

# CONSULTATION

## Quality of Service measurement and improvement

# Quality of Service measurement and improvement

A Consultation issued by the Telecommunications  
Regulatory Authority  
12th May 2003

**Purpose:** To provide a framework for the establishment of service performance measurement and improvement by operators delivering telecommunications services in the Kingdom of Bahrain.



هيئة  
تنظيم  
الاتصالات  
Telecommunications  
Regulatory  
Authority

# CONSULTATION

## Quality of Service measurement and improvement

### Table of contents

<b>1</b>	<b>Introduction.....</b>	<b>1</b>
1.1	Consultation process.....	1
1.2	Scope .....	2
<b>2</b>	<b>Meeting the Authority's duties under the Law .....</b>	<b>3</b>
2.1	Approach to protection of users and subscribers.....	3
2.2	Stages on the route to managed performance .....	3
2.3	Provision of information to users .....	4
<b>3</b>	<b>The measurement of service performance.....</b>	<b>6</b>
3.1	Convention 1 Responsibilities of the Authority .....	6
3.2	Convention 2 Responsibilities of the operators.....	6
3.3	Convention 3 Qualification of operators and services.....	6
3.4	Convention 4 Service types .....	7
3.5	Convention 5 Market types .....	8
3.6	Convention 6 Performance types.....	8
3.7	Convention 7 Common approach to measures .....	9
3.8	Convention 8 Use of industry forae.....	12
3.9	Convention 9 Provision of information.....	12
3.10	Convention 10 Stages of development.....	13
<b>A 1</b>	<b>Background paper on service model approach to measurement.....</b>	<b>14</b>
A1.1	Service types .....	14
<b>A 2</b>	<b>Background paper on service level mathematics.....</b>	<b>21</b>
A2.1	Introduction.....	21
A2.2	References.....	22
A2.3	Engineering design process.....	22
A2.4	Parameters with distributed values.....	24
A2.5	Test level setting and risk of failure .....	29
A2.6	Impact of sampling on guarantees.....	33
A2.7	Steady state and real time measurements .....	34
A2.8	Handling multiple sourced systems .....	37
A2.9	Some practical techniques.....	39
A2.10	Summary.....	46
<b>4</b>	<b>Schedules to the Consultation.....</b>	<b>47</b>
<b>Schedule 1</b>	<b>Measurement methods .....</b>	<b>48</b>

# CONSULTATION

<b>Quality of Service measurement and improvement</b>	
Schedule 1.1	Implementation .....48
Schedule 1.2	Operation.....48
Schedule 1.3	Performance .....48
Schedule 1.4	Other .....49
<b>Schedule 2</b>	<b>Measurement definitions .....50</b>
Schedule 2.1	Subsidiary parameters .....50
Schedule 2.2	Initial response time .....51
Schedule 2.3	Delivery performance .....52
Schedule 2.4	Delivery Time .....52
Schedule 2.5	Complaints .....52
Schedule 2.6	Billing .....52
Schedule 2.7	Customer satisfaction .....53
Schedule 2.8	Mean Time To Repair (MTTR) .....53
Schedule 2.9	Availability .....53
Schedule 2.10	Security (breaches) .....54
Schedule 2.11	Mean Time Between Failures (MTBF) .....54
Schedule 2.12	Probability of Blocking.....54
Schedule 2.13	Probability of Loss of Circuit .....54
Schedule 2.14	Mean and 95%ile packet delay .....54
Schedule 2.15	Probability of Packet Loss .....55
Schedule 2.16	Jitter .....55
<b>Schedule 3</b>	<b>Enforcement schemes.....55</b>
<b>Schedule 4</b>	<b>Performance metrics and targets .....56</b>
Schedule 4.1	Implementation targets.....56
Schedule 4.2	Operation targets .....57
Schedule 4.3	Performance targets.....58
Schedule 4.4	Other targets .....59
<b>Schedule 5</b>	<b>Glossary and definitions .....59</b>

# CONSULTATION

## Quality of Service measurement and improvement

### 1 Introduction

Legislative Decree No. 48 of 2002 promulgated the Telecommunications Law (the “Law”) for the Kingdom of Bahrain. The Law formed the Telecommunications Regulatory Authority (the “TRA” or “Authority”) which has responsibility for the regulation of the telecommunications of Bahrain during its transition to a competitive industry. One of the roles assigned to the TRA in section 3, (b), 1 of the Law is that it shall:

*“...carry out its duties ...in the manner best calculated to:*

*1. protect the interests of subscribers and users in respect of....*

- .....
- *availability and provision of service*
- *quality of services; and.....”*

In pursuance of these duties, the Authority has prepared this Consultation document which aims to prepare mechanisms for ensuring that the availability, provision and quality of telecommunications services provided in the Kingdom shall be consistent with the protection of the interests of the users and subscribers of those services.

The Authority recognises that there is a wide range of services and service types that may be provided within the Law. The characteristics of these services differ depending on the technology used and the type of service being delivered, and without care, the administration of Quality of Service (QoS) standards could become onerous. Therefore the Authority wishes to establish a common framework for the measurement, improvement and, where applicable, guarantees of levels of service that can apply to all types of service, both current and future, without the need for the underlying framework to be modified. The performance levels specified will be the same for all operators of equivalent services.

This document, therefore, contains a discussion on the background to the needs for performance measurement, the methods that could be adopted and the mathematics involved. It provides the rationale for the Authority’s thinking in addressing the issue of Quality of Service in the provision of telecommunications services in Bahrain. It also includes the core of the measurement system that the Authority would like to adopt following the consultation process. Additional background material is provided in the Attachments to ensure a common basis of definition for the consultation.

The document contains a number of Conventions that define overall principles to be applied, and then a series of Schedules that provide specific information relating to one or more detailed aspect of the provision of good quality services. These Conventions and Schedules will form the core of the measurement system once the consultation is complete.

#### 1.1 Consultation process

Interested parties are invited to respond in writing to the TRA with comments and suggestions on this document prior to the close of business on Wednesday 21<sup>st</sup> June 2003. Following consideration of the responses the TRA will then issue the formal quality of service regulation.

# CONSULTATION

## Quality of Service measurement and improvement

The address for responses to this Consultation is:

The Market Operations Unit  
Telecommunications Regulatory Authority  
PO Box 10353  
Manama  
Kingdom of Bahrain

Alternatively, formal e-mail responses may be sent to the Authority's e-mail address at [consult@tra.org.bh](mailto:consult@tra.org.bh)

### 1.2 Scope

While it remains in consultation form, this document has no legal status other than as a basis for discussion. Once the consultation is complete, the Conventions and Schedules, as modified following the consultation, will be issued in the form of a regulation or other equivalent formal instrument. In such form it will be the principal structure for the measurement and reporting standards to be followed by all of the operators in the Kingdom that fall under its ambit. Certain exemptions are proposed herein that minimise the overhead of reporting for statistically insignificant services, and also for operators in the process of starting up operations in Bahrain.

The Authority recognises that some of the measurements of the type proposed herein will take time to settle in and become part of normal operating practice so the process will be introduced in stages depending on the maturity of the operators and the services being delivered. This process is described in section 3.10 below.

Once issued, the QoS regulation will define the performance aspirations for the provision of telecommunications services in the Kingdom of Bahrain, and the TRA will generally make use of competition to achieve these goals. Where an operator is declared as dominant in any specific area, the services associated with that area will not be subject to effective competition. For these services, the aspirations defined herein will become mandatory requirements on the dominant operator, acting as a proxy for competition until competition becomes effective.

# CONSULTATION

## Quality of Service measurement and improvement

### 2 Meeting the Authority's duties under the Law

#### 2.1 Approach to protection of users and subscribers

The Authority regards its primary goals in meeting its duty under the Law as follows:

- To ensure that, consistent with any special conditions applicable to Bahrain, the performance of telecommunication services in the most general sense should meet or exceed equivalent levels of performance available in developed countries elsewhere in the world
- To provide users with comparative information to enable them to make informed decisions as to which of the competing operators available is most suited to their needs
- To act on behalf of subscribers to impose performance standards where necessary on services where competition is not yet effective

The Authority regards the first of these goals as the more important in the initial stages of the introduction of competition into Bahrain since until competition is effective, the second goal will have no data to support it.

Service measurement is a complex task and in many countries the adoption of a fragmented approach has meant that the primary goals outlined above are not being met. The Authority is keen to see a single, consistent and agreed approach to service measurement that is applicable to all service providers and which applies from the very start of competition in The Kingdom. In this way common forms of measurement will be used, data will be truly comparable and operators will understand from the outset what is expected of them (see Schedule 1 and Schedule 2 below).

The Authority is also aware that it is all too easy to impose heavy and complex measurement rules on the operators to produce data that is not entirely relevant to the primary goals defined above. To this end the Authority wishes to keep the measurement structure simple with a few strong principles running throughout. These principles will take as their baseline that any well run operator would wish to monitor and manage their performance quality effectively and efficiently and so the measurement requirements should not represent any significant additional administrative load to an efficiently run organisation.

#### 2.2 Stages on the route to managed performance

The Authority does not expect that telecommunications services in Bahrain will necessarily be able to meet the aspirations set for them immediately and a series of stages are proposed that will allow these to be met within practical periods and without wholesale disturbance to the infrastructure providing the services. Different services from different providers may well be at different stages of this process at the same time (see Figure 1).

The first part of the process is to introduce measurement and reporting of the current performance of each qualifying service. This is described in section 3.10 below.

Once the measurement and reporting is in place and can be seen to be producing reliable results, these will be compared with similar metrics from around the world and achievable long term targets for Bahrain, taking into account local conditions, will be

# CONSULTATION

## Quality of Service measurement and improvement

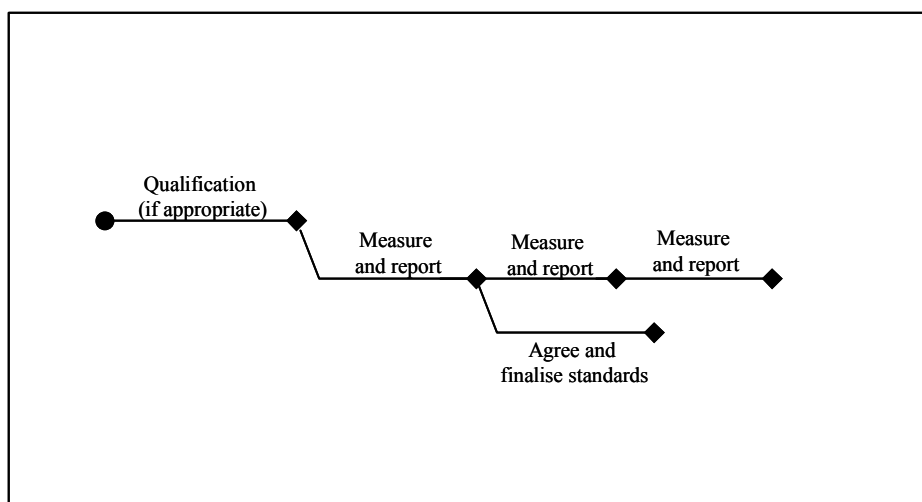
established and published. Initial proposals for these targets are given in Schedule 4 below and comments are invited as to their feasibility.

The Authority believes that with the introduction of competition, it will be sufficient for the competing operators to be aware of the targets that have been set for all of the operators to meet to impose the discipline of achieving them. Where competition is not effective (as for example where dominance occurs) the TRA may impose the aspirations as formal targets to be met.

However, the Authority does not rule out the possibility of introducing some further means of enforcing compliance should this prove necessary. This may take the form of imposed financial penalties for failing to meet the targets (see Schedule 3 below). In principle, the Authority would prefer to operate without the need for such penalties, since they remove funds from the operation of the networks, but will impose them if it finds that they are necessary.

Wherever possible the Authority will adopt a consistent approach to the introduction of measurements, whatever the service type.

Regulators in some countries have introduced the concept of a performance forum in which the operators, the users and the regulator meet regularly to discuss what can be measured and what information is required to be published. Having seen the results of this approach the Authority is not in favour of it. It is seen to lead to over-complex measurements and the provision of almost meaningless amounts of information to the public. The Authority wishes to retain a strong engineering-based approach to the whole process and concentrate on what it is practical and useful to measure at all times. Further discussion on this subject is given in section 3.8.



*Figure 1 Stages in setting performance standards*

### 2.3 Provision of information to users

In the early stages of the introduction of competition, the Authority does not believe that any useful purpose is served by publishing measurement information. For a significant period there will be no competition in many of the services and so the Authority proposes to use international comparators, suitably adjusted for Bahrain, as the aspirations to be aimed for.

# CONSULTATION

## **Quality of Service measurement and improvement**

Once competition is well established, the Authority will consider publishing some subset of the measured information provided that it serves the purpose of enabling users to make informed decisions about which service to use. Given the likely lack of technical knowledge of users making these decisions, such published information will need to be highly edited and presented in a form suitable for the general public to use.

Some performance information is required by the ITU for publication in its reference documentation. The Authority will meet its obligations in this respect, where the information is available.

# CONSULTATION

## Quality of Service measurement and improvement

### 3 The measurement of service performance

This section of the consultation document, together with the Schedules forms the proposed QoS measurement and control method that is being submitted for consultation. It takes the form of a number of Conventions that define matters of principle against which the measurement process will be developed. The Conventions and Schedules will be reviewed on a regular (at least annual) basis by the Authority and any proposed alterations submitted for consultation in a similar manner to the current document.

#### 3.1 Convention 1 Responsibilities of the Authority

In meeting its responsibilities under the Law, the Authority will:

- Set a standard framework for the measurement of services for all operators to follow
- Define realistic short and long term performance targets in pursuance of its legal obligations
- Consult with interested parties on methods of measurement and the targets that are established
- Compare performance of services provided in the Kingdom of Bahrain against international benchmarks
- Ensure that the measurement of performance of services does not impose an unnecessary additional load, over and above normal good practice, on the operators that deliver the services
- Introduce appropriate measures to ensure that performance targets are met both in the short and long term. This may involve the introduction of sanctions, either for failure to provide the required information or for failure to meet targets (in the case where competition is ineffective)
- Publish comparative performance figures when such are available and of value to users in selection of services from operators

#### 3.2 Convention 2 Responsibilities of the operators

In meeting the Authority's requirements, operators who qualify, and whose services qualify in accordance with section 3.3 below, will:

- Set up measurement systems consistent with the framework proposed in this document
- Provide regular quarterly returns of the measurement results for all qualifying services to the Authority (see section 3.9 below)
- Aim to meet the short and long term service performance targets set by the Authority

#### 3.3 Convention 3 Qualification of operators and services

The Authority recognises that operators newly entered into the Bahrain market will be concentrating on building their infrastructure, setting up interconnection arrangements

# CONSULTATION

## Quality of Service measurement and improvement

and ensuring their channels to market work. It accepts that this is in the interest of the introduction of competition into Bahrain and therefore proposes a period of exemption for new entrants from the provision of performance measurements during this period of establishment. This period will be one year from the award of the licence to an operator, and the first return of performance measurements will be expected at the end of the first complete quarter that occurs 12 months after the award of the licence (see section 3.9 below).

It is also important, for the achievement of statistical significance, that the population of repeated similar services is large enough for the measures to be meaningful in relation to the number of events being measured. There are many different approaches to this threshold in the international community, based on a range of parameters. The base reason for setting the threshold is for statistical significance and the Authority believes that this is most simply achieved by setting a numeric threshold of the number of services in use. This number (the Qualifying Population – see Schedule 2.17 below) will be 500 and will count the number of statistically independent instances of the service that are in operation. This may, or may not, equate to the number of users of the services. As soon as any given service or service group achieves this number it will become liable for reporting performance measurements. This figure applies equally to services from all operators.

The Authority reserves the right, should it become necessary, to impose measurement and performance requirements on services that do not meet the qualifying population. This may arise where a service is of a particularly critical nature – such as an interconnect point. It may also arise where artificial service differentiation is used in order to avoid exceeding the qualification threshold.

### 3.4 Convention 4 Service types

Attachment A 1 below contains a background paper on the categorisation of services in telecommunications. The Authority proposes to use this categorisation as a means of structuring and grouping services into similar types to which similar measures may be applied. Taken together with the types of measures defined in section 3.6 below, a relatively simple measurement class model emerges that is shown in Figure 2.

All services will be assigned to a layer (row in Figure 2) and within that layer measurements will be applied in a consistent manner. Within any one layer, measures will be one of the four types defined in section 3.6 (columns in Figure 2). Within a column, a similar approach will be taken throughout for that type of measurement.

In this way the Authority expects a simple framework to be applicable to a wide and complex technological structure and further expects that this structure will be able to accommodate new, and as yet unknown, services without major reconstruction of the measurement system.

For specific services, the measures will always be conducted in the same manner, but the performance targets may vary. The targets are specified in the Schedule 4 below.

As far as possible, services will be categorised in the largest possible groups consistent with meaningful measurement. Sub groupings will only be used where there is an identifiable and significant difference in performance expected from the sub group.

# CONSULTATION

Quality of Service measurement and improvement				
	Implementation	Operation	Performance	Other
Application	Implementation SLAs	Operational SLAs	Performance SLAs	
Network				
Grooming				
Bearer				
Raw b'width				
PoP				

*Figure 2 The measurement class model*

### 3.5 Convention 5 Market types

In some administrations services are measured and reported on in groups relating to different markets. These include retail, residential, business and interconnect services. The Authority sees no need to differentiate on measurement and reporting by market. If a service is supplied to several sectors, then its measurement should be across the whole population with no requirement for sub division.

If a service is designed and provided for a single market sector, such as the business sector, then the service will stand alone as a result of the different performance that it offers, not as a result of the market it serves.

### 3.6 Convention 6 Performance types

Figure 2 shows four types of measurement of fundamentally different type:

- Implementation
- Operation
- Performance
- Other

#### *Implementation*

Implementation service measures relate to the response by a service provider to the receipt of a request for a service to be delivered. Where this involves a major infrastructure build with special plans and procedures, the Authority does not regard the provision of service measurements as beneficial. Normal processes of project management and reporting are perfectly satisfactory and would be set up on a case by case basis outside the QoS measurement regime.

Where a service can be regarded as a commodity – for example the provision of a new telephone or a leased line – then Implementation service measurement is appropriate.

# CONSULTATION

## Quality of Service measurement and improvement

### *Operation*

Operation service measures relate to the Availability of the service or the ability of the service operator to keep the service fault free and to mend it when it fails. Annex A2.9 contains a description of what falls within the definition of Availability and what does not, and a formal definition is given in Schedule 2.9 below. Performance related failures specifically do not fall into this category, neither do issues of geographic coverage.

### *Performance*

Performance measures relate to the objective performance delivered by the service under specific load conditions where that performance is not specified by the service itself. For example, a leased circuit rated at 2Mbit/S will always deliver that line speed or it will be considered to be faulty. There is no requirement for a Performance measure in respect of such a service. However, a packet switched IP service, that shares its infrastructure among a number of users, would have a variable performance depending on the load to which it is subjected. Services such as this do require a Performance measurement.

### *Other*

The Other category of service measurement refers to the softer aspects of services such as customer satisfaction and billing.

The standard measures that will apply to each layer and each measure type are given in section 3.7 below. The targets for the individual service types are specified in the Schedules. It should be noted that at this stage the targets are mean values in the terms of the definitions in Attachment A 2. When thresholds are eventually set against a formal measurement process, these averages will be adjusted for the variability of measurements across different measurement periods in accordance with the mathematics in Attachment A 2.

## **3.7 Convention 7 Common approach to measures**

This section defines the range of parameters that will be used to measure services. Two of the following categories cover whole vertical columns in the class model (Implementation and Other), while the others are grouped by technology layer and then sub divided where necessary into the appropriate measurement type. Definitions of the measurement methods and the parameters are given in Schedule 1 and Schedule 2 below. The Authority will keep the measures under review and may propose additional measures or alterations to existing measures should this become necessary.

### *Implementation*

This set of measures applies to all technology layers. Where a service can be considered to be a commodity (i.e. there is no requirement to set up special and distinct plans or processes in order to deliver it), the following measures will be maintained for all such commodity services:

- Initial response time: The time between the service provider becoming aware that a requirement exists to the receipt by the client of a firm offer of a delivery date

# CONSULTATION

## Quality of Service measurement and improvement

- Delivery performance: The percentage of deliveries made on or before the promised delivery date and the percentages that are late by daily increments until the whole population is accounted for.
- Delivery time: The time between the client confirming the order and receiving a satisfactory service for the first time.

### *Other*

This set of measures also applies to all technology layers and addresses some of the 'softer' issues surrounding service delivery. Three sets of measures will be maintained in relation to qualifying services under this category:

- Complaints: A log of complaints together with their time to satisfactory resolution will be maintained. These will be analysed by service type
- Billing: A similar log of billing complaints will be maintained, also with their time to satisfactory resolution, by service type.
- Customer satisfaction: On an annual basis, or such other period as the Authority may direct, the service providers will conduct a user satisfaction survey through a third party organisation using a questionnaire/script approved by the Authority.

The following measures are proposed for specific layers in addition to the Implementation and Other types mentioned above.

### *PoP Layer*

The Authority does not wish to be proscriptive in this layer and no additional mandatory measures are proposed. Should operators wish to enter into service agreements at this layer, the Authority would suggest the following categories of measure:

- MTTR of any maintenance services offered
- Availability of any core services such as power and air conditioning that are provided
- Security of the PoP in terms of the numbers of breaches that occur to the security controls of the PoP

### *Raw bandwidth*

Only Operation type additional measures are proposed for this layer, and they include no more than two of:

- Availability: The measured Availability of the population of similar service types
- MTBF: The measured Mean Time Between Failures of the population of similar service types
- MTTR: The measured Mean Time To Repair of the population of similar service types

The reason that only two of the three possible measures are required is that the third measure is always a function of the other two and the third will be assumed to be the value derived from using the formulae in Schedule 2.9 below.

# CONSULTATION

## Quality of Service measurement and improvement

### *Bearer*

#### Permanent services

For services of this nature, only Operation type measures are proposed using the same parameters as for Raw Bandwidth above.

#### Switched services

For services delivered as a result of switching, such as ATM or IP Switched Virtual Circuits (SVCs), then in addition to the measures for Permanent services specified above, the following measures will be taken:

- Probability of blocking: The probability that, in the peak hour, a service will not be available as a result of blocking of the capacity by other users (see ITU-T X.131)
- Probability of loss of circuit: The probability that, in the peak hour, a circuit that has been established might suffer from a premature drop (see ITU-T X.136)

### *Grooming*

Several measures are proposed for Grooming services. Operation type measures will apply to all services and, as before they will include no more than two of:

- Availability: The measured Availability of the population of similar service types
- MTBF: The measured Mean Time Between Failures of the population of similar service types
- MTTR: The measured Mean Time To Repair of the population of similar service types

In addition to the Operation measures, Grooming services will also have the following Performance measures applied to them:

- Packet Delay (Mean and 95<sup>th</sup>ile): The average and 95<sup>th</sup> percentile delay that would be experienced by a representative number of 128 byte packets crossing the service in the peak hour (see ITU-T X.135).
- Packet Loss: The average packet loss in the peak hour expressed as a percentage of total traffic submitted.
- Jitter: The Standard Deviation of packet delay during the peak hour.

For services set up by switching (Switched Virtual Circuits), the following measures will be applied:

- Probability of blocking: The probability that, in the peak hour, a service will not be available as a result of blocking of the capacity by other users (see ITU-T X.131)
- Probability of loss of circuit: The probability that, in the peak hour, a circuit that has been established might suffer from a premature drop (see ITU-T X.136)

# CONSULTATION

## Quality of Service measurement and improvement

### *Network*

Network measures will be the same as those for Grooming.

### *Application*

The Authority does not propose to apply measures to application services on the grounds that these are generally not within the scope of telecommunications services. It reserves its position and may introduce them should this basic assumption alter at some time in the future.

### **3.8 Convention 8 Use of industry forae**

It is the practice in some Administrations to support industry forae attended by users and the operators. In the opinion of the Authority this approach tends to lead to muddled thinking and the documentation produced is by and large of little value to users and has little technical rigour. The Authority would prefer, at least in the early stages of the introduction of competition, to concentrate on well defined engineering measures of performance which any well run operator would wish to produce in any case. These measures will be used to ensure that Bahrain receives the best possible service in comparison with international achievements consistent with any overriding local constraints.

The data provided by the operators will be made available to all service providers so that they can see what improvements, if any, are needed in their own performance to be able to compete in the market.

The Authority will, from time to time, conduct independent measurements of the services provided in order to confirm, or otherwise, the accuracy of the data provided by the operators.

### **3.9 Convention 9 Provision of information to the Authority**

The Authority recognises that the measures provided under this regulation may have some commercial sensitivity. However, it takes the position that the open availability of such information is essential to the nurturing of competition and it is only by making such data available that the beneficial effects of competition will be felt by all of the suppliers.

Where a qualifying operator fails to provide the necessary data, the Authority may, after due warning, make arrangements for a third party to carry out independent measurements at the expense of the failing operator.

Information will be provided to the Authority by all qualifying operators in respect of all qualifying services on a quarterly basis. These returns will fall due at the end of March, June, September and December in each year. They will be provided in both hard copy and electronic form so that they can be edited into a consistent format for publication.

As the measurement process settles down, the Authority will maintain it under review and may consider extracting user relevant information to assist users to make informed choices between different service providers. Should this prove necessary and beneficial, the Authority will consult before taking action to follow this path.

# CONSULTATION

## Quality of Service measurement and improvement

### 3.10 Convention 10 Stages of development

The Authority recognises that service measurements taken initially may not meet the standards that are internationally recognised as satisfactory so it proposes three stages of development of the measurement process leading to a long term steady state system. The target figures that are given in the Schedules are the initial suggestions for the steady state performance measures.

The stages of development are:

- Measure
- Baseline, compare and set standards
- Meet the required performance standards

Different operators and different services may be at different stages in this process at different times, but the Authority is keen to see all qualifying services reach the steady state as rapidly as possible.

#### *Measure*

This stage consists of the establishment of measurement systems and the provision of returns to the Authority. During this stage, no targets will be applicable. The purpose of this stage is to establish accurate and dependable measurement systems and to generate an initial body of reliable data. The time spent by any service in this state may not exceed two quarters.

#### *Baseline, compare and set standards*

Once two quarters of data exist for a service, they will be compared with the targets set in the Schedules to this document. In consultation with the service provider(s), any difference between the achieved performance and the target performance will be considered and, if necessary, the target itself may be adjusted. Targets will be set at the same level for the same service types across all operators.

#### *Meet the required performance standards*

The Authority wishes to avoid the need for penalties to be introduced to enforce adequate performance. Provided that progress is made towards achieving the target measures, then the Authority would wish to allow competition to be the main method of enforcement of performance. Performance targets will be published and it is expected that this will put pressure on operators to meet the agreed targets.

Where the Authority considers that competition is not effective (either as a result of an operator being declared dominant or for other reasons), the Authority may consider the introduction of sanctions should the performance targets not be met.

Should this competitive approach not prove to be effective, then the Authority reserves its position and may consider the wider introduction of targeted sanctions or service credits to ensure progress towards the provision of high quality services.

# CONSULTATION

## Quality of Service measurement and improvement

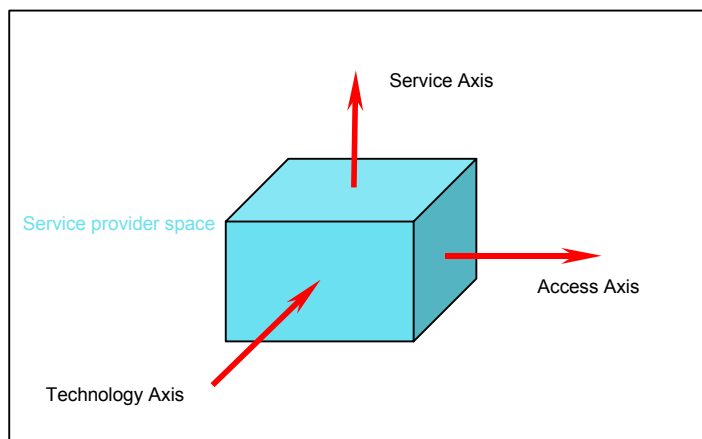
### A 1 Background paper on service model approach to measurement

This paper was written by Intercai Mondiale Ltd as an annex to a treatise on Service Level Agreements (SLAs). This copyright material is used with the permission of the owners with minor editing to relate it to the consultation. It does not form a part of the long term QoS requirement, but provides a rationale and framework for the approach suggested by the Authority.

#### A1.1 Service types

Telecommunications and information services are complex integrated constructions of technology, functions, services, processes and topography that are difficult to describe clearly without a structured approach. In this Attachment the Authority sets out to constrain the complexity by imposing a structure on the visualisation of telecommunications systems. Whilst this still leaves grey and overlapping areas, it does help to visualise the constituent elements and in particular it makes it easier to apply a common approach to measurement across a wide and changing range of technologies. The Authority believes that such a common approach is of value to operators and users alike in that it reduces the range of different approaches that can be taken and ensures that the original purpose of the taking of measurements does not become lost in the need to produce greater volumes of reports.

Figure 3 shows the manner in which the service provider space is presented in this consultation document. The technology axis comprises the common underlying technologies that underpin all telecommunications and information services. The access axis addresses the many different ways in which users can gain access to the underlying core technologies. The service axis represents the element that the user sees.



*Figure 3 Approach to analysis*

It is the service axis on which the approach to service measurement is based in this Attachment.

As an aid to the visualisation of the relationships between the technologies, Figure 4 shows, in a hierarchical manner, the different architectural levels in a telecommunications network. Whilst not intended to be exhaustive, it can be seen from this diagram where particular technologies fit. It can also be inferred where new technologies will fit and hence the approach that will be taken by the Authority to the measurement of service for that technology.

# CONSULTATION

## Quality of Service measurement and improvement

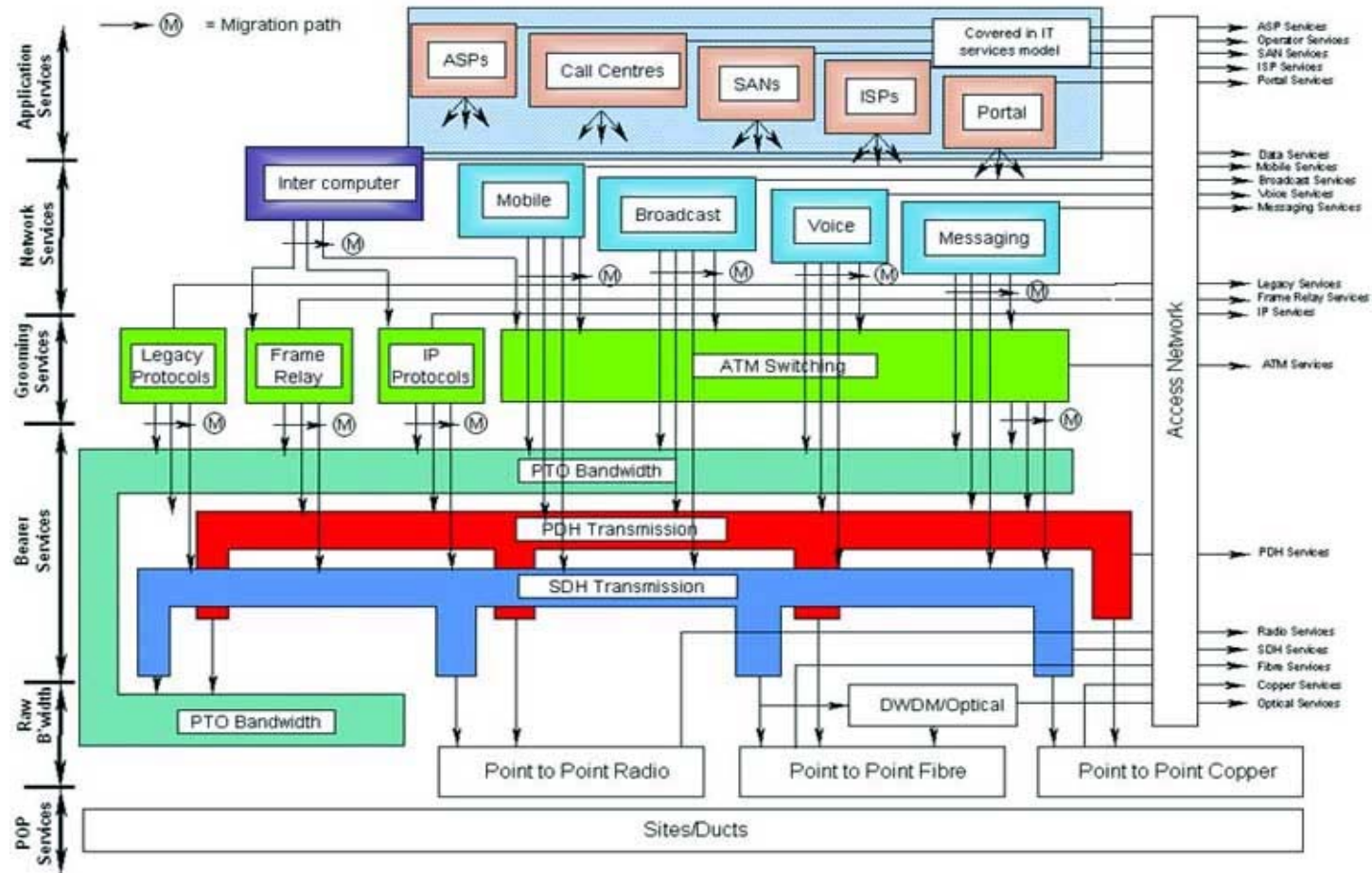


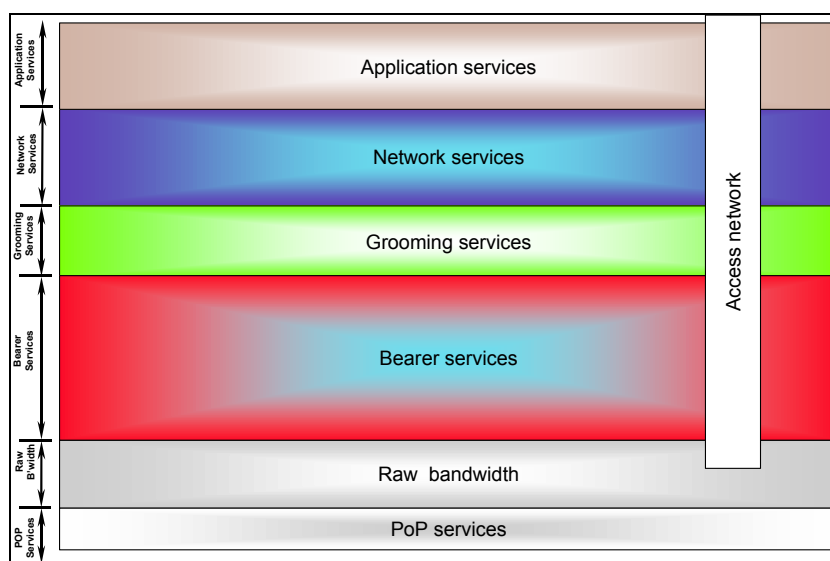
Figure 4 Technology architecture

# CONSULTATION

## Quality of Service measurement and improvement

### *Generic architecture*

A telecommunications service that is delivered to an end user will have been built from a whole range of different technologies. There is a hierarchical structure to the manner in which the different types of equipment are strung together to form services, and it is this hierarchy that is used as the basis for the service model derived in this Attachment. Figure 4 is a diagram that shows the relationship between the different types of technology used in telecommunications. Figure 5 shows a simplified version of that diagram with less detail to enable general principles to be illustrated without the very fine detail of the technology diagram. It will be seen that the interfaces, both vertical and horizontal, between the parts of the diagram are the natural boundaries at which services are handed off between service providers, or between service providers and their clients. It follows that they form the natural boundaries over which measurements apply. This feature enables us to collect together similar services (and hence similar measurements) for purposes of description.



**Figure 5 Generic technology architecture**

Figure 5 may be considered to be the plane at the technology face of the cube shown in Figure 3. It is divided along the left-hand edge of the diagram into six distinct and separate layers, each of which serves a useful purpose in the overall delivery of services to users. Starting from the bottom of the diagram, the functionality of each layer increases successively, resulting eventually in the application service delivered to the user. Not all of the layers need to be present in any given service structure, but the totality of the layers represents a general model for the application of measurements.

Services of different types can be picked out of the technology at any of the different layers and the Access Network box towards the right hand side of the diagram illustrates this process. The Access Network also contains different technologies to deliver services to users.

The highest layer of the architecture is at the top of the model – the Application Services. These perform the ultimate functions for which the network was built. This layer of the model includes end to end services and also what might be called internal network services such as Domain Name Servers (DNS) and Network Address

# CONSULTATION

## Quality of Service measurement and improvement

Translation (NAT). Services at this level are generally outside the scope of telecommunications services.

Each layer of Figure 5 fulfils a specific purpose and these are described in turn below.

### *PoP Services*

The lowest layer of the network hierarchy consists of the basic civil engineering on which the network is built. This includes the exchange or hosting buildings, the street-works (roadside cabinets) and the underground ducting or overhead distribution equipment that links them together. It also encompasses radio base stations and towers as well as data centres. Typical services in this layer of the model would include collocation services.

### *Raw Bandwidth*

This layer comprises the basic transmission medium that underpins the network. It may consist of copper, radio, microwave, laser or, increasingly, fibre. Different media provide different degrees of capacity, but the variability between the media is generally hidden from the user by the intervening layers. The Raw Bandwidth layer is typically built making use of the facilities of the PoP layer and interconnects the buildings or nodes in that layer. Typical services in this layer of the model would be the unbundled local loop, or dark fibre.

### *Bearer Services*

To provide a usable transmission service, raw bandwidth needs to have information transmitted across it and this is achieved by the Bearer Service layer. Typically this provides point-to-point operation over a fixed infrastructure of some form and is almost universally carried over the Synchronous Data Hierarchy (SDH) transmission protocol regardless of the underlying raw bandwidth. Increasingly Optical Processing is taking a role in this layer. Services provided at this layer are typically fixed capacity leased lines such as a 2Mbit/S E1 service.

### *Grooming Services*

Grooming services arose from the IT industry which found itself constrained to use low capacity, expensive, unreliable transmission media and developed a series of protocols for using this limited and costly resource efficiently. As voice and data services have converged and the demands of data have become more bandwidth intensive, the role of the grooming layer has taken on that of a convergence layer which also provides many to many connectivity over the fixed infrastructure provided by the Bearer layer. Unfortunately competition between the suppliers and weak or almost non-existent regulation of the IT industry has led to there being many different competing technologies within this layer and it is here that many of the requirements for performance measures start to emerge. The biggest difficulty facing this layer of the measurement model is the estimation of traffic levels. Rapid change in the applications that use the infrastructure makes it impossible to define traffic levels in all but the most constrained of circumstances.

Grooming service technology can also be used to provide bearer like services but they are categorised in this layer for purposes of measurement. Typical services provided in this layer are IP or ATM switched or permanent circuits.

# CONSULTATION

## Quality of Service measurement and improvement

### *Network Services*

There is an overlap between Network Services and Application Services and the boundaries between them are blurred. Network services are real applications that are delivered as a result of the inherent underlying functionality of the network, rather than being derived from some overlay. Network services include fixed and mobile voice services.

### *Application Services*

Application services are divided, for purposes of this annex, into two sub-layers. The first of these is the set of applications needed to run and manage the network itself. They include the features and facilities necessary to design, provision, operate and charge for the network.

Secondly, at the top of the hierarchy, are the specific applications services that can provide the recognisable service for which the user has contracted. These can include services such as e-mail, databases and function specific applications.

### *Access*

The last box in Figure 5 relates to access services. These are the technologies that enable users to gain access to the services at each layer. A typical example might be xDSL, which connects an end user to an ISP, say. In some layers of the model the access device has its own dedicated technology that is access specific. In others access is inherent in the layer itself (such as the PoP layer) and there is no specific access technology to describe.

Much of the technological change that has taken place in the transmission layer has happened as a result of investments from the telecommunications industry. Much of that in the application layer has come from the IT industry. In the middle, it is the grooming layer where these two forces come together and where the major ATM v IP debate is taking place.

The Authority is concerned to ensure that the model used to derive a structure for the measurement of services does not rest on a single technological snapshot that will date rapidly and become ineffective. For this reason it has sought a way of representing technology, and the measures that apply to that technology, that enables them to be considered in generality and, to a large degree, independently of the technology that delivers them. The model is shown in diagrammatic form in Figure 6.

This diagram shows layers of technology around which the service provision industry organises itself in the rows (derived from Figure 5), and types of measure that apply to the layers in the columns. Within each cell of the table, the measures that apply to the services within the cell are broadly consistent and technology independent. This means that there are few varieties of measures and those that are needed are easy to understand and administer. The Authority proposes that all telecommunications services will be considered to be within one of the six layers of the technology model as far as measurements are concerned.

With the exception of the Grooming and, to a lesser extent, the Network layers, all of the other cells can have their measures expressed simply and consistently. This arises because the service that is being supplied in most of the cells is very well and accurately specified. Where difficulties arise is in the layers that have traffic dependent performance, where the measures are more difficult and additional measures need to be taken.

# CONSULTATION

## Quality of Service measurement and improvement

Four classes of measure are identified for purposes of this consultation, namely:

- Implementation
- Operation
- Performance
- Other

	Implementation	Operation	Performance	Other
Application	Implementation SLAs	Operational SLAs	Performance SLAs	
Network				
Grooming				
Bearer				
Raw b'width				
PoP				

*Figure 6 The measurement class model*

Analysis shows that measures can be well behaved in all Implementation and Operation cells and that most of the difficulties that arise relate to the Performance measures that apply to the Grooming and to some extent Network layers of the model. In these layers there is no prospect of any guaranteed level of performance if the infrastructure has to be shared with other users without some means of protecting the resources allocated. For example, techniques using ATM can provide this protection, whereas those using IP cannot. IP technologies are still immature in very significant areas such as Flow Control, Security, Addressing, Traffic Engineering, re-Transmission and Performance and means of addressing these issues are being introduced into networks on a piecemeal basis.

A tabular version of the model is shown in Table 1 and the ticks indicate where the Authority believes that measurements of any particular type are relevant. Ticks in brackets show where measures are practical but the Authority does not, at this stage, propose imposing them.

**Implementation** measures are concerned with the initial delivery or configuration of services and although this can apply to large bespoke projects, the Authority does not believe that measures are either appropriate or necessary in such cases where normal project management techniques are more than adequate. In this model the Authority considers these measures to apply to multiple instances of commodity services (e.g. delivery of a leased line) where existing design and installation processes should deliver a consistent delivery performance.

# CONSULTATION

## Quality of Service measurement and improvement

**Operation** measures apply to the fault performance of the service. These include parameters such as failure rates and repair times and measure the long-term steady state Availability of the operational status of the service.

**Performance** measures apply to the capacity of the service to deliver consistent performance under defined load conditions. These measures are mainly applicable at the Grooming layer where multiple users share the same infrastructure and there is the possibility that different traffic streams will interfere with each other.

**Other** measures embody those types of special measure that fit nowhere else. In this category we would include measures such as security, cleanliness, customer satisfaction, billing accuracy and quality of reporting.

Service type/SLA type	Implementation	Operation	Performance	Other
Application		(✓)	(✓)	(✓)
Network	✓	✓	✓	✓
Grooming	✓	✓	✓	✓
Bearer	✓	✓		✓
Raw bandwidth	✓	✓		✓
PoP	(✓)	(✓)		(✓)

*Table 1 Table of measurement types against service types*

The Authority does not consider it appropriate to apply measures to services that are in a state of flux or initial roll out before a steady state has been achieved. Problems occurring during the roll out of a new service are primarily of a development nature and should be treated as such. It is proposed that measures taken should apply to services that are in a steady state of operation.

In many of the cells of Table 1, the service is extremely well specified. A leased line bearer, for example, has an internationally agreed specification and the only measure that is appropriate to it in the steady state is whether it is delivering its service or not. This is why the Performance measures do not apply to all of the cells in Table 1.

# CONSULTATION

## Quality of Service measurement and improvement

### A 2 Background paper on service level mathematics

This paper was written by Intercai Mondiale Ltd as an annex to a treatise on Service Level Agreements (SLAs). This copyright material is used with the permission of the owners with minor editing to relate it to the consultation. It contains useful background material on the mathematics of measurement and the setting of service level guarantees, together with the impact of statistical significance. The contents is not directly applicable to the Quality of Service consultation, but the Authority gives notice that it will be using techniques similar to those described herein to calculate the target performance thresholds. Similarly, when the approach to QoS measurement matures in Bahrain and guarantee levels are introduced, the same mathematics will be applied.

Much of the mathematics used in this paper is also used by the ITU in its recommendations on performance as applied to packet switching. The Authority believes that these standards still have general, if not specific, applicability to modern systems and may well refer back to them as source material. These standards include:

- X.131: Call blocking in public data networks when providing international synchronous circuit switched services
- X.134: Portion boundaries and packet layer reference events: Basis for defining packet switched performance parameters
- X.135: Speed of service (delay and throughput) performance values for public data networks when providing international packet-switched services
- X.136: Accuracy and dependability performance values for public data networks when providing international packet-switched services
- X.137: Availability performance values for public data networks when providing international packet-switched services
- X.138: Measurement of performance values for public data networks when providing international packet switched services
- X.140: General quality of service parameters for communication via public data networks.

#### A2.1 Introduction

This Attachment addresses the mathematics and statistics that apply directly to the setting, measurement and operation of Service Level Agreements (SLAs) in the Operational area. It does not include the mathematics of Performance design, but focuses on the implications for SLAs of sampled measurements of parameters with inherently random values and infrequent occurrence rates. Performance mathematics follows similar principles and will be used by the Authority where relevant.

In order to provide a sound basis for the remainder of the Attachment, a brief overview of random events and the mathematics that describe them is given. This does not purport to be a detailed exposition of the subject, but is intended to provide an intuitive feeling for it and at the same time to provide some rules of thumb that enable practical engineers to work effectively in the field.

The practical measurement of SLA parameters involves the regular sampling of specific values such as traffic carried or network faults. The sampling process can

# CONSULTATION

## Quality of Service measurement and improvement

affect the values collected and this effect needs to be taken into account in setting SLA parameters. Typically the impact of sampling is much greater on parameters that happen infrequently (e.g. faults) than on those which occur in large numbers (such as computer based transactions). The mathematics involved in calculating these effects is addressed in this Attachment.

In modern commercial SLA parlance, the term Availability is used in a loose engineering sense and actually includes several mechanisms, each subject to different analytical techniques. In this Attachment we define the terminology that we are using and show how each aspect of the mechanisms that contribute to 'unavailability' can be analysed.

Once the firm basis of definitions and techniques has been established, we provide some guidelines as to how to set performance thresholds, and the risk that is being taken when they are set in this manner. Intercai has built a simple illustrative model of the impact of sampling on service level risks, which is described in the Attachment. This model can be made available to the consulted parties should this be thought necessary.

Finally we provide a short list of techniques that can be used to overcome some of the issues that are raised by the mathematics of SLAs.

### A2.2 References

Reference 1: Systems Analysis for data communications systems. James Martin

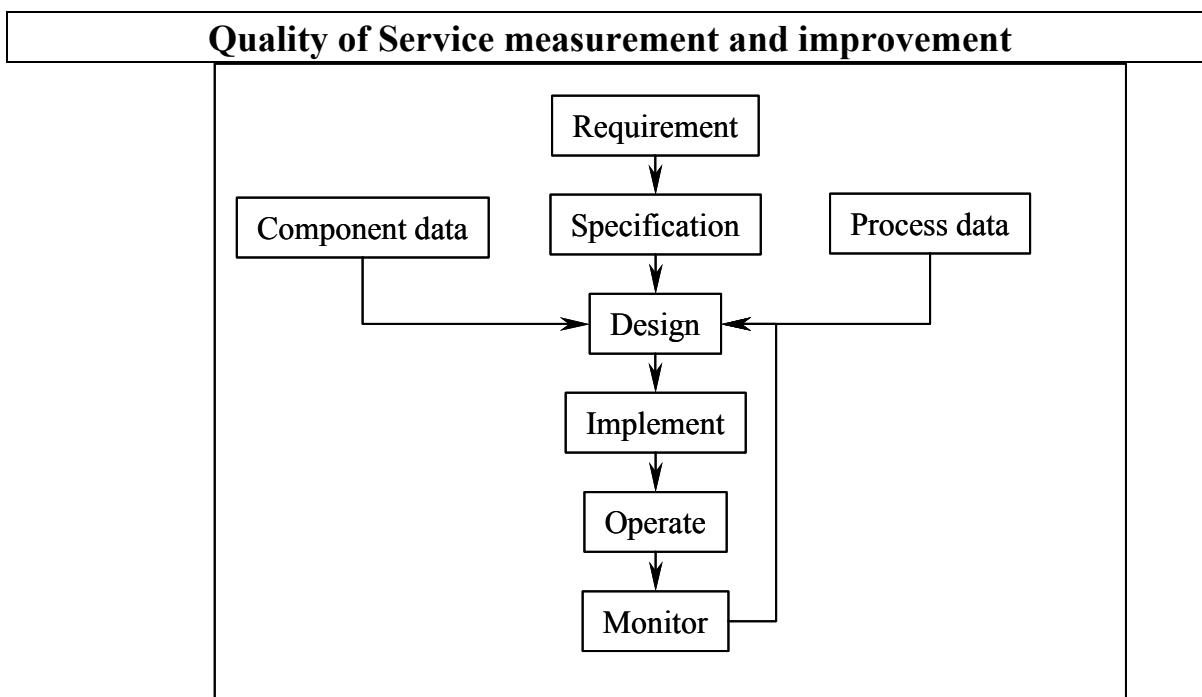
Reference 2: <http://davidmlane.com/hyperstat/A14043.html>

Reference 3: <http://www.rand.org/methodology/stat/applets/clt.html>

### A2.3 Engineering design process

In an ideal world a system, consisting of a set of components and processes, would be designed to meet a number of specific requirements laid down by the ultimate user. The design process is outlined in Figure 7 and is applicable to all types of engineering design, regardless of the application or the technology.

# CONSULTATION



*Figure 7 The 'ideal' design process*

In this process, the end user specifies what is required of the system at the user level, in terms that they can understand. In response to this the designers can generate an engineering specification that describes the requirement in terms that can be engineered to, and, importantly, can be tested against. There may well be iterations of the process at this stage to trade off, say, performance against cost.

Working from the specification, designers can then develop a system design and plug in knowledge about the performance and loading of the components that contribute to the design in order to establish whether the design will meet the requirement. Following completion of this process, the system is built, tested and operated and some monitoring process provides information about how well it is meeting the requirement. A good system design process will then allow for improvements (which can include cost reductions) to take place in response to the performance measures being taken. A good design team will expect their system to perform to the level that they specified and there should be no surprises when the performance comes to be measured.

As we said above, this all occurs in an ideal world. In real life, there are fewer and fewer operations that can afford to pay for designs to be used solely for their own purposes. These opportunities are limited to major bespoke procurements where design visibility is easy to impose. There is a significant trend towards outsourcing those parts of the design that could be considered to be commodities and that deliver some cost benefits from being designed and delivered in bulk from a single source. In such cases, the user loses visibility of the design process and can no longer impose an overall requirement specification. It is more likely that he will be offered a range of services from which he selects the performance levels that are closest to what he needs and accepts that this may involve some sacrifice in order to achieve the potential savings.

Once a user has adopted this approach, then the visibility of the design process that is implied by Figure 7 is lost and some alternative means of gaining confidence in the delivery of the requested service is needed. This is the function of the Service Level Agreement (SLA). It attempts to provide a control framework that provides the user

# CONSULTATION

## Quality of Service measurement and improvement

with some degree of confidence that his requirement is being met. Unfortunately, this disconnection of the user from visibility of the design process provides an opportunity for the marketing function to take control of SLAs. There is a trend in the market place for the engineering basis of SLA thresholds to disappear in favour of their use as a competitive tool. Consequently there are SLAs being offered by some suppliers that owe more to the need to compete than to any sound engineering foundation.

It is important to be aware that the figures that are specified in an SLA are not those that define the actual service to be delivered. It is the performance that the Service Provider is prepared to guarantee and there is a significant difference between the two. Given the advent of marketing control of SLAs and, in some instances either ignorant or cynical suppliers, our point is that the SLA on its own is not to be trusted without some further investigation on the part of the user.

It is also true that if the supplier cannot deliver against the SLAs that have been put in place, there is little option to follow the 'improvement' loop in Figure 7 because the costs to the supplier could be out of proportion to the revenue from any given user. The trend, therefore, is for suppliers to live with whatever performance their system actually delivers and then make payments in compensation to users whose requirements are not met, since this is cheaper than attempting to re-engineer their system to meet the original needs.

It is important, therefore, that users and potential users of systems make every effort, prior to procurement, to ensure that the performance of the system will meet their needs because subsequent correction will prove to be extremely difficult.

In summary, the SLA is a form of substitute for the visibility of engineering rigour in the design process and it is worth bearing this in mind as the rest of this report unfolds. If we assume that the original requirement document was a rational and accurate description of the real need (and this is not necessarily always the case) then the user is more interested in achieving the performance he actually requested than in receiving some form of financial compensation for a less than perfect service. The migration to this type of process puts a greater onus on the designers at the original specification stage since there is a much-reduced opportunity to correct for early errors through the monitoring and correction loop. If an SLA can be agreed with the necessary degree of design visibility, then SLAs can have a role to play in controlling the quality of delivered service performance.

### A2.4 Parameters with distributed values

We have set the scene for SLAs in general in the preceding sections. They are set up as a mechanism for measuring the performance of some system in order to provide confidence to the user that the intentions of the original requirement are being met. In real systems, though, measures of performance are not consistent and they often exhibit variability of behaviour from one measurement to the next. Mathematical techniques have been developed to describe such behaviour and it is on these techniques that the basis of the mathematics of this Attachment rest. It is important first, though, to ensure that the measures being taken apply to a single population of broadly homogeneous services. Whilst it would be possible to apply them to any mix of services, we are interested in extracting information about the characteristic performance of services and we can only do that if each set of measures applies to a population of consistently performing services. The first requirement necessary in any SLA activity, therefore, is

# CONSULTATION

## Quality of Service measurement and improvement

to draw boundaries around groups of services that can be considered to perform in a similar manner so that useful conclusions about their performance can be drawn.

Now, if a specific parameter (say MTBF) is measured regularly in a repeatable manner for such a group of services, we would expect to get a list of numbers that varied about some average value. If the performance were truly repeatable then we would expect the measures to have some consistency about an average value and be able to state that consistency in a mathematical expression. This is what the mathematics of statistics provides and a good introduction to the subject will be found in Reference 1, from which much of the mathematics in this Attachment is derived.

The mechanism we are about to describe occurs in many instances in normal life. For purposes of description and illustration we use the example of a journey to work, undertaken on a regular basis with all the uncertainty and variability introduced by traffic, the weather and the time of departure. The 'service' meets our criterion of homogeneity in that it always delivers the same result, albeit with different degrees of performance. If we now define the measure of performance of the journey to work as the time of arrival and we plot the arrival time over a long period of time (say a year or more) we will have a population of results from which some conclusions can be drawn.

First of all we can calculate the mean value of the whole population of measurements. This defines the average time of arrival over the whole year's journeys. There is more useful information in the sample than this. If we were to take each individual journey and count the number of times that the time of arrival fell within a given timeslot, then we could plot a Probability Distribution Function (PDF) which would tell us something about the variability of the times of arrival of the journey. The PDF would probably be fairly irregular and may be difficult to analyse mathematically. Real life PDFs are generally complex but they can be described by characteristics of the shape that are known as Moments. The First Moment is the best known and is the Average, or Mean, value. The Second Moment is also quite well known. It is called the Variance and is a measure of the width of the distribution. A derivation of the Variance, which is even better known, is the Standard Deviation (SD). This is the square root of the Variance.

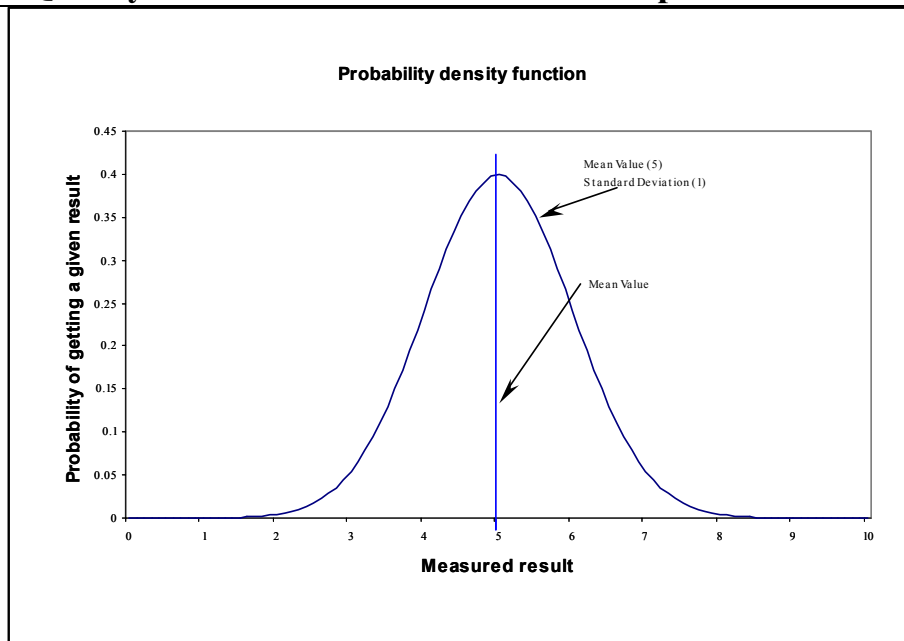
There is an infinite set of moments that can be used to describe a PDF, but only the first few have intuitive meanings. The Third Moment is a measure of the symmetry (or otherwise) of the PDF and is called the Skew. The Fourth Moment describes the 'lumpiness' of the PDF and is called the Kurtosis. The series carries on to infinity but these are the most well known Moments. Unfortunately the mathematics for operating with distributions with multiple Moments is extremely complex.

However, there is a very convenient theorem in statistics that simplifies the whole process of sampling and provides us with mathematical tools that enable us to analyse and describe the performance of systems of this nature without resorting to complex mathematics. If we group the events into sets of samples, say a week or a month's worth of journeys in our example, and take an average for each set and then plot these averages as a PDF, then the Central Limit Theorem (CLT) states that the resulting PDF approximates to a Normal Distribution and that this approximation improves in accuracy as the number of measures in each mean sample increases (see Figure 8).

The URL in Reference 2 provides a formal definition of the CLT, while Reference 3 provides an applet that illustrates the effect of statistical significance in a clear, dynamic and graphical manner.

# CONSULTATION

## Quality of Service measurement and improvement



*Figure 8 The probability density function*

We do not intend here to prove the CLT. Suffice it to say that provided the rules about sample sizes are observed, then we can make use of an easily manipulated form of mathematics to describe the performance of our measurement systems. The 'normal' distribution is often called the 'bell' curve, a name arising from the shape of its PDF. Its great benefit is that it has a Mean and Standard Deviation, but it is otherwise symmetrical (no Skew) and has no secondary lumps (no Kurtosis) and can be described relatively easily. The PDF is generally drawn in a normalised form in which the area under the curve is always 1 since the sum of the probabilities of all the possible values the event can take must add up to exactly 1 (because the event does happen).

Taking the shape of Figure 8 and relating it to our example, the horizontal axis represents the time of arrival, and the vertical axis the number of occasions that a specific time of arrival occurs. Now for a practical number of events, the curve would not be smooth sided in the way that it is drawn, it would actually be slightly stepped for relatively small sample sizes.

There are different types of distributions that deal with discrete parameters that can only take certain values (such as a number of faults where the answer can only be an integer) or can be continuous (such as a Time To Repair, which can take any value at all). We do not propose to enter into this discussion here, but simply to present results based on the different types of distribution and use them to draw conclusions.

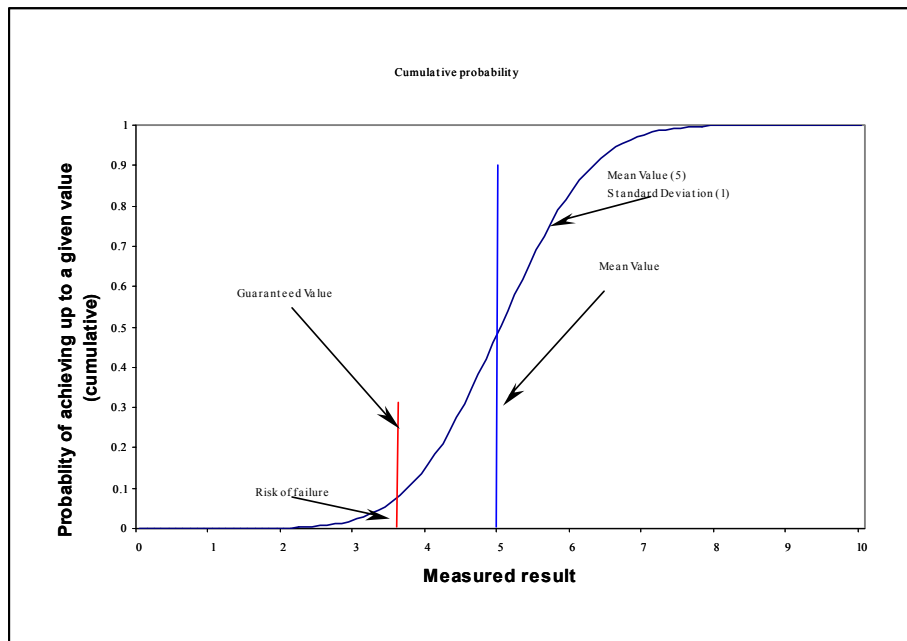
Returning to Figure 8 the highest point on the curve is the arrival time that occurs most often. For a Normal Distribution, it also happens to be the average arrival time and might be, for example, 9 a.m. The tails of the distribution show the probability of arriving at times other than 9 a.m. and give a picture of the variability of the distribution of arrival times.

Now a practical question that is often asked is 'what is the probability of arriving before (or after) a specific time?' To answer this question, the PDF is often drawn in a

# CONSULTATION

## Quality of Service measurement and improvement

slightly different form in which the PDF values are added together successively to form the 'cumulative' PDF whose form is shown in Figure 9.



**Figure 9 The cumulative probability function**

This form of the curve can have rising or falling values depending on whether the question addresses the probability of arriving before or after the defined time. The two curves are mirror images of each other.

In the curve of Figure 9, the probability of arriving very early is quite low. The value increases slowly and then more rapidly until the mean position of 9 a.m. is reached. At this point the probability of having arrived at some time up to 9 a.m. is 50%. The curve continues upwards slowing its rate of rise until the probability of arriving before some time late in the day becomes sufficiently close to 1 to warrant no further plotting of the value.

Suppose that we wanted to set a time at which we guaranteed to arrive at work. It is clear from an inspection that if we set it to 9 a.m., then the probability of arriving before (or after) this time is 0.5 or 50%. If we wanted to have a time that gave a better degree of confidence then we might select a time as shown in Figure 9 where the risk of exceeding that time was quite small. This is illustrated as the Guaranteed Value in the diagram and the area under the curve to the left is the risk of failing to meet that guarantee.

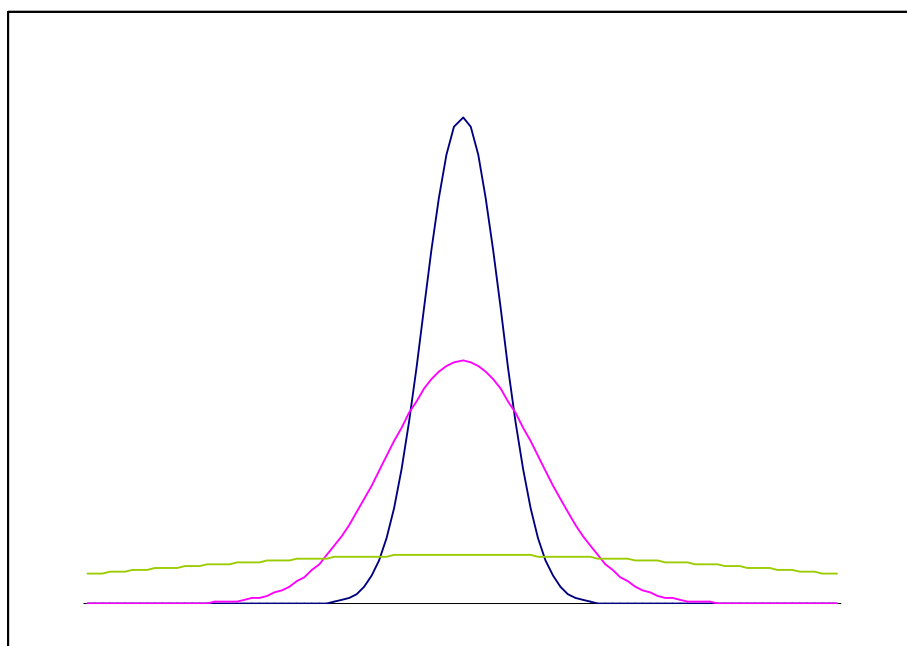
Now the Normal Distribution is sufficiently well behaved that it is possible to look up or calculate the areas shown and hence calculate the risk involved in providing a guarantee. It is the calculation of this risk level that is the basis for setting SLA thresholds and measurement standards.

There is one last characteristic of the PDF curve that we would like to use our example to illustrate. We have mentioned above the process of dividing the samples up into smaller groups over which the average is taken. As the groups get smaller (and ignoring the diminishing impact of the CLT) the Standard Deviation of the PDF curve gets larger. Figure 10 shows three PDFs, all with the same mean, but all derived from

# CONSULTATION

## Quality of Service measurement and improvement

different sample sizes. Relating this to the example of the journey to work, if we were asked to guarantee to arrive at work before a specific time of day averaged over a whole year, we may be fairly relaxed about it. If we were asked to make the same commitment averaged over a week, or even on a specific day, then we would probably decide to add some additional allowance to the journey time to be certain of meeting the guarantee. What we are doing instinctively is making an allowance for the statistical fact that over a small sample, the distribution is much wider and the risk of failure to meet a specific guarantee is higher than if we are allowed to average it over a longer period. We return to this feature in section A2.6 below and provide some rules of thumb for putting numerical values to this effect.



*Figure 10 Effect of sample size on distribution*

The effect of sampling on the measurement of SLAs depends very much on the size of the sample. The variability of a measure increases as the sample size decreases and this can have a significant effect on the risk of meeting any given test threshold. As the sample size increases, there comes a point where the impact of sampling can be assumed to be negligible and this means that measurements with very large populations (for example delay times for a specific transaction that takes place very frequently, like lottery ticket purchases), are substantially unaffected by sampling. As the sample size reduces (as, for example with a relatively small group of telecommunications circuits where the number of faults occurring is being measured in a finite time) then the effect is significant.

In practice it is only the parameters that happen broadly in 'human real time', such as building or repairing a system, that need to have this effect taken into account. This means parameters such as Delivery, Availability, Mean Time Between Failures (MTBF) and Mean Time To Repair (MTTR) are affected, while services that are carried out automatically, such as delay of a packet across a network, are not.

# CONSULTATION

## Quality of Service measurement and improvement

### A2.5 Test level setting and risk of failure

We have seen from the preceding sections that the probability of achieving a value of up to the Mean of a normal distribution is 50%. This means that if we were to set an SLA threshold at the Mean level, we would have a 50% chance of failing the test. Neither the supplier nor the user would be very impressed with a test that failed 50% of the time, so it is common practice to set the guarantee level at some value that is offset from the Mean in order to provide some higher confidence of passing the test.

In order to decide at what level the test threshold should be set, the organisation responsible for the test should decide what level of risk of failure is acceptable. This can be a broadly commercial decision, but there are some engineering factors that suggest where the risk level should be set. Referring back to Figure 9, it can be seen that the cumulative PDF curve is quite steep in its central sections, and becomes much flatter towards the edges. If the threshold is selected in the steeper central section of the curve, then small errors in the design calculations will produce significant alterations in the risk level. In the flatter areas, the threshold will be much more robust to errors of this nature.

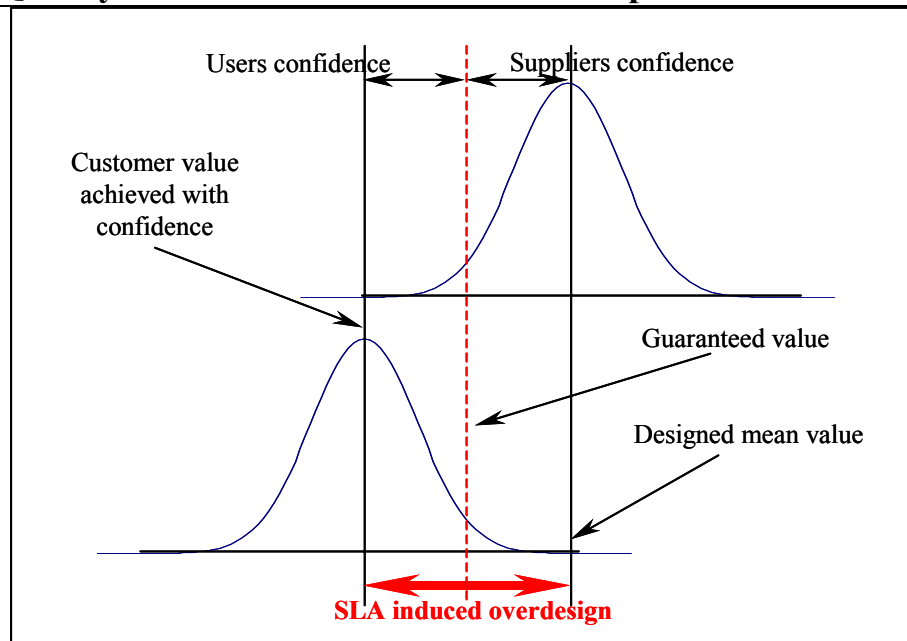
In the absence of other imperatives, we would normally recommend setting the risk level at 1%. The acceptable risk will obviously be affected by the commercial implications of failure and the compensation regime that has been agreed as a part of the SLA.

What we have just described is the process whereby the supplier of a service sets the test thresholds of the SLAs so that they are close enough to the Mean level to provide an incentive to improve if the design turns out not to have the performance that was expected, but not so close that there is a steady stream of failures of the test even when the system is performing correctly. The gap between the Mean and the test or Guarantee level is known as the supplier's confidence level. If, for example, a system is designed to exhibit an MTBF of 10,000 hours, then the supplier might set the test threshold at, say, 8,000 hours in order to provide the supplier's confidence.

Now a user, seeing the test level of 8,000 hours will, quite correctly, say that a test that demonstrates only 8,000 hours gives him no confidence that the 10,000 hour confidence level is being met. He may apply his own confidence level to the test (the user's confidence) and say that it only demonstrates that a figure of, say, 6,000 hours is being achieved. Both of these positions is perfectly justified and supported by the mathematics. They are illustrated in Figure 11.

# CONSULTATION

## Quality of Service measurement and improvement



*Figure 11 Impact of user and supplier confidence on design*

In a rigorous procurement, this effect can lead to something called SLA induced over-design. Suppose that the user wants 6,000 hours MTBF and further wants it demonstrated with a high degree of confidence, he might call for the test threshold to be set at 8,000 hours. The supplier would now look at this requirement and say that if he is required to pass a test at 8,000 hours he had better design for 10,000 hours. In this way the actual design has almost doubled in performance – with a consequential increase in cost – and will outperform any of the users requirements because of the presence of the SLA.

Now the calculations of these test thresholds and the risks that they represent are not a difficult mathematical task, but they are tedious, repetitive and prone to error. In order to assist both users and providers of services IML has developed a small illustrative model of the effect which can be made available to the consulted parties should this be felt necessary and useful.

The model is based on reliability calculations for a notional carrier's carrier operating in Northern Europe. It is written in Microsoft Excel with a great deal of transparency so it could be applied to any other type of service by any engineer with a reasonable grasp of Excel. The model initially takes instructions from the operator as to:

- Which service is to be calculated
- Which countries it connects with
- Which parameter the SLA should be based on
- What the pass/fail threshold is to be
- What the population of services is that the test can be distributed over and
- What the length of the test period should be

It then calculates the Mean performance of the service and, using the population, test threshold and test duration, it calculates the risk of failing or passing the test. This is

# CONSULTATION

## Quality of Service measurement and improvement

plotted for values either side of the actual probability so that the operator can assess whether he is close to a steep part of the curve, and also assess whether a small change in the SLA will result in a major improvement in the risk profile. The model uses the Poisson distribution for discrete events (faults). This distribution closely models the situation where events happen randomly but with a known (fixed) probability, or frequency of occurrence. It is also known as an exponential arrival pattern or an Erlang 1 arrival pattern. All of these names describe the same distribution.

For the continuous events (MTTR for example) the model makes use of the Exponential distribution, which may be considered a little pessimistic, but has the advantage of being easy to manipulate. There are some simplifying assumptions in the model, such as assuming only two parameters can be distributed at the same time, but the results achieved in practice justify them. The purpose of the model is to illustrate the shapes of the curves and to assist non-mathematical staff to assess the commercial implications of different SLA strategies.

A sample output page from the model is included at Figure 12. It illustrates an MTBF test of 10 services that individually have an MTBF of 2401 hours with a fail level set at 2000hrs and a sample period of one year. The result shows that the probability of failing the test is 0.124181(12.42%) with the chosen values of MTBF and MTTR. The model can be instructed to make similar calculations for MTTR and Availability and all of the variables surrounding the test can be adjusted.

# CONSULTATION

## Quality of Service measurement and improvement

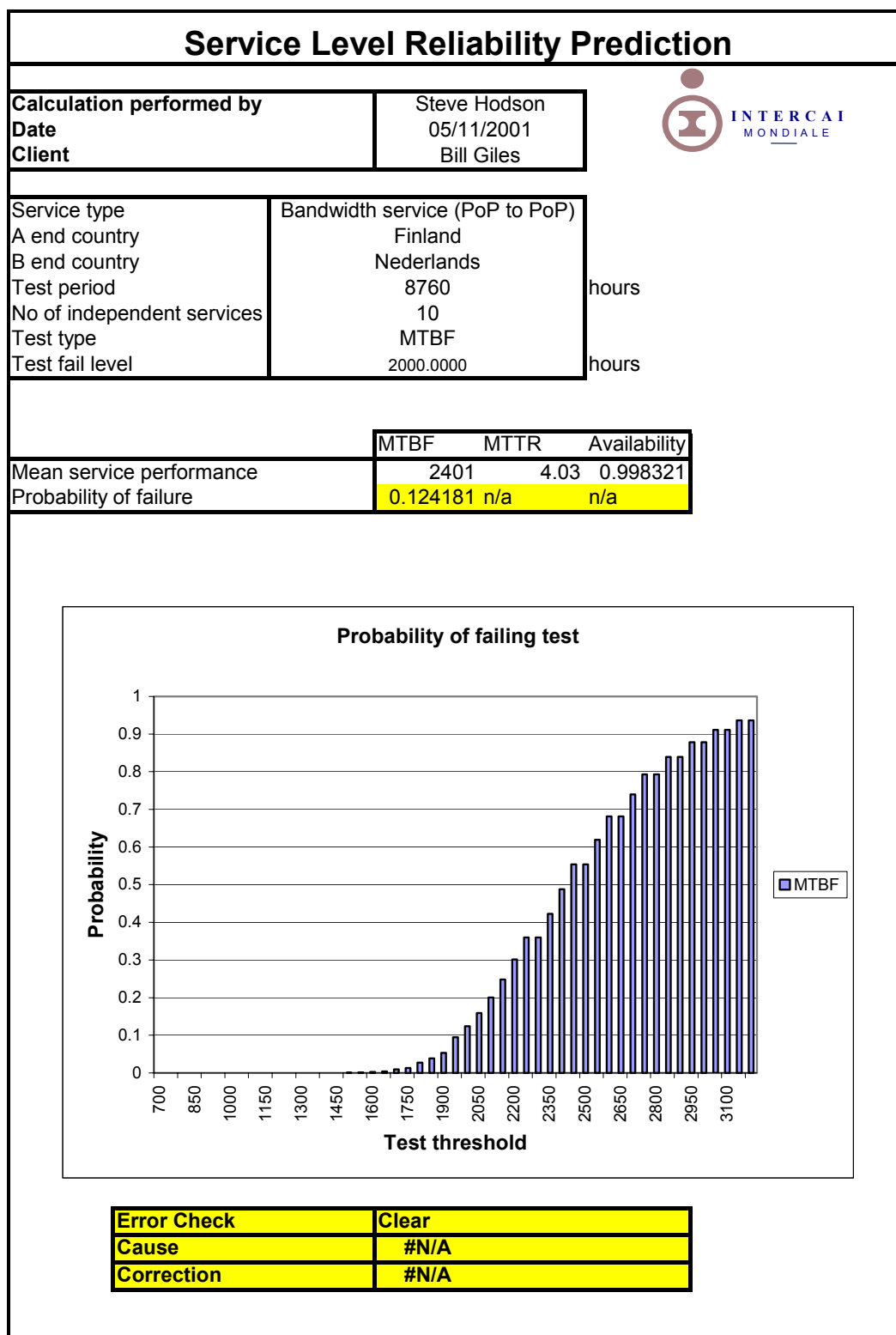


Figure 12 Sample model output sheet

# CONSULTATION

## Quality of Service measurement and improvement

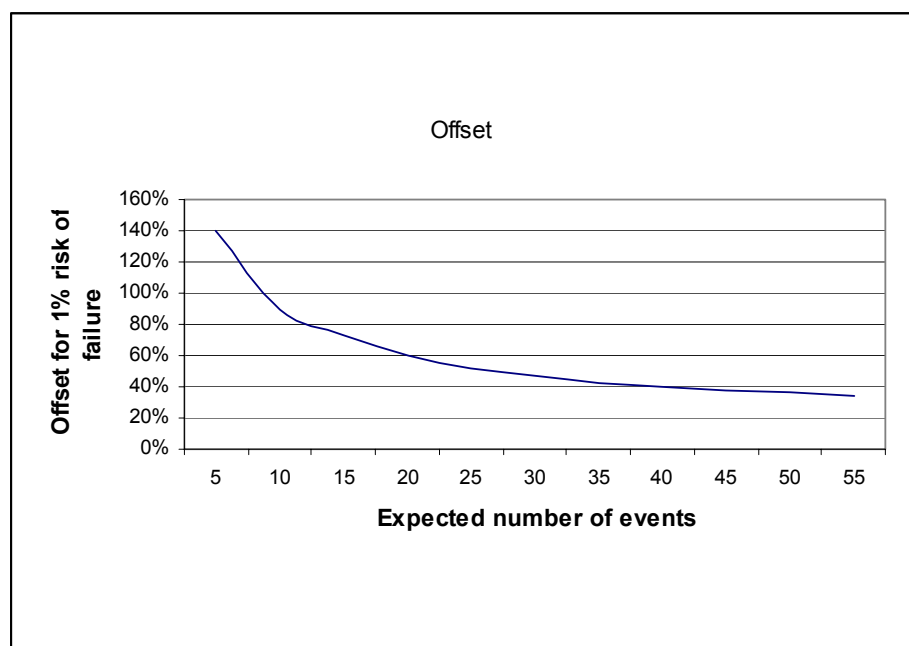
### A2.6 Impact of sampling on guarantees

We have discussed the issue of sample size above and the need to ensure that the samples are sufficiently large to allow the CLT to operate effectively. There is also a need for the sample size to remain large in order to minimise the effect of the widening SD of PDFs with small sample sizes.

We have also discussed the need for an offset between the guaranteed level of performance and the Mean performance in order to provide the supplier's (or user's) confidence. In this section we provide some rules of thumb and a chart to simplify the process of threshold setting and to illustrate the effect of failing to keep a sufficient sample size. The process of taking enough measurements to provide meaningful results is known as achieving "statistical significance".

Taking a discrete variable and the Poisson distribution, Figure 13 shows a chart of the guarantee offset needed from the Mean value (in percentage terms) for different sample sizes and for a constant supplier's confidence of 1%. Translating this into practical terms, suppose that a test for MTBF is being constructed with a known test period and a known population. From knowledge of the MTBF it is possible to calculate the number of faults that are expected in the test period.

What Figure 13 shows is the percentage offset that is needed to achieve a constant risk of failure of 1% for different numbers of expected faults. If, for example, the test period and population implied an expected number of 5 faults, then the test level would need to be offset by approximately 140% (i.e. 7 more faults) to achieve a 1% risk of failure.



*Figure 13 Constant risk offset for varying sample size*

If, on the other hand, the number of faults expected was 50, then the offset would only need to be 40%. These values apply to the same service delivering the same performance. The variation in test threshold comes about solely as a result of statistical significance. This has a direct relevance to the values that will be established for service measurement in the remainder of this consultation.

# CONSULTATION

## Quality of Service measurement and improvement

This is a very important factor in the setting of test thresholds, and is particularly onerous where the services are highly reliable. Modern telecommunications systems easily achieve MTBFs of 10,000 hours, which is in excess of a year. In order to achieve a test in which, say, 30 failures are expected, then a population of more than 30 service-years would be necessary within the test. This can only be achieved with large populations or a very extended test time, and calls into question systems where Availabilities are offered that are measured over short periods such as a month, or even less.

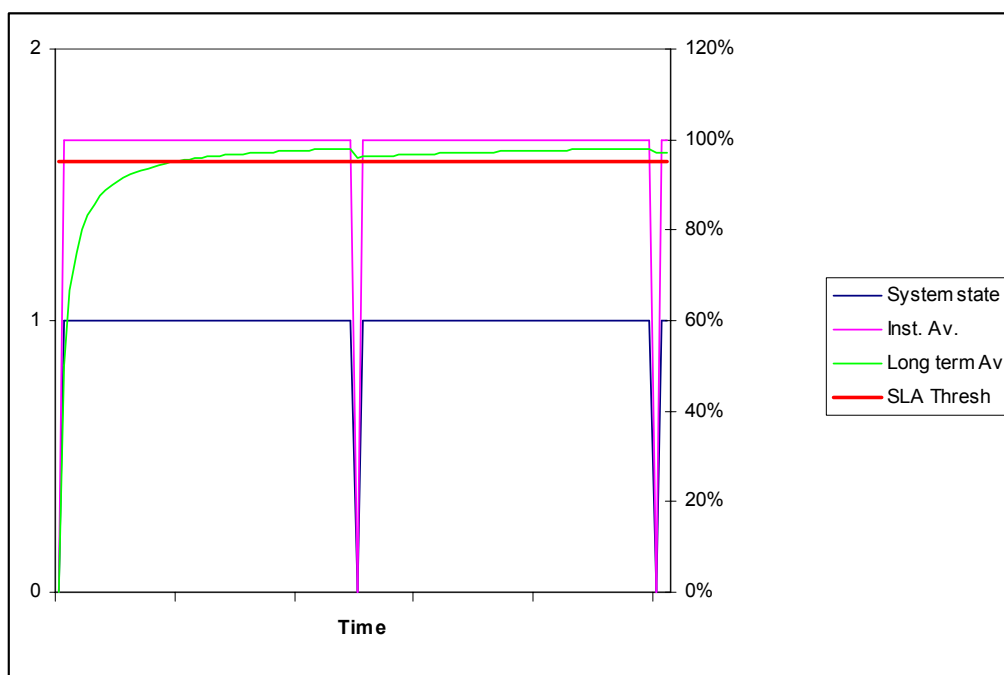
Whilst we have used the Availability example as an illustration of this effect, it has general application across the field of SLAs.

### A2.7 Steady state and real time measurements

There is a tendency with modern systems to attempt to report everything, including the SLA parameters in real time with little thought for the consequences. We have no issue with designing management processes that keep a user informed as to the current status of the service being provided; indeed these are an essential part of service provision. However we do believe that reporting against SLA thresholds too rapidly can be misleading, can lead to information overload and also can encourage faulty decision-making based on incorrect data. We use two real-life examples to illustrate this process. One is based on instantaneous reporting of Availability and the other on instantaneous reporting of traffic overload (i.e. excessive delay).

#### A2.7.1 Real time reporting of Availability

Availability is a parameter that defines the probability of a system being working at any given instant. However, that is not the same as reporting whether a system is Available at any given instant. Consider a system whose performance is described in Figure 14.



*Figure 14 Meaning of Availability*

# CONSULTATION

## Quality of Service measurement and improvement

This system is either Available or not in accordance with the black 'System State' series on the chart. It is Available when the system state is one and Unavailable when the system state is zero (against the left hand vertical axis).

Now if Availability were to be reported instantaneously, then the value would be the same as the system state parameter and would be represented by the mauve series 'Inst. Av.' (against the right hand vertical axis). In practice this would be identical with reporting that a fault had occurred, and the instantaneous value of availability would either be one or zero. Whilst there may be value in reporting that a fault has occurred, there is little value in reporting Availability in this way against a threshold, because the actual Availability cannot be calculated at least until another fault has occurred and, in practice, some significant time after that. Now the steady state Availability in this particular example is 98%, but it can be seen from the green series 'Long term Av.' (this is Availability measured from the beginning of the chart), that this level is not reached until some significant time after the fault has been cleared because of the need for a denominator in the Availability formula.

If now we compare the long term availability with a notional SLA threshold of, say 95% (see the red series 'SLA threshold'), then it can be seen that in the long term the Availability performance is perfectly satisfactory, but the instantaneous Availability, and any short term reporting of Availability would be indicating highly unsatisfactory performance.

It is essential in our view that any reporting of SLA parameters should take sufficient time to ensure that reliable data is available before decisions, that may prove expensive, are taken. The example given above shows that any decisions taken on the basis of instantaneous Availability reporting would be likely to be flawed. In practice we recommend that parameters such as this should be measured over a statistically significant period as described in section A2.5 above.

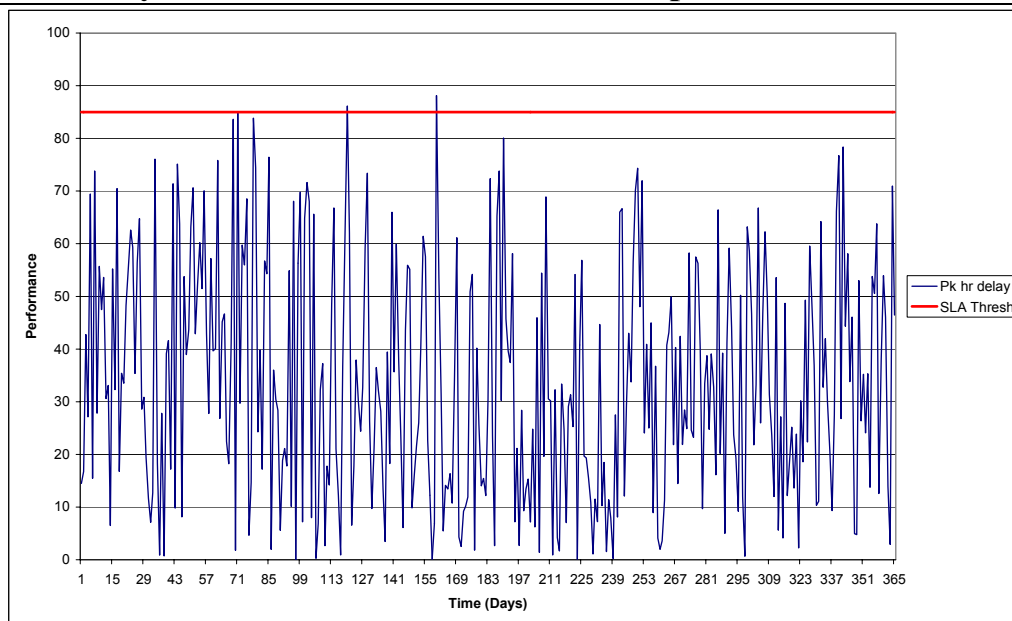
### A2.7.2 Real time reporting of performance data

A similar situation to that described in section A2.7 above arises with Performance SLAs. Again we recommend an approach that allows the SLA parameters to be viewed in a relatively long-term manner rather than instantaneously. As before we have no issue with the need to report issues as they arise – that is normal good service management practice, but decisions affecting performance and cost do need to be based on solid foundations and this can only be achieved in our opinion by taking a longer term view.

As an example of this effect, consider a system with a performance parameter related to delay. As delay is likely to vary across each day, the measured parameter is Peak Hour Delay, and this is compared with a performance threshold in Figure 15. This particular parameter is affected by the submitted load, and the comparison against the SLA threshold is shown for a whole year. It exceeds the threshold on two occasions – something less than 1% of the peak hours.

# CONSULTATION

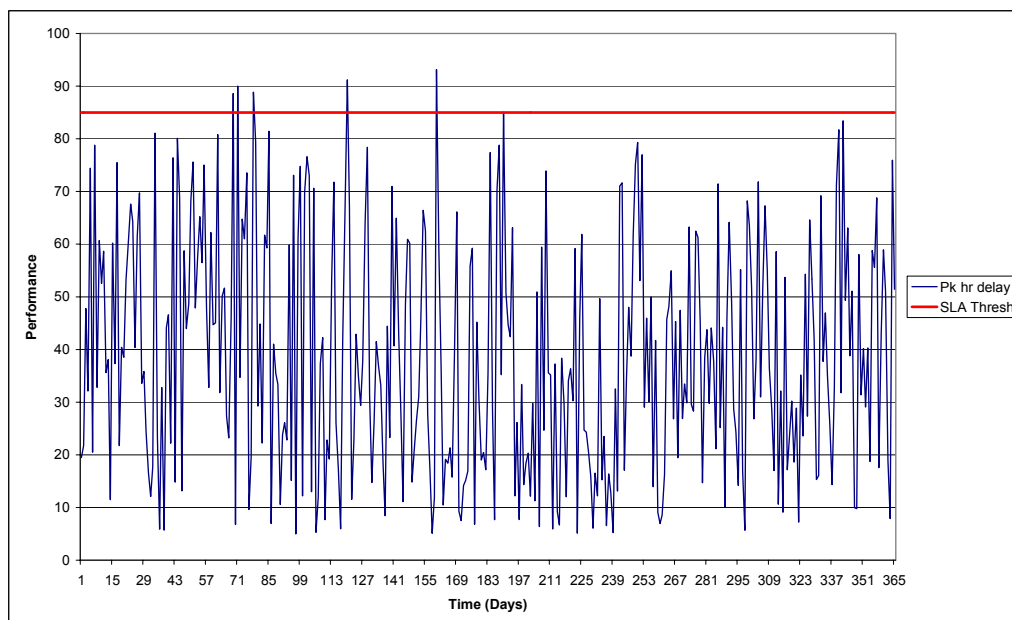
## Quality of Service measurement and improvement



*Figure 15 Performance based SLAs*

Now most users (except those with extremely stringent requirements) would probably accept that this was a broadly satisfactory service. They would expect to review the causes of the occasional breaches of the threshold with the Service Provider, and provided the explanations were satisfactory, then apart from the payment of some minimal service credit no further action would be taken.

Consider now the case where the traffic has risen by 5%, as shown in Figure 16.

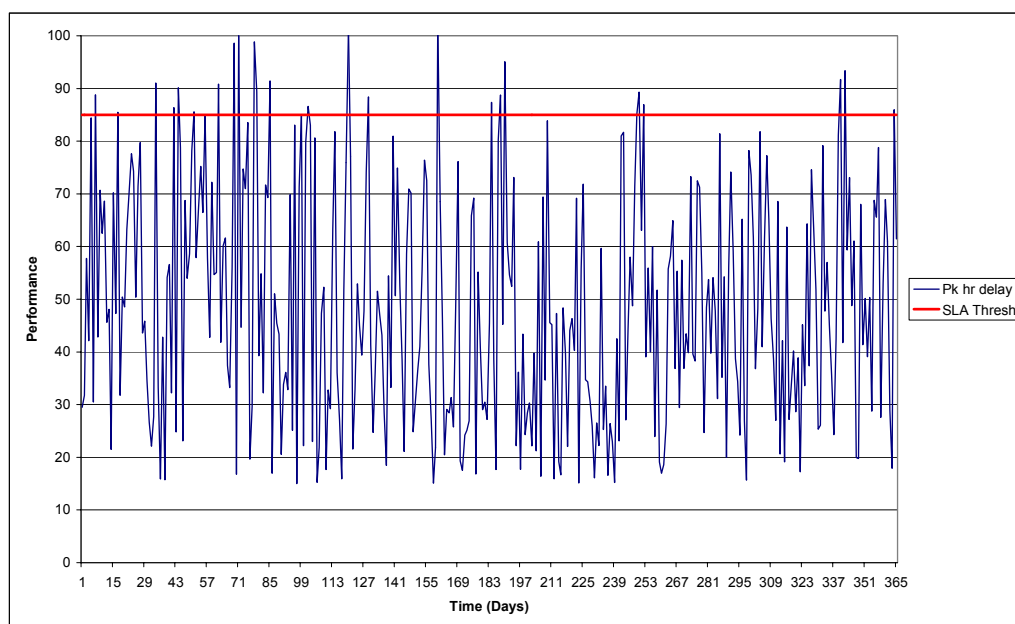


*Figure 16 Performance based SLA with 5% more traffic*

The rise in traffic has caused the number of breaches of the SLA threshold to increase, but the number of occasions is still quite small.

# CONSULTATION

## Quality of Service measurement and improvement



**Figure 17 Performance based SLAs with 15% more traffic**

If we now extrapolate the increase in traffic further, as shown in Figure 17, a level of traffic is eventually reached where the number of breaches of the SLA is such that consideration needs to be given to changing some of the underlying parameters of the service. This may mean the addition of further capacity to the service, an alteration to the usage pattern of the user or some other resolution of the issue. The point we are making with this and the preceding section is that SLAs involve making decisions that may well involve significant cost and the context of SLA breaches can have a significant influence on their meaning. We recommend, therefore, that while there is a need for reporting to take place in real time, the SLA parameters and the decisions that rest on them should be taken in slower time and should allow sufficient time for the full context to emerge before actions are taken on the measures.

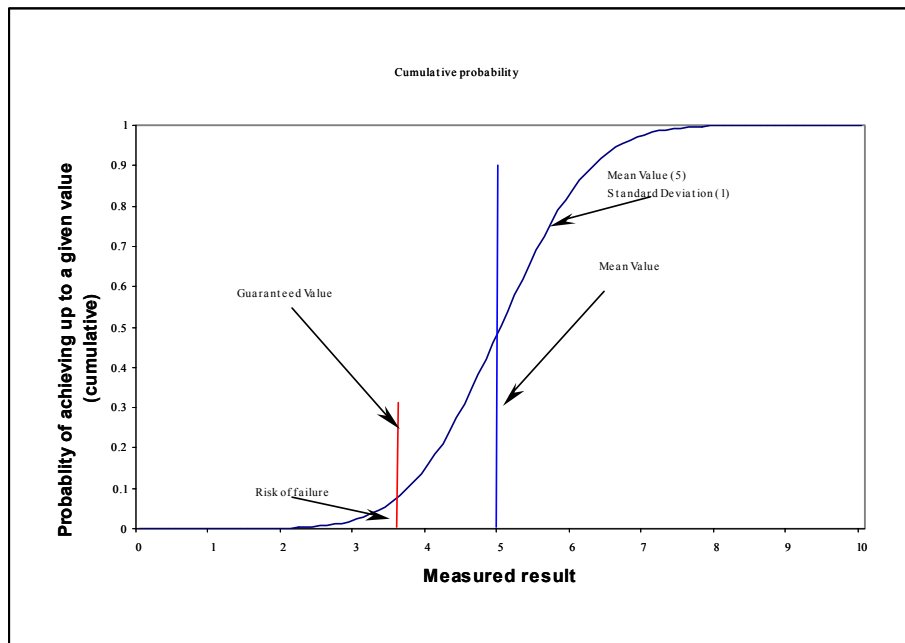
### A2.8 Handling multiple sourced systems

It is frequently the case that a service is delivered by several suppliers operating in series. Whilst it is preferable in such a situation for one of them to take responsibility for the end-to-end service performance, it is important to know how guarantees from several suppliers can be combined mathematically. The issue, and the techniques that can be deployed to resolve it, are equally applicable to the slower moving parameters such as Availability, as to the faster parameters such as delay or latency. The general principle at work here is that no service provider can 'guarantee' to meet a mean value of specified performance (whatever the parameter) because the very nature of sampling means that 50% of the tests applied will fail. Any SLA parameter, therefore, will include an element of risk allowance (whether conscious or unconscious on the part of the provider) to ensure that if he is meeting the requirement as stated in the specification, then the test levels he uses do not force him into breach of the SLAs too often. In this section we examine the general mathematics that describes this situation, using standard parameters as an example.

# CONSULTATION

## Quality of Service measurement and improvement

If each of the constituent elements of a service has a common parameter (using, say, MTBF as an example in this case) with a distribution that can be expressed in a similar manner to that shown in Figure 18, then the guaranteed level of performance will be a known percentile of the distribution. For purposes of description, let us assume that the guarantee is at the 99th percentile of the distribution.



**Figure 18 Individual service guarantees**

Now if there are, say, three services being delivered in series as building blocks for an end-to-end service (possibly using several different suppliers), then the probability that any one of the measurements taken is at the 99th percentile point is 0.01 or 1%. However, the probability that two will be at the 99th percentile at the same time is  $0.01 \times 0.01$  or 0.0001, i.e. 0.01%. Similarly the probability that all three will be at this extreme at the same time is 0.000001 or 0.0001%.

What this means in practice is that combining the guarantee (percentile) values of serial constructions such as this is not possible by the normal combinatorial methods that would be suggested by, say, reliability engineering.

If the characteristic, or mean, values for each element in the service are known, then it is perfectly possible to calculate the characteristic performance of the whole service. If the parameter being measured, for example, is MTBF, then summing the inverse values of the MTBF (i.e. the Failure Rates) for each element and re-inverting the answer will give the MTBF of the system as a whole, thus:

$$\frac{1}{MTBF_s} = \frac{1}{MTBF_a} + \frac{1}{MTBF_b} + \dots + \frac{1}{MTBF_n}$$

where  $a$  to  $n$  are the individual elements and  $s$  is the overall system.

Similarly, for delay across a multi element system, the overall mean delay is the sum of the individual mean delays.

This principle applies no matter what parameter is being examined and it is the basic foundation on which end-to-end SLAs can be calculated, whether the underlying

# CONSULTATION

## Quality of Service measurement and improvement

component services are delivered by a single or by multiple suppliers. The process steps can be described as follows:

- Identify individual service SLAs
- Work backward from the SLAs to the underlying mean performance (by understanding the architecture and the risk calculations)
- Combine the mean performance parameters to produce an end-to-end mean performance figure comprising potentially dissimilar services from multiple sources.
- Re-calculate the end to end SLA that can be offered using the same technique in the forward direction

If, now, the user wishes to combine the known data to provide the probability of passing some end-to-end performance threshold, some further information is required. This is some measure of the variability of the distribution of the parameter. Whilst this is often expressed as the Standard Deviation, the more useful parameter is the Variance, which is the square of the Standard Deviation. It is possible to calculate this value from the knowledge of the 99th percentile and the assumption of a given distribution. It is legitimate to add Variance values across a multi element system and this summation provides the Variance of the overall service, which can be converted back to a Standard deviation by taking the square root. Reference 1 shows a method whereby the ratio of the Mean and Standard Deviation can be used in combination with the incomplete Gamma function to calculate any given percentile for the performance of the overall system. Suffice it to say that it provides a means whereby either the user himself, or the service provider taking overall responsibility for the service, can calculate the risk of failing any test threshold that is specified in the SLA.

Although it was confined to a single supplier, this approach was adopted successfully for the Government Data Network (GDN) in the UK, where each service was specified in terms both of its characteristic (mean) performance, and its guaranteed performance. Whilst there was no guarantee that the mean value of performance would be met, the Characteristic performance represented the best engineering estimate of what the long-term average performance would be. Failure to achieve it did not mean that sanctions were imposed. This was reserved for the guaranteed performance level.

The advantage of specifying the characteristic performance level is that SLAs from different suppliers can be combined without, at the same time, combining their risks cumulatively, and incorrectly.

## A2.9 Some practical techniques

### A2.9.1 The meaning of Availability

In this section we examine some practical techniques that we, and others, have used over the years in order to ameliorate some of the implications of the statistics that we have discussed above. Some of them are good techniques, one or two we would not recommend. However, before addressing these issues, one further point of a practical nature must be confronted. This is the definition of the term Availability.

Availability is a term defined in many textbooks and in BS 5760. It is a measure of the up time of a system and has a whole infrastructure of tools, techniques and mathematics that support it. There are two forms of expression used in BS 5760 that depend on the

# CONSULTATION

## Quality of Service measurement and improvement

precise definition of MTBF. We prefer to define MTBF as total time divided by the number of faults. This, in our view, defines the average time between faults, or MTBF. The alternative definition is to define MTBF as up time divided by the number of faults. This in our view defines the average Time To Fail (MTTF), not the time between failures.

Both systems are perfectly valid and self consistent, but they are mutually exclusive, and it is important to be clear about which definition is being used at all times. (We have seen one requirement document from a public body that used both definitions in different parts of its specification with consequent confusion). We have used our preferred definition throughout this Attachment and that is the definition given above and gives rise to the relationship:

$$\text{Availability} = \frac{MTBF - MTTR}{MTBF}$$

Note that this definition of Availability refers to a steady state condition and not to instantaneous values. It is our view that instantaneous reporting of parameters such as Availability, while useful as a management tool, should not form a part of the SLA process and this view is argued, together with some examples in section A2.7 above.

Unfortunately, the SLA industry has taken over the term Availability and tends to use it for purposes outside this tightly controlled definition. The issue hangs on the definition of a fault, and the modern interpretation, in our view, is much wider than is allowed for in the mathematical relationship above and leads to confusion. More seriously it leads to faulty analysis of the root causes of problems and seriously blurs the boundaries of responsibility for resolving issues in an outsourced environment.

We do not take this position from a purist mathematician's point of view, but from a pragmatic engineering standpoint from which we believe that the SLA industry is damaging both itself and its clients by using the term incorrectly. We recognise the need for additional types of faults to be analysed in modern systems, but we prefer to carry this out by introducing a new term rather than trying to bend an existing one out of shape to meet our needs and failing in the process. We have seen claims by companies who make use of the new definition to define the functionality of their service or software, but then fall back on the classical definition in claiming very high levels of reliability. This confusion of definitions can be highly misleading to users of services of this nature and damages the credibility of companies who make the claims.

As we mentioned above, the issue hangs on the definition of a fault. The industry today takes the view that a system is faulty if a user cannot use it as he wishes to. We believe that there are some four different, independent mechanisms that can operate that prevent the user from using a system as he wishes, and not all of them are due to what we call faults, or are capable of analysis using normal reliability techniques. The four mechanisms are:

- (a) The system design is incorrect
- (b) The system is faulty
- (c) The system is overloaded
- (d) The system has been operated incorrectly by the supplier or the user

Any or all of these events can prevent a user from using a system as he sees fit. We believe that only the second one causes a system to be Unavailable and susceptible to

# CONSULTATION

## Quality of Service measurement and improvement

proper reliability analysis. That is not to belittle the impact of the system failing to work as the user wishes; it is simply to question the term that is used for it.

For purposes of this Attachment we introduce the term Usable. A system is Usable when none of the four factors listed above are present. The occurrence of any one (or more) of them causes the system to become Unusable. The numerical value of Usability is a dependant variable that can be calculated from the effects of each of the four contributing parameters.

Our reasons for taking this approach are as follows:

- (a) If a system design is faulty, then the time it will take to alter the design is not subject to the rules of random behaviour. Commercial factors come into play and no amount of 'repair' action by an operator can correct the problem. Responsibility for the resolution of the problem clearly rests with the original designers of the system, who may well not be the current operators of the service. In our view every system introduced into service should undergo acceptance testing to show that the design is fit for the purpose to which it is put. This is a fundamental precursor to running an SLA and it is intended to prove that there is a sound basis for the SLA during the life of the system. Trying to correct design faults in service is near impossible, extremely expensive and disruptive to users. Ideally a system should have been tested before release and should enter service with a negligible number of design faults. Should such faults be discovered in live operation, then some form of redesign process needs to be undertaken. This has its own means of management and analysis
- (b) If a system is truly faulty, then some event has happened that causes it to depart from the designer's intent. A component has failed in some way (and software is included as a component) and repair action is required to return the system to the state that its designer intended. Such faults occur randomly in a manner that is capable of analysis using standard reliability engineering tools. The terms Availability, MTBF and MTTR properly apply only to this type of Unusability. The responsibility for fixing the system in line with the SLA as a result of these types of fault rests clearly with the service operator.
- (c) If a system is overloaded, then it cannot be said to be faulty in the sense that it has departed from the designer's intent. There is no doubt that it is Unusable, but the responsibility for that unusability lies squarely with the user, who is using it outside the design parameters, or with the operator who is trying to carry too much traffic across the infrastructure. This may arise as a result of deliberate intent by the user or the operator; as a result of some external event; or simply as a result of gradual growth in the usage of the system. Depending on the cause of the overload, the user may wish to live with the resulting performance degradation if it occurs infrequently. Alternatively he may wish to pay the service operator to increase the capacity available to the application, or he may wish to alter the way in which users make use of the system to avoid the problem. Mathematics exists that can describe the systems performance under any given load condition. It is based on queuing theory, but is completely different from the mathematics of reliability. A system that is overloaded will suffer degradation of its performance parameters – possible delay or latency. This is a Performance SLA, not an Operational one.

# CONSULTATION

## Quality of Service measurement and improvement

- (d) Finally, a system may become unusable as a result of some inadvertent operation of the system by the service provider (or in some cases, the user himself). Again the responsibility for the problem lies with the agency that caused it, and there is no mathematics to describe its performance adequately. Unfortunately, modern systems are so complex that this form of loss of service is coming to dominate some performance statistics. In our view, this type of fault is entirely avoidable. The reason that it occurs is usually that some change is being made to the system, and the consequences of that change have not been fully considered or tested. The fault may arise because the change has not been adequately tested; because the operator has not been adequately trained; or because uncontrolled access to the system has been allowed to some third party. None of these should occur to any significant extent in a well-run service operation.

It can be seen from the above that there are four different mechanisms that can operate to cause Unusability. Cause (a) should only have an effect during the design testing stage. Causes (b) to (d) can occur in normal operation, but they are all completely independent mechanisms that are caused by different agencies and, in our view, cannot adequately be described by a single parameter or SLA. We would strongly recommend that each of these mechanisms should be addressed by a different SLA, and that mechanism (d) be subject to extremely stringent penalties because of its avoidable nature.

One additional reason for preferring this approach is that it is also consistent with the definition of Availability given by the ITU in Recommendation X.137: Availability performance values for public data networks when providing international packet-switched services.

By way of illustrating how the combination of the different mechanisms can cause confusion, consider the case where a service provider has agreed an availability SLA with a user that includes outages due to overload. The user can cause that SLA to fail at any time simply by sending more traffic than has been contracted across the service. The operator of the service has no control of that traffic and cannot, in our view, be held responsible for the consequences.

### A2.9.2 Use of groups to protect users and suppliers

We have described the need for statistical significance in the preceding sections and the pressure for SLAs is to make the population of services as large as possible so as to minimise the length of the test period to achieve statistical significance. This is a valid technique that was used to very good effect in the Government Data network (GDN) in the UK where the whole population of users of a Government department could be a member of a Group for purposes of an SLA.

The disadvantage of this approach is that if a Group becomes very large, while the SLA performance level for any given risk improves significantly, it is also possible for a single user to be permanently off the air without significantly affecting the Group average. A useful technique to protect individual users in this scenario is to use two different SLAs for the same service, one that applies to the Group as a whole, and has a relatively stringent performance threshold, and the other that applies to all individual users in the Group, which has a less stringent threshold in order to accommodate the effects of the smaller sample. Of course, it is important to ensure that the relative importance of these two SLA types is preserved. If SLA monitoring is automated, the

# CONSULTATION

## Quality of Service measurement and improvement

group SLA must be prioritised over the individual SLA in the event that both demand attention at the same time.

### A2.9.3 Use of rolling periods to increase statistical significance

A commonly used technique to increase the population within an SLA measurement period is to make the period quite long (say six months) but to measure the parameters on a more frequent basis (say monthly) using a rolling six-month window. Whilst this does apparently provide more data on which to base the measurements, the approach is flawed, because the data is not independent. Once the measurements for one month have been used, if they are used again in the succeeding month's calculations, then independence has been lost and the statistical significance is not there. This approach also has an unfortunate effect in that if an operator has a major disaster in one month, and that problem is resolved, then it remains in the statistics for six months until it has worked its way out of the rolling window. We do not recommend the use of rolling windows.

### A2.9.4 Taking over existing services

When a service is outsourced, it is frequently the case that records either do not exist, or the new supplier does not have sufficient confidence in the records that do exist to offer a commercially binding SLA with stringent penalties involved. Under these circumstances, it can be useful to invoke a calibration period. This was used to good effect with the London Underground project Connect, where existing systems were taken over and then upgraded. Immediately after the takeover of the systems, a measurement system was put in place and measures taken of the actual performance parameters that were to be used for the SLA. Once statistical significance had been achieved, the operator could put an SLA in place with a high degree of confidence that it could be met. This measurement period was called the calibration period and allowed both the user and the operator to come to an agreement about the SLA on a fair basis using accurate information.

### A2.9.5 Provision of higher performance without major additional cost

Where an operation is critical to a business, it may be thought necessary to provide a very high degree of resilience, or performance to all of its users. This can be expensive, especially if it involves duplication of the communications paths to the end-user's site. In many cases though, there will be a large number of users on the same site and the pressing business need can be met if some reduced number of them is operating even under fault conditions. A technique that was used for the GDN to address this scenario was to add a further layer of SLA performance above the individual user and Group performance thresholds. This was called the Network Access Service (NAS) and its effect was to split the users on a site into two (or more) Groups. If one of the Groups was working then the NAS was regarded as working even if the individual Groups were not. This caused the operator to provide two independent routes into the site, but not to have to provide additional circuits. This was a very cost effective method of achieving the business need at the lowest possible cost.

A similar technique, which does involve some cost, is to provide a service over a leased line, but to provide a dial up option that is held in reserve in case the main link fails. With narrow band networks this could be a perfectly acceptable technique with no performance penalty. With modern broadband systems, the dial up option is likely to be at a lower level of performance and the SLA will need to take this into account. A point

# CONSULTATION

## Quality of Service measurement and improvement

to watch in this configuration is to ensure that the management system is aware when the backup option has been invoked, because dial up circuits typically charge by connect time and the cost of the backup can rise very rapidly if repair action is not initiated immediately.

### A2.9.6 Achieving statistical significance by decomposition of the service

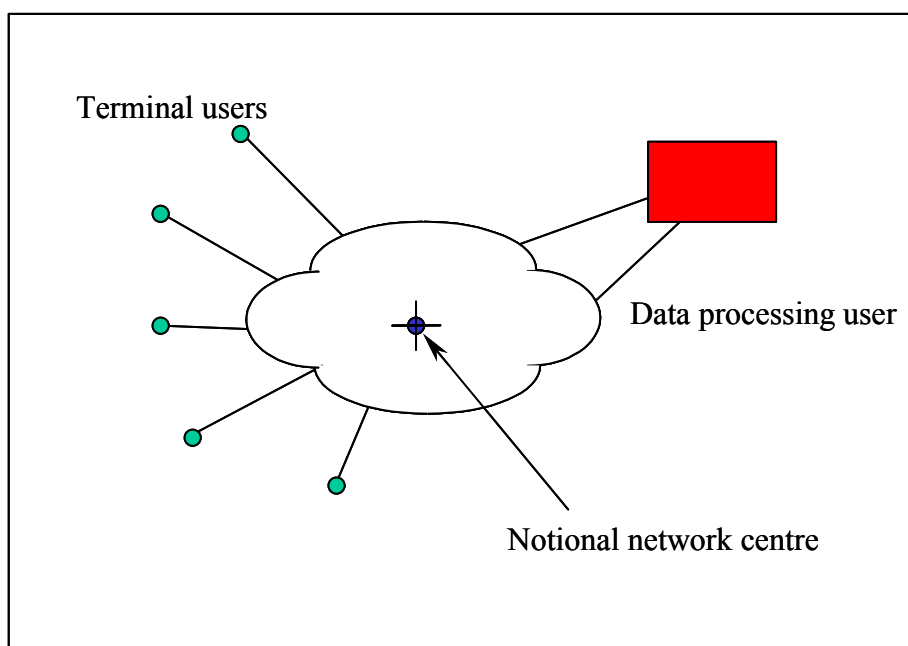
When an extremely high performance service is implemented, it may prove to be impossible to construct a test that adequately provides a measurable SLA for the service. Consider an SDH ring, for example. With its ring configuration providing parallel redundancy, Availability figures above 99.9999% can be achieved. This implies MTBFs of 100,000 hours and more and achieving statistical significance can take an impractically long time.

In these circumstances, it is legitimate to decompose the service into constituent paths – in this case, each of the redundant paths - and measure them separately, relying on mathematical analysis to provide the overall performance figure.

### A2.9.7 Use of end to end Availability

A requirement that frequently arises is for an SLA that manages ‘end to end’ communications across some network infrastructure. There are inherent dangers in this approach that need to be managed.

Consider a system such as that shown in Figure 19:



*Figure 19 End to end Availability definition*

In this scenario, a number of users at terminals are accessing an application that runs at a data centre. The terminal users are connected using single, low speed, circuits, while the data centre is connected using high-speed independent duplicated circuits. Both the terminal users and the data centre could be considered to be users of the network service.

# CONSULTATION

## Quality of Service measurement and improvement

Putting some trivial figures to the configuration for purposes of illustration, assume that the terminals and the data centre are perfectly reliable and that circuits are 99.9% Available. Assume also that the data centre can carry all of the traffic sent to it over just one of the circuits connecting it to the network and that these two circuits are independently routed.

If we measure end-to-end Availability (i.e. terminal to data centre) then the Availability will be almost exactly 99.9% (driven by the terminal tail circuit), and if there are 1000 terminals, then on average the data centre will always be Unavailable to one terminal and could thus be considered to be down all the time. This is an example where an apparently simple construction can lead to confusion. It is also worth noting that the terminal-to-terminal Availability is different from that to the data centre.

A closer look at the configuration shows that the services used by the terminal and data processing user are fundamentally different. The speeds are different but, crucially, so are the reliability figures. The data centre is configured to provide extremely high Availability, while the terminal connection is configured for minimum cost. Lumping these services into a single Group breaks our rule about homogeneity of services and confuses the picture.

Our preferred approach to this common construction is to invent some notional central point in the network that can be considered to be at the highest reliability level within the network. Individual service types such as the terminals or the data centre should then be measured against their Availability to that central point. It will then be seen clearly that the terminal Availability is 99.9% while the data centre is achieving an Availability of 99.991%. Combining the two provides an end-to-end Availability of very close to 99.9%, but the picture is much clearer and is now an aid to diagnosis rather than a hindrance. For this type of configuration, we recommend the use of individual SLAs for the different services, with a combination by mathematics should this be required to achieve the end-to-end need. This provides a much clearer indication of where any problems might lie with achieving the required performance.

### A2.9.8 Time of day criticality

Many applications exhibit behaviour that makes some specific time of day or day of the week more critical than other times. A good example is the UK National Lottery network, which carries 40% of its traffic in the few hours leading up to the draw on a Wednesday or Saturday evening. Although there is much talk about bandwidth on demand, we do not see that this is as yet a practical service since the fundamental infrastructure must be in existence to provide it and if it is there and unused, then it must be paid for. If some other party can share it, then there is a risk that it will not be available when it is needed anyway. These arguments apply marginally to the core of a network, but are irrelevant at the edges where the cost of duplication or backup is probably prohibitive.

The service operator has little in the way of design options to provide higher SLA performance during critical business periods. Failure rates are built into systems, so he will be unlikely to be able to manage them by any management action. Repair times are generally controlled by access to spares and travel times. Placing more staff on call would increase the capacity for repairing faults, but not do much for the MTTR. In any case there should be adequate staff available at all times to deal with faults, so adding further capacity is unlikely to help. Establishing the optimum operating point in cases like this is a matter of achieving a balance between cost and performance.

# CONSULTATION

## Quality of Service measurement and improvement

Our advice is that systems will need to be designed for optimum performance under the worst case conditions and if this means that it is over designed for other times, then that is the price that must be paid for a successful system. This is certainly the case with the lottery network.

### A2.9.9 Definitions and exclusions

Setting and operating an SLA system is a painstaking and accurate activity that involves careful measurements. The accuracy of these measurements can be called into question if the definitions of the parameters that are being measured are unclear. We cannot over-emphasise the need for clear and unambiguous definitions of all the parameters involved.

We present here a number of common difficulties that can arise and which can be avoided by careful definitions:

- What is the definition of a fault? Is a service faulty when it can carry out a part of its function?
- When does a fault start and stop? What happens if the operator sees a fault and the user doesn't? What happens if the user cannot get through to the operator to report the fault? What if the operator cannot inform the user when the fault is cleared?
- Which formula is being used for MTBF and Availability?
- What happens if the Group is not big enough to achieve statistical significance?
- What may be legitimately excluded from the calculations? For example, can the operator declare Scheduled Maintenance Periods (SMPs)?
- Are there specific operating hours for the service? What happens outside these hours?

### A2.10 Summary

In this Attachment we have looked at the mathematics of sampling for the measurement of SLAs. It is clear that this mathematics applies to relatively slow moving parameters, such as Availability, MTTR and MTBF, but not to high volume parameters such as transactions and applications. It is also applicable to less predictable activities such as the implementation of a new service.

We have examined the mathematics of distributed parameters and how test levels should be set for any specific commercial risk level. We have also provided rules of thumb and a model for non-technical staff with a little training to carry out basic assessments of the impact of SLAs on their business.

We have taken a strong line on the meaning of Availability and suggested alternative names for lack of functionality at the service level.

Finally we have provided a number of mathematical tricks and techniques (both good and bad) to enable the authors of SLA specifications to avoid some of the difficulties that can arise with this whole area.

# CONSULTATION

## Quality of Service measurement and improvement

### 4 Schedules to the Consultation

The Schedules to this Consultation will form part of the eventual definition of the Quality of Service measurement requirements. The Schedules provide a common basis for the definition of parameters to be measured and also the methods of measurement that would be acceptable. In this first iteration of the Consultation, values that are quoted are long term average benchmark values based on international experience. They are not, at this stage, the values that would act as thresholds against quarterly returns, which will be derived from the long term average benchmark values and the mathematics of section A 2 above.

Once the initial returns are received, the relationship between what has been achieved and the long term average benchmarks will be used to derive a new long term average target for Bahrain (see section 3.10 above). Where necessary, a timetable of improvement may be worked out to enable efforts to be directed towards improvement against a practical timetable.

As new services are introduced to the market, further Schedules will be included in this document and its successors that address those services. The basic approach will remain static regardless of the service types. Values and thresholds may be service specific, but the Authority intends that the parameters to be measured and the definitions of those parameters will remain constant, regardless of the service type. If a new service is introduced that does not fit within these parameters, then the Authority will consult on the means by which it will be measured before introducing it to this document.

The first two Schedules contain common definitions and standards that apply to all measures. The remaining Schedules are service specific and contain performance thresholds that apply to each service type.

For modern services there is little in the way of standardisation that is available to apply to these measurements. However, the ITU-T has carried out significant amounts of work on this subject for packet based systems that made use of earlier generations of technology. Much of the methodology of this work is applicable to modern systems even if the terminology and actual performance values are not. Where additional information is needed, the Authority may make reference to those aspects of the following standards that are of general applicability:

- ITU-T Recommendation X.131, Call blocking in public data networks when providing international synchronous circuit switched services
- ITU-T Recommendation X.134, Portion boundaries and packet-layer reference events: Basis for defining packet-switched performance parameters.
- ITU-T Recommendation X.135, Speed of service (delay and throughput) performance values for public data networks when providing international packet switched services.
- ITU-T Recommendation X.136, Accuracy and dependability performance values for public data networks when providing international packet-switched services
- ITU-T Recommendation X.137, Availability performance values for public data networks when providing international packet switched services.

# CONSULTATION

## Quality of Service measurement and improvement

- ITU-T Recommendation X.138, Measurement of performance values for public data networks when providing international packet switched services.
- ITU-T Recommendation X.140, General quality of service parameters for communication via public data networks.

### Schedule 1 Measurement methods

This Schedule describes the methods for deriving the measures that the Authority would find acceptable. The Authority expects to carry out audits from time to time on the collection of measurements and will expect the underlying data to be retained by the service providers in the event that an audit is to be undertaken. Measurement methods will be consistent across all service types.

#### Schedule 1.1 Implementation

Implementation measures are used to track the performance of a service provider in responsiveness to customer requirements. The method used for measuring these parameters should be based on the management systems used by the service provider. These systems must be capable, at the minimum, of identifying and recording the parameters defined in Schedule 4.1 below. Internal reports and documentation should be retained by the service provider in support of any possible audit requirement.

Implementation measures will be averaged over the whole of a reporting period.

#### Schedule 1.2 Operation

Operational measurements are based on the periods during which a service is available to the user. It must be possible to identify the point at which service starts and stops in terms of order fulfilment, and be able to track faults from their initial recognition to their eventual resolution. Periods such as scheduled maintenance, and service cover periods should be recognised by these systems.

These measures should use the order processing, maintenance scheduling and trouble ticketing systems used by the service provider and should retain sufficient detail to enable subsequent auditing to take place for the parameters specified in Schedule 4.2 below.

Operational measures will be averaged over the whole of the Service Cover Period during the reporting period.

#### Schedule 1.3 Performance

Performance measures may be derived in one of two ways, by direct measurement of specific packets (whether inserted for the purpose or not), or by mathematical derivation from more easily measured parameters such as occupancy and packet size Variance. The parameters to be measured are defined in Schedule 4.3 below.

##### *Specific packet measurement*

If necessary, specific traffic may be instrumented, or test traffic may be injected into the information stream. These packets must be representative of the traffic being carried and must be statistically significant. Information about the performance of the test traffic will be collected by normal measurement techniques such as SNMP.

Mathematical methods may then be used to extract the specific performance measures

# CONSULTATION

## Quality of Service measurement and improvement

required by this document. The mathematics to be used in this instance must be recorded and made available to the Authority for audit purposes.

### *Proxy performance measurement*

As an alternative to the specific packet measurement method, a service operator may elect to use other, more easily measured, parameters as a proxy for the direct measurement of performance. Analytic mathematical measurements of the type described in section A 2 above would then be used to derive the specific performance measures that are required. Typically, parameters such as occupancy of nodes and links and packet variance statistics can be used as primary inputs to an analytic model, and the model used to provide the measures needed by this document. Where this approach is adopted, the Authority will expect to have sight of the mathematics and also expect to see a calibration test that shows that the measurement method is an acceptable proxy for the measurements.

All Performance measures will be averaged over the peak hour of each day and reported individually at the end of each reporting period. Long term average benchmarks apply to the worst case day in each period. For parameters that apply to many to many end point services, the output value will be traffic weighted in accordance with the mathematics of Attachment A 2 above.

### **Schedule 1.4 Other**

The 'Other' category of measurements addresses what might be considered to be the softer and more subjective issues associated with service delivery. The Authority recognises the lack of technical rigour associated with such measures and wishes to minimise their use as far as possible. However, it does accept that if there is an underlying discontent on the part of users with the services offered by a service provider, then this may be an indicator of more serious difficulties with the service(s) being provided. Three categories of measure are proposed in this area and these are defined in Schedule 4.4 below.

#### **Schedule 1.4.1 Complaints**

All service providers are required to maintain a log of complaints, categorised by service type so that they can be related to the performance measures against those same services. This log will be used as the basis for the analysis of complaints and the detailed entries in the log must be maintained in the event an audit is required.

#### **Schedule 1.4.2 Billing**

The Authority is keen to ensure that service providers operate accurate and fair billing systems so that users are charged the amount they expected to pay. To this end, the Authority will wish to see regular tests of the accuracy of billing systems by subjecting them to known test data and checking their accuracy. These tests can be carried out off line. In addition, as a special category of the Complaints register mentioned in Schedule 1.4.1 above, a record will be maintained of all complaints related to billing, and reports will be made against these complaints in the regular returns.

#### **Schedule 1.4.3 Customer satisfaction**

The third category of 'soft' measurement is that of Customer Satisfaction. The Authority will expect to see operators test their performance in a subjective manner on

# CONSULTATION

## Quality of Service measurement and improvement

a regular basis by consulting their users and following up on user complaints. The Authority wishes to be consulted on the methods proposed for testing Customer Satisfaction in the interests of promoting a common approach. These activities will be recorded and the results submitted to the Authority, together with the methods used.

### Schedule 2 Measurement definitions

This Schedule contains formal definitions of the main parameters that will be used in deriving the measures listed in the remaining Schedules. In every instance of their use, the Authority expects the same definition to be used across all services and by all service providers. Some of the definitions given in these Schedules are for subsidiary parameters (such as a Fault) which are not for direct measurement, but whose definitions affect parameters that are to be measured.

Where, for convenience, a service provider decides to group a set of services together for reporting purposes, that group will be used for all aspects of measurement. It is not acceptable for different groupings to be used for different measures.

#### Schedule 2.1 Subsidiary parameters

Parameters defined in this sub-Schedule are used to derive the primary measures elsewhere in the Schedule. All measures of time should be made to the nearest minute (or smaller) in this sub-Schedule.

##### Schedule 2.1.1 Fault

The presence or absence of a Fault is used to define whether a service is Available or not. A service is regarded as faulty when it enters a state where it does not meet the service specification and some repair action is required. For the avoidance of doubt, a service is not considered to be faulty when it is overloaded, that is a separate condition addressed under Performance.

A Fault is considered to start when the service provider first becomes aware of its existence, whether or not the user is aware of the fault. It is considered to come to an end when the user has agreed that the fault has been cleared. Start and finish times will be recorded to the nearest minute.

##### Schedule 2.1.2 Service Cover Period

The Service Cover Period (SCP) is the time during which it is agreed that the service will be operated to the defined performance standards. This may be 24 hours per day, or such other period as is defined in the service specification.

##### Schedule 2.1.3 Scheduled Maintenance Period

A Scheduled Maintenance Period (SMP) is any pre-publicised period during which the service provider needs to carry out maintenance work on the service infrastructure and during which it is possible that the service will not operate to the intended standard.

##### Schedule 2.1.4 Lost Access Time

Lost Access Time (LAT) is time in which the service provider is prevented from clearing a fault as a result of user action or inaction. For example LAT includes time when access is denied to customer premises in order to clear a fault. It may also include

# CONSULTATION

## Quality of Service measurement and improvement

time during which the service provider is unable to contact the user to inform him that a fault has been cleared.

### Schedule 2.1.5 Downtime

Downtime is used in the calculation of Availability, MTBF and MTTR. It is measured on a service by service basis over a reporting period. For each service, Downtime is the sum of all the time during the reporting period when a Fault exists on the service. Downtime may exclude time outside the SCP, time within SMP and any LAT.

Downtime for a group of similar or identical services is the sum of all the Downtime for all the services in the group.

Downtime will be measured in hours to the nearest minute.

### Schedule 2.1.6 Operating Time

Operating Time (OT) is the sum of the total operating time of all the services in a group during the reporting period. It may exclude periods outside SCP, and periods within SMP.

Operating Time will be measured in hours to the nearest minute.

### Schedule 2.1.7 Peak Hour

All performance measures will vary depending on the traffic load submitted to the network resources. Most networks experience daily, weekly and other periodic variations. To avoid substantial variation over short periods, all Performance based measures will be measured on an hour by hour basis and averaged over each hour. The Peak Hour during a day is that hour which experiences the worst value in the parameter being measured. The Peak Hour during a reporting period is the hour within the period that experiences the worst value for the parameter.

### Schedule 2.2 Initial Response Time

The Initial Response Time (IRT) is an Implementation parameter. It measures the time between the service provider becoming aware that a requirement for service exists and the time at which a delivery promise is made to the user.

IRT starts when:

- A postal request for service is opened, or
- An Internet/SMS/e-mail or other on-line request for service is received, or
- A telephone call requesting service is completed

IRT ends when:

- A postal response with a firm delivery date is put into the post, or
- An Internet/SMS/e-mail or other on-line response with a firm delivery date is sent to the user, or
- A telephone call providing a firm delivery date is completed.

IRT will be measured in hours to the nearest 5 minutes.

# CONSULTATION

## Quality of Service measurement and improvement

### Schedule 2.3 Delivery Performance

Delivery Performance (DP) is a measure of the performance of deliveries against the promised delivery dates. It will be measured for each group of services of a similar nature, and will be reported on each reporting period. A delivery is considered to be complete when the user has accepted it as satisfactory.

For the reporting period, the DP will have a range of values as follows:

- The percentage of services that were delivered on time or early compared with the delivery promise during the reporting period.
- The percentage of services that was delivered up to one day late compared with the promise.
- The percentage of services that was delivered up to two days late compared with the promise
- And so on until all deliveries are accounted for

### Schedule 2.4 Delivery Time

Delivery Time (DT) is a measure of the absolute time between the end of IRT and the completion of delivery as defined in Schedule 2.3 above. It will be recorded for each individual service and reported on in groups with the average delivery, and the 95%iles around the average in each direction (early and late).

Delivery Time will be recorded in days (working business) and hours to the nearest 15 minutes.

### Schedule 2.5 Complaints

The Authority does not wish to be proscriptive about the means of measuring Complaints. It does expect to be able to inspect the register of Complaints and on the basis of this inspection may consider introducing measures at some stage in the future. It will consult on this prior to taking any action with regard to complaints.

The minimum data that must be maintained in the register includes:

- Contact details of the complainer
- Type of complaint
- Service affected
- Start and end times of the Complaint

### Schedule 2.6 Billing

The Authority regards it as essential that bills provided to users of services be both timely and accurate. In Schedule 1.4.2 above the methods of checking billing accuracy are defined as accuracy and complaints.

The parameter to be measured for Billing accuracy will be the percentage error introduced by the billing process. This will be the percentage difference between the 'correct' total billed and the 'actual' total billed.

# CONSULTATION

## Quality of Service measurement and improvement

The other parameter measured in respect of billing will be the number of complaints relating to billing. The parameters to be measured for this measure will be the same as those defined in Schedule 2.5 above.

### Schedule 2.7 Customer satisfaction

The Authority will expect service providers to check their customers' perceptions of the service that they provide on a regular (at least annual) basis. Parameters to be included should include performance in relation to:

- Provisioning
- Fault repair
- Support
- Range of products
- Price
- Customer handling

### Schedule 2.8 Mean Time To Repair (MTTR)

For any group of similar services, the MTTR will be measured over the reporting period by taking the total Downtime for all of the services in the group (excluding time outside SCP, time within SMP and LAT if necessary) and dividing it by the number of faults that were recorded.

MTTR will be measured in hours to the nearest minute.

### Schedule 2.9 Availability

For any group of similar services, the Availability will be calculated using the formula:

$$Availability = \frac{OperatingTime - Downtime}{OperatingTime}$$

where Downtime and Operating Time have the meanings defined in Schedule 2.1.5 and Schedule 2.1.6 above.

Availability is a dimensionless number with a value between 0 and 1. For convenience it may be expressed as a percentage to at least two decimal places.

There are two generally accepted definitions of the relationship between Availability, MTBF and MTTR. For the avoidance of doubt, the Authority uses the following relationship:

$$Availability = \frac{MTBF - MTTR}{MTBF}$$

When only two of these parameters are measured for any population of services, the third parameter will be assumed to have the value derived from this formula, which is compatible with the definitions of MTTR (Schedule 2.8 above) and MTBF (Schedule 2.11 below).

# CONSULTATION

## Quality of Service measurement and improvement

### Schedule 2.10 Security (breaches)

Where security is an issue, the parameter to be measured will be the number of breaches of security. This will include physical breaches, where some secure area has been accessed by unauthorised personnel; or it may include electronic breaches where some part of a computer's data has been accessed by unauthorised personnel. Breaches may be categorised by type for ease of understanding.

### Schedule 2.11 Mean Time Between Failures (MTBF)

For any group of services the MTBF will be measured over the reporting period by taking the total Operating Time for the group (excluding time outside SCP and time within SMP if necessary) and dividing it by the number of faults recorded for the group.

MTBF will be measured in hours to the nearest minute.

### Schedule 2.12 Probability of Blocking

For telecommunications services that are set up by switching, there is a probability that network congestion may prevent the connection (logical link, call, circuit etc..) from being set up - see ITU-T X.131 and ITU-T X.140).

The Probability of Blocking is a Performance parameter that may be measured by either of the methods specified in Schedule 1.3 above. It will be averaged over each hour and recorded for the Peak Hour of every day. Over the reporting period the Probability of Blocking is that experienced on the worst day. The Probability of Blocking threshold may be associated with a defined worst case traffic level.

The Probability of Blocking is a dimensionless number between 0 and 1. It may be expressed as a percentage for convenience.

### Schedule 2.13 Probability of Loss of Circuit

For telecommunications services that are set up by switching, there is a probability that network congestion may cause the connection (logical link, call, circuit etc..) to fail prematurely - see ITU-T X.136 and ITU-T X.140).

The Probability of Loss of Circuit is a Performance parameter that may be measured by either of the methods specified in Schedule 1.3 above. It will be averaged over each hour and recorded for the Peak Hour of every day. Over the reporting period the Probability of Loss of Circuit is that experienced on the worst day. The Probability of Loss of Circuit threshold may be associated with a defined worst case traffic level.

The Probability of Loss of Circuit is a dimensionless number between 0 and 1. It may be expressed as a percentage for convenience.

### Schedule 2.14 Mean and 95%ile packet delay

For any telecommunication service carrying packet information, each packet will experience delay which will vary with packet size, occupancy and packet size variance. Delay is a Performance measure that can be measured across a group of meaningful paths by the use of traffic weighting. The parameter that is measured has the start and end points defined in ITU-T X.135 and may be measured by either of the methods specified in Schedule 1.3 above. Both the mean and 95%ile delay will be measured for

# CONSULTATION

## Quality of Service measurement and improvement

a nominal packet size of 128 bytes during the Peak Hour of the reporting period. The Delay thresholds may be associated with a defined worst case traffic level.

The Mean and 95%ile delay figures will be measured in milliseconds (mS).

### Schedule 2.15 Probability of Packet Loss

Some modern protocols (such as IP) will discard packets as a means of flow control. The number of packets discarded will vary with load. For any group of routes, the Probability of Packet Loss is a Performance parameter that defines the percentage of the total packets submitted that may be discarded in the Peak Hour. The Probability of Packet Loss threshold may be associated with a defined worst case traffic level.

The Probability of Packet Loss is a dimensionless number between 0 and 1. It may be expressed as a percentage for convenience.

### Schedule 2.16 Jitter

For a specific data path that requires a steady delivery of data (e.g. a Voice over IP data stream) Jitter may be specified as a measure. Jitter is a measure of the Variance of the inter arrival time of the packet stream. It will be measured in the Peak Hour and may use either of the methods defined in Schedule 1.3 above.

Inter arrival time is the average time measured between the beginning of a stream of packets. Jitter is the measure of the variability of this measure and is expressed as the Standard Deviation of the inter arrival time. The Jitter threshold may be associated with a defined worst case traffic level.

Jitter will be measured in milliseconds.

### Schedule 2.17 Qualifying Population

The Qualifying Population of a service is defined as a threshold above which that population becomes liable to meet the defined performance standards. For purposes of this definition, the Qualifying Population is the number of instances of services with similar type and performance that are in service. Typical categories could be:

- Digital leased lines
- Voice services
- International leased lines

## Schedule 3 Enforcement schemes

The Authority has expressed the view that it does not wish to impose sanctions for non-achievement of Quality of Service thresholds, preferring rather to rely on competition to achieve the same results. Should this prove not to be effective, then the Authority may consider the introduction of financial penalties for non achievement. The principles on which any such penalties will be based are as follows:

- Where it is possible to identify the user affected by the failure to meet performance thresholds, financial penalties will be paid to that user. Failing this they will be paid to the Authority in the form of a fine, or other suitable levy allowed under the Law

# CONSULTATION

## Quality of Service measurement and improvement

- Financial sanctions will be set at levels that provide the service provider with an incentive to improve the service performance without being punitive
- Financial sanctions will be progressive and related to the degree of failure to meet the performance thresholds
- The sanction process will allow for the event that if service fails significantly and consistently to meet its performance thresholds. Under these circumstances consideration may be given to re-setting the service definition, its price or the performance thresholds. This would be considered as a breach of the service provision contract.

### Schedule 4 Performance metrics and targets

The following Schedules provide indicative long term average figures that the Authority considers are general good practice in the provision of international telecommunications services. The list is not exhaustive and will be enhanced as new services are introduced to the market. The approach adopted throughout is to aim for consistency throughout the application of measures to services.

Where performance measures are traffic and capacity dependent, targets will be set on a mathematical basis using the assumptions of:

- Occupancy of 75%
- Packet Variance equal to the mean packet delay

Reports based on this approach will include statements about the actual occupancy and Variance that was achieved.

As real services are introduced to the market, thresholds will be set that take into account local conditions and the population of services that can be measured within the reporting period. The mathematical basis for these new thresholds will be that defined in Attachment A 2 above. These thresholds will represent the aspirations which the Authority will encourage the services to meet. Where competition is deemed to be ineffective, the Authority may declare that the thresholds become formal performance requirements that must be met.

#### Schedule 4.1 Implementation targets

##### Schedule 4.1.1 PoP services

No thresholds are proposed for this type of service.

##### Schedule 4.1.2 Raw bandwidth services

Three scenarios are identified for the provision of these services as follows:

###### *Service provisioned automatically*

Measure	Threshold
Initial response	24 hours
Delivery performance	95%
Delivery time	One day

# CONSULTATION

## Quality of Service measurement and improvement

*Service provisioning involves adding equipment to existing infrastructure*

Measure	Threshold
Initial response	24 hours
Delivery performance	95%
Delivery time	Five days

*Service provisioning involves building new infrastructure*

Measure	Threshold
Initial response	24 hours
Delivery performance	95%
Delivery time	No threshold

### Schedule 4.1.3 Bearer services

Thresholds as for Raw bandwidth services specified in Schedule 4.1.2 above.

### Schedule 4.1.4 Grooming services

Thresholds as for Raw bandwidth services specified in Schedule 4.1.2 above.

### Schedule 4.1.5 Network services

Thresholds as for Raw bandwidth services specified in Schedule 4.1.2 above.

### Schedule 4.1.6 Application services

No thresholds are proposed for this type of service.

## Schedule 4.2 Operation targets

### Schedule 4.2.1 PoP services

No thresholds are proposed for this type of service. Where service agreements are entered into outside the scope of the QoS measurement regulation, the Authority suggests the following thresholds:

Measure	Threshold
MTTR	1 hour (on site maintenance)
	4 hours (remote maintenance)
Availability (Power and Air conditioning)	99.9%

### Schedule 4.2.2 Raw bandwidth services

For this type of service the following Operation thresholds are proposed:

Measure	Threshold
Availability	99.95%
MTTR	4 hours

# CONSULTATION

## Quality of Service measurement and improvement

### Schedule 4.2.3 Bearer services

For this type of service the following Operation thresholds are proposed:

Measure	Threshold
Availability	99.95%
MTTR	4 hours

### Schedule 4.2.4 Grooming services

For this type of service the following Operation thresholds are proposed:

Measure	Threshold
Availability	99.95%
MTTR	4 hours

### Schedule 4.2.5 Network services

For this type of service the following Operation thresholds are proposed:

Measure	Threshold
Availability	99%
MTTR	4 hours

### Schedule 4.2.6 Application services

No thresholds are proposed for this type of service.

## Schedule 4.3 Performance targets

### Schedule 4.3.1 PoP services

No Performance thresholds are proposed for this type of service.

### Schedule 4.3.2 Raw bandwidth services

No Performance thresholds are proposed for this type of service.

### Schedule 4.3.3 Bearer services

For Bearer services that are set up by switching the following Performance thresholds are proposed:

Measure	Threshold
Probability of Blocking	1%
Probability of Loss of Circuit	1%

# CONSULTATION

## Quality of Service measurement and improvement

### Schedule 4.3.4 Grooming services

For Grooming services the following Performance thresholds are proposed:

Measure	Threshold
Packet mean and 95%ile delay <sup>†</sup>	Use method defined in Schedule 4 above
Probability of Packet Loss <sup>†</sup>	Use method defined in Schedule 4 above
Jitter <sup>†</sup>	Use method defined in Schedule 4 above
Probability of Blocking*	1%
Probability of Loss of Circuit*	1%
Actual peak hour occupancy	<75%
Actual packet delay variance	< Mean Delay

<sup>†</sup> Data services only

\* Switched services only

### Schedule 4.3.5 Network services

Performance thresholds for Network services will be the same as for Grooming services specified in Schedule 4.3.4 above.

### Schedule 4.3.6 Application services

No thresholds are proposed for this type of service.

### Schedule 4.4 Other targets

Measures that fall into the Other category will apply at all levels of the model. In general, the Authority does not propose thresholds for these softer parameters (with the exception of Security, which is addressed below) other than to require that records be kept and data provided to the Authority. The areas that have been discussed above include:

- Complaints (see Schedule 2.5 above)
- Billing (see Schedule 2.6 above)
- Customer Satisfaction (see Schedule 2.7 above)
- Security (see Schedule 2.10 above)

For security, no mandatory threshold is proposed, but where such is entered into the following threshold is proposed:

Measure	Threshold
Security	Zero breaches

## Schedule 5 Glossary and definitions

This Schedule contains a glossary of terms and cross references.

**ATM:** Asynchronous Transfer Mode

# CONSULTATION

## Quality of Service measurement and improvement

**Authority:** The Telecommunications Regulatory Authority (see section 1)

**Availability:** Measure of the operational performance of a service (see Attachment A2.9 and Schedule 2.9)

**Bearer:** Standard point to point transmission services (see Attachment A1.1)

**BS 5760:** British Standard. Reliability of systems, equipment and components

**CLT:** Central Limit Theorem (see Attachment A2.5)

**DNS:** Domain Name Service

**Downtime:** Period during which service operation is faulty (see Schedule 2.1.5)

**DP:** Delivery Performance (see section 3.7 and Schedule 2.3)

**DT:** Delivery Time (see section 3.7 and Schedule 2.4)

**DWDM:** Dense Wave Division Multiplexing

**Fault:** Occurrence that causes a service to require repair (see Schedule 2.1.1)

**GDN:** Government Data Network (UK)

**Grooming:** Many to many technology transmission layer (see Attachment A1.1)

**Group:** A set of services of similar performance that are measured together in order to achieve statistical significance.

**Guarantee level:** A measurement threshold which a service provider is prepared to accept against a known risk of failure (see Attachment A2.5)

**Implementation:** A class of measures related to the delivery of services (see section 3.6, Attachment A1.1 and Schedule 1.1)

**IP:** Internet Protocol

**IRT:** Initial Response Time (see section 3.7 and Schedule 2.2)

**ISP:** Internet service Provider

**ITU-T:** International Telecommunications Union - Telecommunications

**Jitter:** A measure of the variability of delay in packet transmissions across a path (see section 3.7 and Schedule 2.16)

**LAT:** Lost Access Time (see Schedule 2.1.4)

**Law:** Legislative decree 48 of 2002 promulgating the telecommunications law for the Kingdom of Bahrain (see section 1)

**Mean level:** The average level across a number of measurements (see Attachment A2.5)

**MTBF:** Mean Time Between Failures (see Schedule 2.11)

**MTTR:** Mean Time To Repair (see Schedule 2.8)

**NAS:** Network Access Service (see Attachment A2.9)

**NAT:** Network Address Translation

**Occupancy:** Measure of the degree to which the capacity of a resource is loaded by traffic.

# CONSULTATION

## Quality of Service measurement and improvement

**Operation:** A class of measures related to the maintenance of services in an operational state (see section 3.6, Attachment A1.1 and Schedule 1.2)

**OT:** Operating Time (see Schedule 2.1.6)

**Other:** A class of unclassified measures (see section 3.6, Attachment A1.1 and Schedule 1.4)

**Packet Delay:** A Performance measure of the time taken for a given packet to transit a telecommunications service (see section 3.7 and Schedule 2.1.4)

**Packet Loss:** A Performance measure of the probability of a packet being discarded (see section 3.7 and Schedule 2.15)

**PDF:** Probability Density Function (see Attachment A2.5)

**PDH:** Plesiochronous Digital Hierarchy transmission system

**Peak Hour:** Time of maximum load and minimum performance for a telecommunications service (see Schedule 2.1.7)

**Performance:** A class of measures related to the capacity of services in a traffic loaded state (see section 3.6, Attachment A1.1 and Schedule 1.3)

**Poisson distribution:** Mathematical means of describing a random arrival pattern (see Attachment A2.5)

**PoP:** Point of Presence

**Probability of Blocking:** Likelihood of failure to make a connection across a telecommunication service due to overload (see section 3.7 and Schedule 2.12)

**Probability of Loss of Circuit:** Likelihood of loss of a connection across a telecommunication service due to overload (see section 3.7 and Schedule 2.13)

**PTO:** Public Telecommunications Operator

**Qualifying Population:** Number of instances of a service above which it qualifies to meet the QoS targets (see Schedule 2.17)

**QoS:** Quality of Service

**SCP:** Service Cover Period (see Schedule 2.1.2)

**SD:** Standard Deviation (see Attachment A2.5)

**SDH:** Synchronous Digital Hierarchy transmission system

**Security:** An Other measure used to monitor the security of a service (see Schedule 2.10)

**Service Provider:** A company licensed to provide telecommunications services in the Kingdom of Bahrain under the Law

**SLA:** Service Level Agreement

**SMP:** Scheduled Maintenance Period (see Schedule 2.1.3)

**SNMP:** Simple Network Management Protocol

**SVC:** Switched Virtual Circuit

**TRA:** Telecommunications Regulatory Authority (see section 1)

# CONSULTATION

## Quality of Service measurement and improvement

**URL:** Uniform Resource Locator

**xDSL:** (Type) Digital Subscriber Loop. A family of high speed transmission technologies for use with copper subscriber loops