TV	Television
UAS	Unavailable Seconds
UBR	Unspecified Bit Rate
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
UNI	User-Network Interface
UPC	User Parameter Control
UPL	Unacceptable Performance Limit
UPS	Uninterruptible Power Supply
URL	Uniform Resource Locator
UTC	Universal Time, Coordinated
UTRA	UMTS Terrestrial Radio Access
VAR	Value Added Reseller
VC	Virtual Circuit
VCI	Virtual Circuit Identifier
VDSL	Very High Bit Rate Digital Subscriber Line
VOATM	Voice Over ATM
VOD	Video On Demand
VODSL	Voice Over DSL
VoIP	Voice over IP

VPI	Virtual Path Identifier
VPN	Virtual Private Network
WAN	Wide Area Network
WAP	Wireless Access Protocol
WDM	Wave Division Multiplexing
WTSA	World Telecommunications Standardization Assembly
xDSL	Any DSL service (ADSL, HDLC, SDLC, VDLC)
XIWT	Cross-Industry Working Team
XML	Extensible Markup Language