International Telecommunication Union

ITU-T Technical Paper

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(1 March 2013)

How to increase QoS/QoE of IP-based platform(s) to regionally agreed standards

Forward

This Technical Paper has been developed by Mr Chaesub Lee.

Contents

Page

1 SCOPE 2

2 DEFINITIONS 2

3 ABBREVIATIONS 4

4 FRAMEWORKS OF NP, QOS AND QOE 5

4.1 NETWORK PERFORMANCE, QOS AND QOE 5
4.2 SLA AND RELATIONSHIP WITH QOS 8
4.3 PARAMETERS FOR QOS 11

5 BUILDING BLOCKS AND MECHANISMS FOR QOS 20

5.1 QOS BUILDING BLOCKS 20
5.2 CONTROL-PLANE MECHANISMS 21
5.3 DATA-PLANE MECHANISMS 22
5.4 MANAGEMENT-PLANE MECHANISMS 24

6 MEASUREMENT AND ASSESSMENT 25

6.1 QOS MEASUREMENT 25
6.2 BASIC NETWORK MODEL FOR MEASUREMENT 28
6.3 QUALITY ASSESSMENT METHODOLOGY 29

7 QOS/QOE IN IP-BASED PLATFORM AND NGN 30

7.1 IP SERVICE PERFORMANCE MODEL 30
7.2 IP PACKET TRANSFER PERFORMANCE PARAMETERS 32
7.3 MEASUREMENT NETWORK MODEL 33
7.4 QUALITY CONTROL IN NGN 37
7.5 MANAGEMENT OF NGN PERFORMANCE MEASUREMENT 43

8 CONSIDERATIONS FOR REGIONAL ENVIRONMENT 47

8.1 CONSIDERATIONS FOR PRACTICAL STATUS 47
8.2 CONSIDERATIONS FOR PERFORMANCE MEASUREMENT AND ITS MANAGEMENT 48

REFERENCES AND BIBLIOGRAPHY 55

List of Tables

Page

TABLE 1 – SUMMARY FEATURES OF NP, QOS AND QOE 7

TABLE 2 – REQUIREMENTS FOR NGN PERFORMANCE MEASUREMENT 44

List of Figures

Page

FIGURE 1 – RELATIONSHIP BETWEEN NP AND QOS 6

FIGURE 2 – THE FOUR VIEWPOINTS OF QOS 6

FIGURE 3 – RELATIONSHIP AMONG NP, QOS AND QOE 7

FIGURE 4 – SLA BETWEEN TWO ENTITIES 8

FIGURE 5 – GENERIC STRUCTURE OF SLA 9

FIGURE 6 – DETERMINATION OF QOS PARAMETERS 10

FIGURE 7 – A SLA IN MULTI-PROVIDER ENVIRONMENTS 11

FIGURE 8 – 3X3 MATRIX APPROACH AND DETERMINATION OF AVAILABILITY STATES 12

FIGURE 9 – INTER-RELATIONSHIP OF QOS PARAMETERS 12

FIGURE 10 – ARCHITECTURAL VIEWS OF QOS BUILDING BLOCKS 21

FIGURE 11 – CUSTOMER PERCEIVED QOS........ 26

FIGURE 12 – BASIC NETWORK MODEL 29

FIGURE 13 – ACCESS NETWORK PORTION 29

FIGURE 14 – LAYERED MODEL FOR IP SERVICE PERFORMANCE........ 30

FIGURE 15 – IP NETWORK COMPONENTS 31

FIGURE 16 – IP NETWORK CONNECTIVITY 32

FIGURE 17 – EDGE-EDGE MODEL 35

FIGURE 18 – SITE-SITE MODEL 36

FIGURE 19 – TE-TE MODEL........ 36

FIGURE 20 – NGN REFERENCE ARCHITECTURE 39

FIGURE 21 – RACF WITHIN NGN ARCHITECTURE 41

FIGURE 22 – FLOW DIAGRAM OF PUSH MODE 42

FIGURE 23 – FLOW DIAGRAM OF PULL MODE........ 43

FIGURE 24 – OVERALL ARCHITECTURE FOR NGN PERFORMANCE MANAGEMENT 45

FIGURE 25 – FUNCTIONAL ARCHITECTURE FOR NGN MPM 46

FIGURE 26 – HETEROGENEOUS NETWORK ENVIRONMENTS 47

FIGURE 27 – QOS/QOE ENVIRONMENT IN THE REGION AND GLOBAL 48

FIGURE 28 – MPM PROCEDURE FOR INTER-DOMAIN 50

FIGURE 29 – AN EXAMPLE FEDERATED INTER-DOMAIN MEASUREMENT MODEL 51

FIGURE 30 – AN EXAMPLE HIERARCHICAL INTER-DOMAIN MEASUREMENT MODEL 52

FIGURE 31 – AN EXAMPLE CASCADING NEIGHBOUR INTER-DOMAIN MEASUREMENT MODEL 53

FIGURE 32 – ACTIVE MEASUREMENT ARCHITECTURE INTER-PROVIDERS........ 54

ITU-T Technical Paper

How to increase QoS/QoE of IP-based platform(s) to regionally agreed standards

Summary

IP platform has been the crucial role for building information society by providing easy way of connecting the world and exchanging information with its simplicity. IP is simple in terms of protocol and operation which normally called as “Best Effort.” This simplicity also contributed to separate services from the underline transmission networks such as features called “everything over IP and IP over anything.”

However this simplicity of IP platform raises some concerns such as quality and security. These are becoming more important issues according to the expansion of using IP platforms. Broadband networks have been seriously developed and deployed to solve the quality related problems. However problems are not yet completely resolved even though Broadband networks have been great impact to enhance Quality of Service (QoS) of IP platform. Various technologies for QoS, especially for end-end QoS and QoE have been developed in addition to the broadband networks. Most of approaches provide partial solution in different ways, becoming heterogeneous network environment in terms of QoS. Because of this, problem space is being widened not only for between providers but also a country level and significantly expanded in case of the regional level. QoS is the forehand condition to meet the requirements of end user’s Quality of Experience (QoE) which is becomes important issue.

ITU-T, as the global standard organization under the United Nation, developed various recommendations for the Network Performance (NP), QoS and QoE. For increasing the QoS/QoE in the region which is a key objective of this Technical Paper, it is emphasized that the most of ITU-T recommendations are useful and valuable, especially to develop regional standards.

This Technical Paper starts with helping common understanding about key factors such as Network Performance (NP), QoS, QoE, Service Level Agreements (SLAs) and their relationships. Information about building blocks to form of QoS and their mechanisms are introduced. And measurement objectives and methodologies for performance and quality including appropriate reference network models are identified and it is noted that these would be candidates for the considerations in the development of regional standards. Then explanation about IP platforms using all those mechanisms, methodologies and models are described including how NGN (Next Generation Networks) support QoS and QoE as one of advanced specific IP based platform. Finally issues need to be considered to the regional developments are summarized and recommended with example scenarios.

Introduction

IP platform, which is used in Internet, has been the crucial role for building information society by providing easy way of connecting the world and exchanging information with “World Wide Web.” Today most of the world is connected together and share the life in on-line world. One of the key features of IP taking this role would be caused by its simplicity. IP is simple in terms of protocol and operation which is called as “Best Effort.” This simplicity also contributed to separate services from the underline transmission networks such as features called “everything over IP and IP over anything.”

However this simplicity of IP platform raises some concerns (becoming serious issues), these are quality and security. Quality is the first concern to use IP platform followed by security. To solve the quality issue, various broadband networks have been seriously developed and deployed.

Broadband networks have been great impact to enhance Quality of Service (QoS) of IP platform, but not complete. Broadband connectivity has been leaded more use of broadband information in any time and any place, especially supported by mobile broadband. In addition, many technologies have been approached in various broadband network technologies to provide QoS solutions. These approaches provide partial solution in different ways, so also contributed for building heterogeneous network environment in terms of QoS. Providers deployed IP platforms adopting various different QoS technologies, some cases, even in the same provider. Consequently a country composed of several different IP platforms using various different QoS mechanisms which caused problems about compatibility and ensuring the end-end QoS. These problems should be expanded when apply to the regional level. QoS is the forehand condition to meet the requirements of end user's Quality of Experience (QoE) which is becomes important issue in this subject.

ITU-T, as the global standard organization under the United Nation, developed various recommendations for the QoS and QoE. For increasing the QoS/QoE in the region which is key objective of this Technical Paper, it is emphasized that the most of ITU-T recommendations are useful and valuable, especially to develop regional standards.

This Technical Paper starts with common understanding about key factors on the subject and their relationships such as Network Performance (NP), QoS, QoE and Service Level Agreements (SLAs). Information about building blocks to form of QoS and their mechanisms are introduced. And measurement objectives and methodologies for performance and quality are identified with appropriate reference network models which would be candidates for the considerations in the development of regional standards. Then explanation about IP platforms using all those mechanisms, methodologies and models are described. Special section has been added to explain how NGN (Next Generation Networks) support QoS and QoE as one of advanced specific IP based platform. Finally issues need to be considered to the regional developments are summarized and recommended with example scenarios.

1 Scope

This Technical Paper introduces key factors to identify issues on quality in the telecommunications such as NP, QoS, QoE and SLAs with their relationships based on ITU-T recommendations. Building blocks and mechanisms to support QoS has been identified. And measurements and assessment of QoS and QoE including reference network models are introduced based on ITU-T recommendations. Using those factors and models of the ITU-T recommendations into IP-based platforms including NGNs is followed with some of examples. This Technical Paper analyzes issues in current IP-based platform (networks) operation in terms of QoS/QoE aspects. Finally key issues require further considerations are also introduced taking into account different environments of the region.

2 Definitions

A number of terms are being used to describe Quality of Service and Quality of Experience as well as IP-based platform (network) related. Following terms are used in this Technical Paper with definitions, mostly based on relevant ITU-T Recommendations.

2.1 cessation [b-ITU-T E.800]: All activities associated with the cessation of a service by a service provider from the instant a contractual agreement is in force between the customer and the service provider to the instant all hardware and software associated with the service is made inoperative and/or removed from the customer's premises.

2.2 class of service [b-ITU-T Y.1401]: A parameter used in data and voice protocols to differentiate the types of payloads contained in the packet being transmitted. The objective of such differentiation is generally associated with assigning priorities to the data payload or access levels to the telephone call.

2.3 customer edge (CE) router [b-ITU-T E.800-Sup.8]: The router at the edge of a customer's network, usually facing towards a provider.

2.4 downstream [ITU-T G.1050]: A transmission from a service provider toward an end user.

2.5 entity [ITU-T E.860]: A generic unit involved in using/delivering a service.

2.6 gateway [ITU-T G.1050]: A network device that acts as an entrance to another network. One function is to convert media provided in one type of network to the format required in another type of network. For example, a gateway could terminate bearer channels from a switched circuit network (e.g., DS0s) and media streams from a packet network (e.g., RTP streams in an IP network).

2.7 IP-based networks [ITU-T Y.1401]: A network in which IP is used as one of the Layer 3 protocols.

2.8 IP performance measurement signature (IPPMS) [ITU-T Y.211]: An IP test packet is a regular IP packet that contains a standardized block of fields needed to perform the measurement. This block of fields is named IP performance measurement signature (IPPMS).

2.9 measurement point (MP) [ITU-T O.211]: The boundary between a host and an adjacent link at which performance reference events can be observed and measured. Consistent with ITU-T Rec. I.353, the standard Internet protocols can be observed at IP measurement points.

2.10 network performance [ITU-T E.417]: The performance of a portion of a telecommunications network that is measured between a pair of network-user or network-network interfaces using objectively defined and observed performance parameters.

2.11 peer-to-peer [ITU-T G.1050]: A distributed application architecture that partitions tasks or workloads between peers. Peers are equally privileged, equipotent participants in the application.

2.12 performance [ITU-T G.1741]: The ability to track service and resource usage levels and to provide feedback on the responsiveness and reliability of the network.

2.13 provider edge (PE) router [ITU-T E.sup8]: The router at the edge of a provider's network, usually facing towards a customer.

2.14 quality [ISO 8402]: The totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs.

2.15 quality of experience [ITU-T P.10 Amd.2]: The overall acceptability of an application or service, as perceived subjectively by the end-user.

NOTE 1 – Quality of experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.).

NOTE 2 – Overall acceptability may be influenced by user expectations and context.

2.16 quality of service [ITU-T Q.1741]: The collective effect of service performances, which determine the degree of satisfaction of a user of a service. It is characterized by the combined aspects of performance factors applicable to all services, such as: service operability performance; service accessibility performance; service retainability performance; service integrity performance; and other factors specific to each service.

2.17 QoS edge routing [ITU-T G.1050]: Routing that typically takes place between the customer premises network and the service provider network based on quality of service **classification values.**

2.18 service level agreement (SLA) [ITU-T E.860]: A formal agreement between two or more entities reached after a negotiating activity with the scope to assess service characteristics, responsibilities and priorities of every part. A SLA may include statements about performance, billing, service delivery but also legal and economic issues.

2.19 sequential packet loss [ITU-T G.1050]: Two or more consecutive lost packets.

2.20 upstream [ITU-T G.1050]: A transmission from an end user toward a service provider.

3 Abbreviations

AN	Access Network
APS	Automatic Protection Switching
BI	Business Interface
CE	Customer Edge
CN	Core Network
CPN	Customer Premise Network
DST	Destination host
ECN	Explicit Congestion Notification
FTP	File Transfer Protocol
HTTP	Hypertext Transport Protocl
IdM	Identity Management
IMS	IP Multimedia Subsystem
IP	Internet Protocol
IPER	IP packet error ratio
IPIBR	IP packet impaired block ratio
IPIIR	IP packet impaired interval ratio
IPLR	IP packet loss ratio
IPPM	IP Performance Metrics
IPRR	IP packet reordered ratio
IPTD	IP packet transfer delay
MOS	Mean Opinion Score
MP	Measurement Point
MPM	Management of Performance Measurement
NACF	Network Attachment Control Functions
NAPT	Network Address and Port Translator
NGN	Next Generation Networks
NP	Network Performance
OWAMP	One-Way Active Measurement Protocol
PDV	Packet Delay Variation
PI	Preliminary Information
PME-FE	Performance Measurement Execution Functional Entity
PMP-FE	Performance Measurement Processing Functional Entity
PMR-FE	Performance Measurement Reporting Functional Entity
PPP	Point-to-Point Protocol
PRS	Performance Reporting System
QoE	Quality of Experience
QoS	Quality of Service
RACF	Resource Admission and Control Function
RIPR	Replicated IP packet ratio
RSVP	Resource Reservation Setup Protocol
RTP	Real-time Transport Protocol
SCF	Service Control Functions

SDH	Synchronous Digital Hierarchy
SLAs	Service Level Agreements
SP	Service Provider
SRC	Source host
TI	Technical Interface
TWAMP	Two-Way Active Measurement Protocol
UDP	User Datagram Protocol

4 Frameworks of NP, QoS and QoE

The term quality of service (QoS) is being extensively used and continuously increased with new telecommunication environments such as regarding broadband, wireless and emerging various multimedia services. In addition, quality of experience (QoE) is being popular taking into account user satisfaction of delivered services. However, the term QoS and QoE are usually used loosely or misused. It is important to understand the QoS and QoE correctly, at least in terms of telecommunications and their services. Therefore this section introduces background information for better understanding of QoS and QoE following the guidance from ITU-T recommendations.

4.1 Network Performance, QoS and QoE

Network Performance (NP)

Identifying the "network performance" should be the first step for studying QoS and QoE. This term has been defined in ITU-T Recommendation E.800 as "The ability of a network or network portion to provide the functions related to communications between users" or defined slightly different way in ITU-T Recommendation E.417 as "The performance of a portion of a telecommunications network that is measured between a pair of network-user or network-network interfaces using objectively defined and observed performance parameters." These two definitions are looked slightly different, but key point which is common should be identified the portion of network parts to perform providing communications between users. Network performance should be measured in terms of parameters which are meaningful to the network provider and are used for the purposes of system design, configuration, operation and maintenance. And it should be defined independently of terminal performance and user actions.

Network performance contributes towards QoS as experienced by the user/customer. Network performance may or may not be on an end to end basis. For example, access performance is usually separated from the core network performance in the operations of single IP network, while Internet performance often reflects the combined NPs of several autonomous networks. [b-ITU-T G.1000]

Quality of Service (QoS)

ITU-T Rec. E.800 defines QoS as "the collective effect of service performance which determine the degree of satisfaction of a user of the service." As noted, QoS should be derived the results from the performing activity of the functions which involved to provide services to the user. This called "Network Performance."

QoS comprises both network performance and non-network related performance. Examples of NP are bit error rate, latency, etc., and for non-network performance provision time, repair time, range of tariffs and complaints resolution time, etc. The list of QoS criteria for a particular service would be dependent upon the service and their relevance could vary among the segments of the customer population. [b-ITU-T E.800] Figure 1 shows the relationship between QoS and network performance (NP).

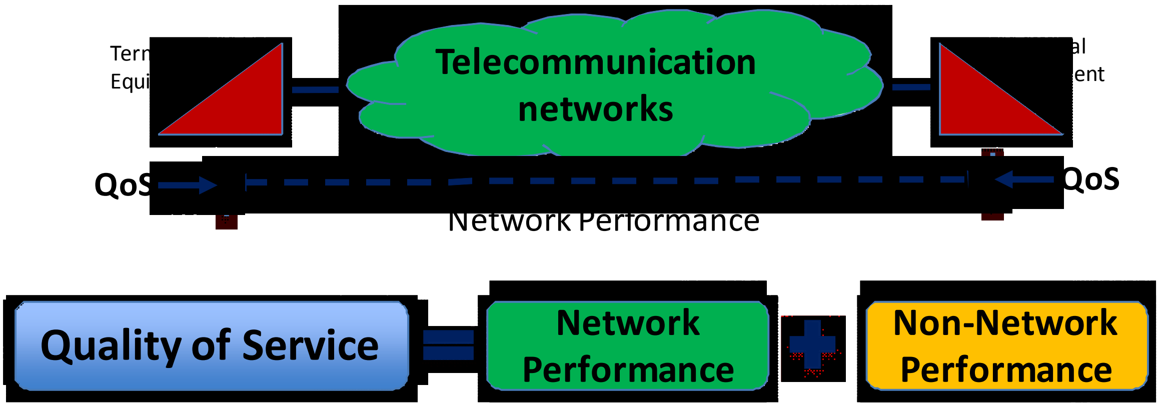


Figure 1 – Relationship between NP and QoS

QoS can be looked from different perspectives for judging the quality of the functions which involved in providing the service. ITU-T Recommendation E.800 and G.1000 introduces 4 different views in this perspective as following with Figure 2 showing the "top down" relationship of these viewpoints:

- Customer's QoS requirements;
- Service provider's offerings of QoS (or planned/targeted QoS);
- QoS achieved or delivered;
- Customer survey ratings of QoS.

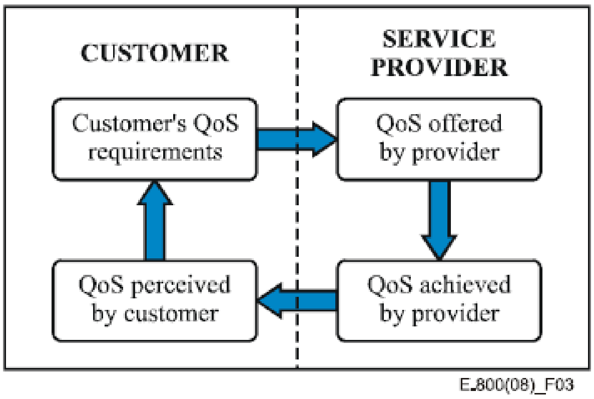


Figure 2 – The four viewpoints of QoS

It is noted that it must be meaningful from these four viewpoints if any framework of QoS to be truly useful and practical enough to be used across the industry. Features and requirements of each perspective are summarized following:

Customer's requirements of QoS: state the level of quality required of a particular service. The customer is not concerned with how a particular service is provided, or with any aspects of the network's internal design, but only with the resulting end-to-end service quality. Quality of service from the customer's view is expressed by parameters, which:

- focus on user-perceived effects, rather than their causes within the network;
- do not depend, in their definition, on assumptions about the internal design of the network;
- take into account all aspects of the service from the customer's point of view;
- may be assured to a user by the service providers, sometimes in contractual terms;
- are described in network-independent terms and create a common language understandable by both the user and the service provider.

QoS offered by the service provider: a statement of the level of quality expected to be offered to the customer by the service provider (expressed by values assigned to QoS parameters). Each service would have its own set of QoS parameters (as in the QoS Classes of ITU-T Rec. Y.1540 for

IP service offers). QoS offered by the service provider can be used in planning documents, to specify measurement systems and also can be used to form the basis of the Service Level agreements (SLAs).

QoS achieved or delivered by the service provider: a statement of the level of quality actually achieved and delivered to the customer expressed by values assigned to parameters. These parameters should be the same as specified for the offered QoS so that the two can be compared to determine what was actually achieved to assess the level of performance achieved. These performance figures are summarized for specified periods of time, e.g. for the previous month.

QoS perceived by the customer: a statement expressing the level of quality experienced and expressed, usually degrees of satisfaction, not in technical terms. Perceived QoS is assessed by customer surveys and from customer's own comments on levels of service. Perceived QoS can be used by the service provider to determine customer satisfaction of the service quality. Ideally there would be 1:1 correspondence between delivered and perceived QoS.

Quality of Experience (QoE)

QoE is quite different with QoS and NP, because it has subjective feature as noted in the definition such as "The overall acceptability of an application or service, as perceived subjectively by the end-user." QoE has a dependency of end user perception as well as features of services, thus it could have quite different ways to specify the value. But it is clear QoE should be impacted from the QoS and NP even though end user subjective. By considering these features, the relationship among NP, QoS and QoE should be illustrated as shown in Figure 3.

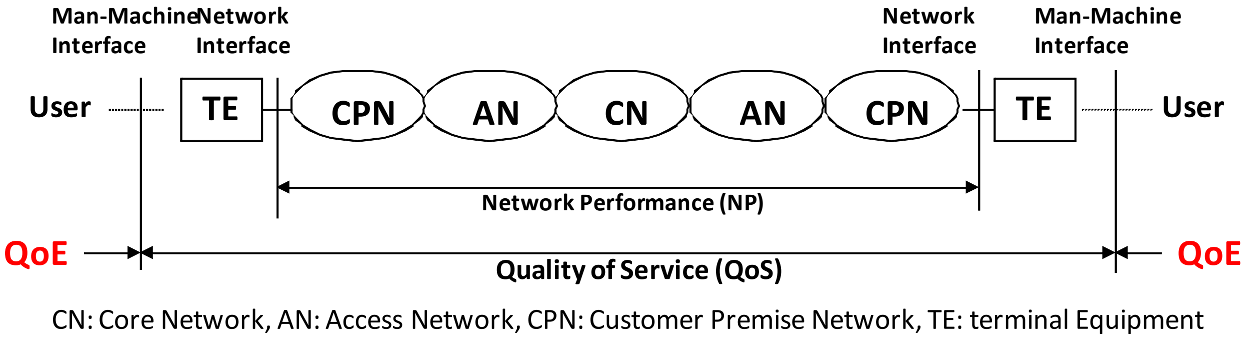


Figure 3 – Relationship among NP, QoS and QoE

Taking into account those differences given by definitions of NP, QoS and QoE, key features of these three should be summarized as Table 1 based on ITU-T Recommendation I.350. QoE and QoS have user oriented view while NP has provider oriented. And QoS has service specifics while QoE has user subjective specifics.

Table 1 – Summary features of NP, QoS and QoE

Quality of Experience	Quality of Service	Network Performance
User oriented		Provider oriented
User behaviour attribute	Service attribute	Connection/Flow element attribute
Focus on user-expected effects	Focus on user-observable effects	Focus on planning, development (design), operations and maintenance

User subject	Between (at) service access points	End-to-end or network elements capabilities

4.2 SLA and relationship with QoS

Service and network providers have big pressures from their customers following the competition of the business in terms of enhancing the quality of services while reduction of costs in order to differentiate their products from those of their competitors. In addition the situation is complicated by the increasing demand of global services which involve in their provisioning several service and network providers. Therefore, roles and their relationship of all entities that take part into service provision have to be described as clear as possible to assure quality of service required from customer.

A useful tool in formalizing the relationships between entities is Service Level Agreement (SLA) that is the result of a negotiation between two or more parties with the objective of reaching a common understanding about the service delivered, its quality, responsibilities, priorities, etc. This section describes frameworks of SLA. [b- ITU-T G.1050]

Structure of a SLA

A SLA refers to all services exchanged between two entities (multi-services SLA) as shown in Figure 4. [b- ITU-T G.1050]

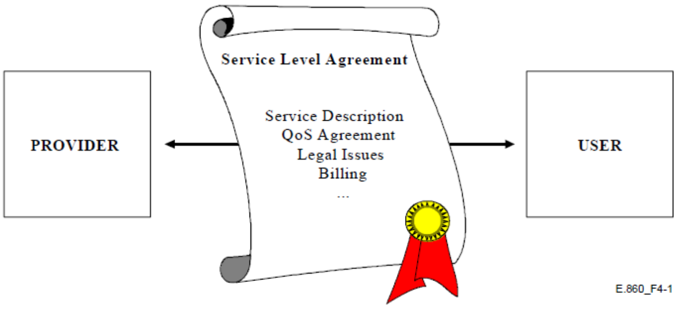


Figure 4 – SLA between two entities

In general, SLA is made up of one common part and of other service specific parts. Following Figure 5 shows a generic structure of SLA which recommended from ITU-T Recommendation.

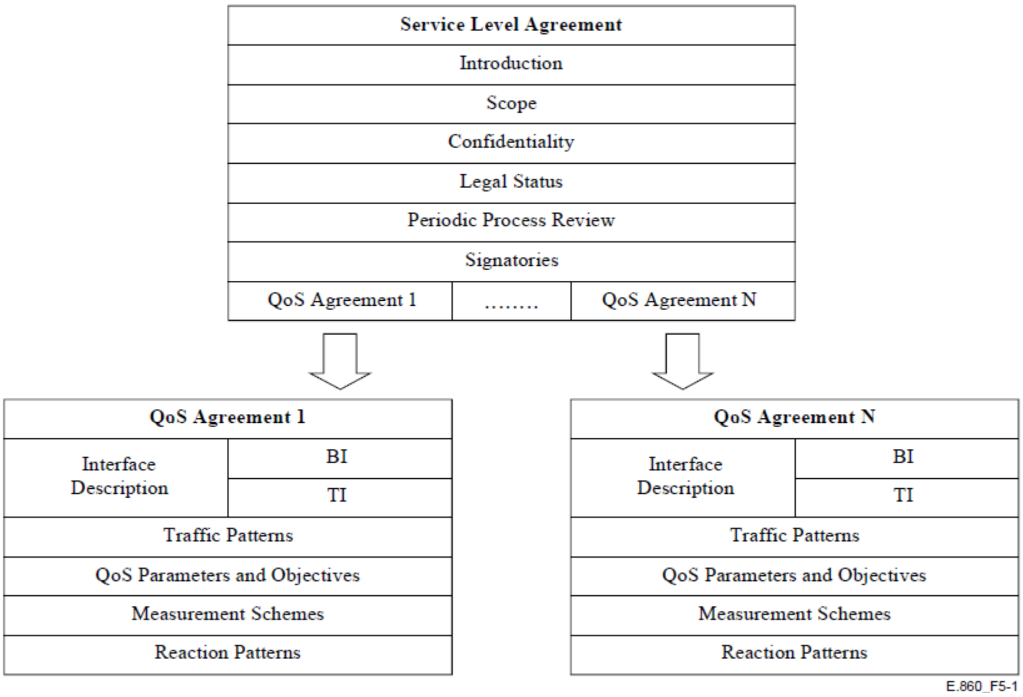


Figure 5 – Generic structure of SLA

Key features of common part of SLA identified as following:

- Introduction: describes the purpose of the SLA that may be:
  - to define service levels that all entities have to guarantee for customer's satisfaction;
  - to assist two entities (User, Service Provider, Network Provider) in exchanging information with suitable QoS and Network Performance;
  - to provide base notions of measurements and parameters for realization of the agreement.
- Scope: describes, in a general manner, services which the SLA deals with and their target performance
- Confidentiality: specifies the treatment of the agreement and the sharing of information between the involved parties (confidential information not be disclosed to entities which are not part of the agreement). This does not apply to public SLA signed with Regulatory authority.
- Review process: defines the frequency (daily, monthly, semi-annual, etc.) and format (paper, electronic) with which QoS information has to be exchanged. It may specify also the frequency of review of the QoS Agreement, so it can be always up to date with actual technology and customer's expectations, but this part may be optional.
- Signatories: A part for sign the agreement to ensure all obligations undertaken by authorized representative.

Key features of specific part of SLA identified as following:

- Interface description: Identify the logical boundary (interface) between two entities. It is composed of a group of interaction points which exchange information with the service provider who, at least virtually, controls all interaction points. Regarding the type of information exchanged, interface description is categorized as:
  - Business Interface (BI): composed of interaction points between user and SP used for specific QoS Agreement functions as well as (re)negotiation, performance reporting and reaction patterns which are triggered when the agreed QoS level is not provided ,or

– Technical Interface (TI): exchange service specific information and allow measurements from which QoS parameters are derived.

- Traffic patterns: Identify the characteristics of traffic that exchanges (receives and send) from other entities (traffic at the ingress points) in order to manage its own resources properly. This includes the conditions (thresholds) that enable activation of reaction patterns from the receiving entity, thus when incoming traffic not conforms the agreed one, the receiving entity may react with mechanism as well as traffic shaping. Description of traffic patterns should be well understood from entities at the both sides of the interface.
- QoS parameters and objectives: Identify relevant parameters to specify the quality of services with their objectives expressed by target values, thresholds and ranges set to QoS parameters.
- Measurements: A description of what, when, where and who should perform measurement procedures and test processes. The methodology to evaluate measurement results is also important and may be included in this part of SLA.
- Reaction patterns: A reaction is a process that is activated in a more or less automated way whenever commitments on traffic patterns and on QoS parameters are not fulfilled.

Determination of QoS parameters from SLA

As explained, SLA is a result of contracting two entities referring all aspects of services, so relevant parameters for QoS should be determined from the SLA. Following Figure 6 shows a way to individuate QoS parameters between two entities taken into consideration of both customer and network parameters. The overall service level requirements which described in the SLA provide relevant information to identify network QoS parameters (both service independent and dependent) and customer QoS parameters. These parameters contribute to identify inter-operator related QoS parameters also both service dependent and independent. Service dependent and service independent parameters allow determining the last ones in a common manner for all services, which simplifies their utilization. [b- ITU-T E.860]

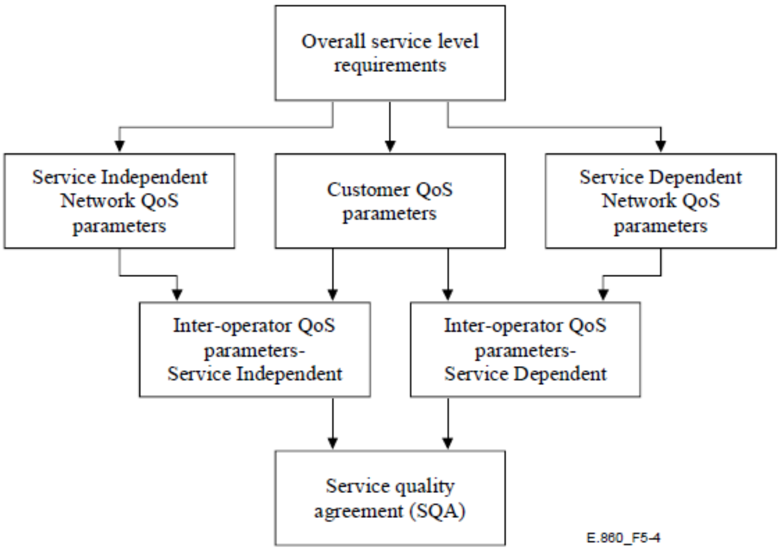

Overall service level requirements
Service Independent Network QoS parameters
Customer QoS parameters
Service Dependent Network QoS parameters
Inter-operator QoS parameters- Service Independent
Inter-operator QoS parameters- Service Dependent
Service quality agreement (SQA)
E.860_F5-4

Figure 6 – Determination of QoS parameters

SLA in multi-provider environment

Several service providers are often involved in a service provision and collaborate together in realizing the various service elements forming a multi-provider environment. Consider the case when a SLA between an end user and a provider, for a connection passing through several service providers (SP) domains, is agreed. Using the one stop responsibility, the end user will require the agreed QoS exclusively from the service provider with whom he agreed upon the SLA, while the latter will have to guarantee that QoS by signing, in its turn, suitable SLAs with its sub-providers. A traditional approach to multi-provider environment consists in making an association of entities (such as SPs) (Figure 7) which all agreed upon a common document dealing with parameters, objectives and techniques of measurement for QoS. Such an approach assures that End-to-End QoS of connections, which pass through several SPs, will fulfill QoS agreed with end user in the SLA. [b- ITU-T G.1050]

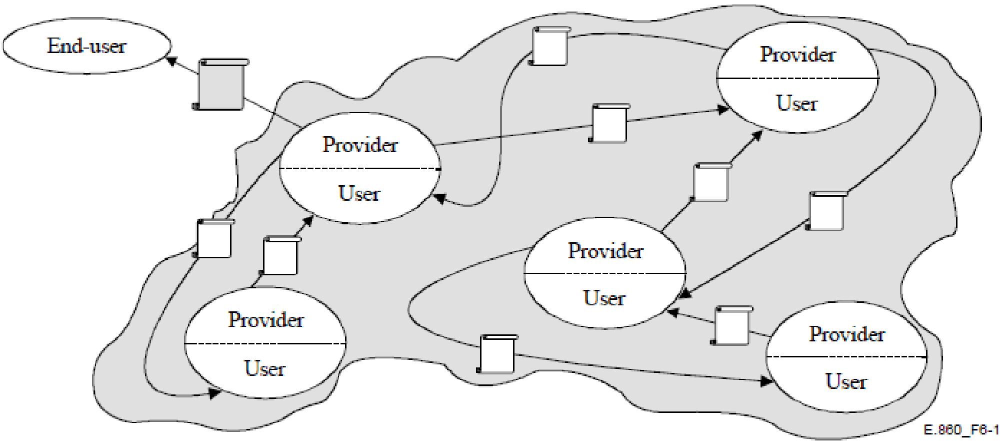


Figure 7 – A SLA in multi-provider environments

4.3 Parameters for QoS

QoS parameters (alternatively named as QoS metrics, QoS indicators, QoS measures or QoS determinants) characterize the quality level of a service being offered, and the level of the customer satisfaction. QoS parameters represent subjective and abstract user-perceived "quality" in terms of numeric (quantified) values. QoS parameters can be used by providers (network and service providers) managing their services offered, as well as by the customers (end users or partner providers) to ensure getting the level of quality that they are paying for.

QoS parameters can be obtained from objective and/or subjective measurement methods. Objective QoS parameters, normally used as internal indicators for service quality characterization and improvement, are obtained from measurement of physical attributes of circuits, networks and signals. The subjective QoS parameters are obtained by from the customer opinion surveys and are normally used as an external indicator, e.g. for customer relationship management. [9]

ITU-T Recommendation I.350 categorizes generic QoS parameters in three functions (Access, User information transfer, and Disengagement) and three performance criteria (Speed, Accuracy, and Dependency). This resulted in nine possible combinations, represented by a 3 × 3 matrixes as shows in Figure 8. Under this generic framework, specific QoS metrics can be defined to serve different applications and service types, such as:

- QoS metrics for different signal types, e.g. voice, fax, video;
- QoS metrics for one signal type (voice), but different service types, e.g. telephony conversational voice, voice-mail and streaming audio, which have different requirements for latency delay, and;
- QoS metrics for the same signal type and service type, but different classes of commercial offering, e.g. premium-price high-quality telephone voice service, as opposed to toll-free best-effort telephone service with advertisements.

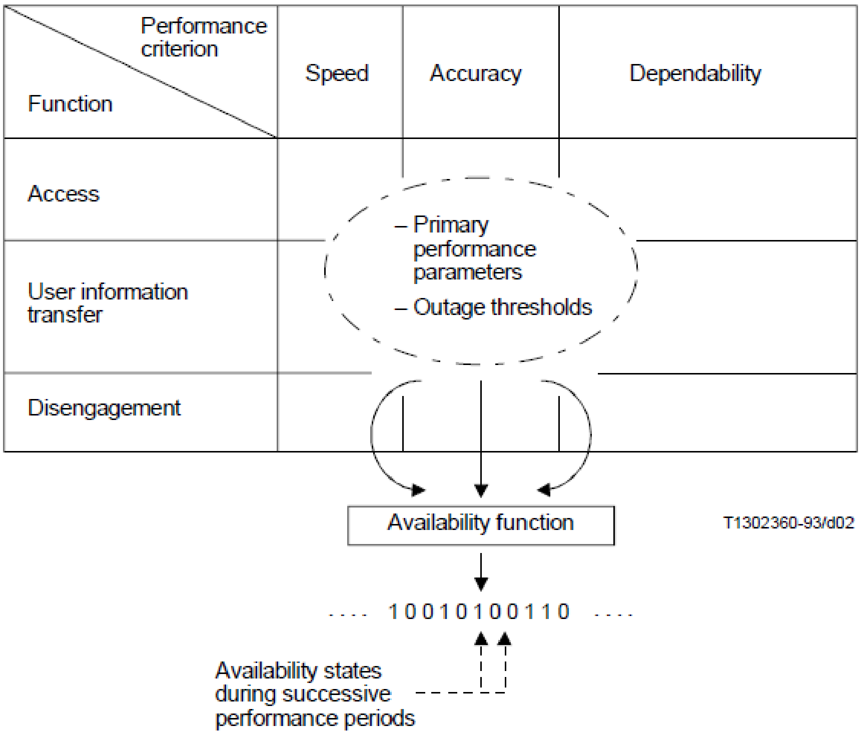


Figure 8 – 3x3 matrix approach and determination of availability states

A prerequisite to effective QoS management is the existence of QoS parameters or metrics, which should be simple to use, proven accurate representations of customer perception, and commonly accepted as standards. QoS parameters have inter-relationship among themselves as shown in Figure 9 [1]. Some parameters specifying network factors can be impacted to application factors. Similarly some parameters belong to application factors also can be impacted to service level of QoS which lead QoE. Thus it is anticipated to identify QoS taking into account such relationships among parameters.

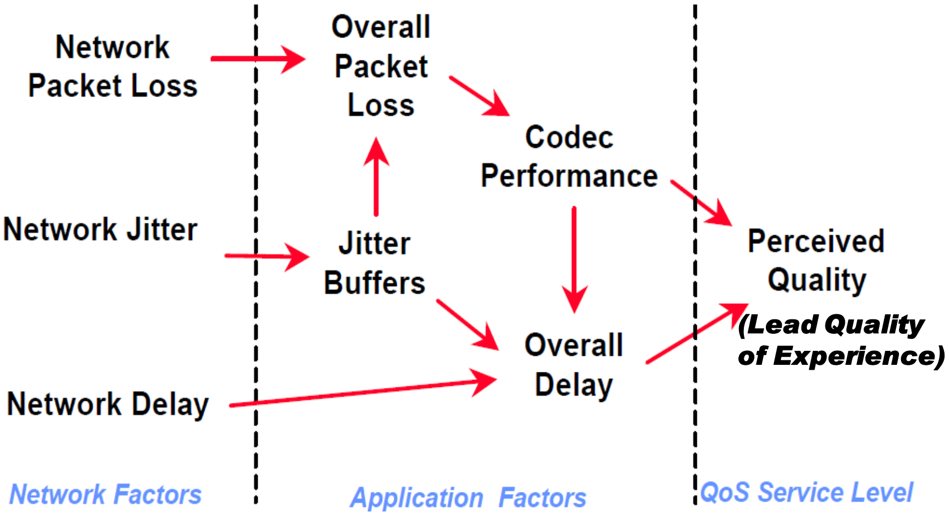


Figure 9 – Inter-relationship of QoS parameters

Following QoS parameters have been identified as being potentially useful for comparison of performance levels of service providers. [b- ITU-T E.803]

Preliminary information on ICT services

[Parameter 1] Integrity of preliminary information: is characterized by a true and fair view of the main points of an ICT service provided to the potential customers by the service provider. This parameter measured as "*Opinion rating*."

[Parameter 2] Pricing transparency: is characterized by clarity, conciseness and unambiguity in every tariff structure for all usage conditions for every service provided by the service provider. This parameter measured as "*Opinion rating*."

[Parameter 3] Availability of PI (Preliminary Information): Ratio of the number of requests for PI from potential users and customers which have been delivered to the total number of requests within a pre-defined time interval. This parameter measured as "*Fraction or Percentage*."

[Parameter 4] Response time for the provision of PI: Time taken from the instant a request for PI was sent to the SP to the instant all requested information was delivered to the customer requesting the information. This parameter measured as "*Time*."

Contractual matters between ICT service providers and customers

[Parameter 5] Integrity of contract information: True and fair view of pertinent information on supply, maintenance and cessation for a telecommunications service provided by a service provider. This parameter measured as "*Opinion rating*."

NOTE 1 – A contractual document describing the supply, maintenance and cessation for a telecommunication service by a SP is clear, accurate, complete, understandable and unambiguous.

NOTE 2 – The language, phrasing and expressions chosen are aimed at maximum understanding for the target customer segment.

[Parameters 6] Compliance of contractual terms with preliminary information: Degree of concurrence of the contents of the contractual document to the PI. This comparison between contractual terms and PI should be based on the PI in force during the period of the contract. Contractual document could have detailed terms which were implicit in the PI. Where differences exist these are not to be considered as errors as long as additional and non-contradictory information is provided. This parameter measured as "*Ratio or Percentage*."

[Parameter 7] Flexibility for customization before contract: The scope and boundary to meet individual customer's specific requirements of service feature(s), service performance(s) and terms and conditions before formal signature on the contract. This parameter measured as "*Opinion rating*."

NOTE – These specific requirements would be departures from the standard service features, performance and terms and conditions normally offered by the service provider.

[Parameter 8] Ease and flexibility to amend terms after formal contract: The scope and boundary of the amendments that could be accommodated to contractual terms to satisfy the post contractual amendments sought by a customer. This excludes contracts which the provider has specifically stated as not considered for amendments. This parameter measured as "*Opinion rating.***"**

Provision of services

[Parameter 9] Meeting promised provisioning date: Successful completion of provisioning of service on the date promised in the contract in relation to the total number of signed contracts with promised service provisioning dates. This parameter measured as "*Ratio or Percentage.*"

[Parameter 10] Time for provisioning: Period of time between the scheduled provisioning time and the actual provisioning time. This parameter measured as "*Time*."

[Parameter 11] Successful provisioning within specified period: Number of successful service provisioning events in relation to all expected provision events within a pre-defined period of time. This parameter measured as "*Ratio or Percentage*."

[Parameter 12] Contract cancelled due to non-fulfillment: Contracts cancelled due to the ongoing non-fulfillment and considered unreasonable to wait any longer to the total number of signed contracts within a given assessment period. This parameter measured as **"***Ratio or Percentage*."

[Parameter 13] Completeness of fulfillment of contractual specification in the provision of a service: Contracts with all network and/or service features specified in the contract fulfilled (after its provisioning) in relation to the number of contracts that have been considered fulfilled for provisioning. This parameter measured as **"***Ratio or Percentage.*"

[Parameter 14] Punctuality of service provisioning: Time difference between the actual service provisioning and that contractually specified. This parameter measured as "*Time*."

[Parameter 15] Punctuality of equipment delivery of service provisioning: Time difference between the actual equipment delivery and the scheduled delivery announced by the service provider for the service provisioning. This parameter measured as "*Time.*"

[Parameter 16] Provisioning not complete and correct first time: Ratio of service provisioning that is either not completely carried out or not correctly carried out in the first attempt, to the total number of contracts where the provisioning is deemed completed. This parameter measured as "*Ratio or Percentage.*"

NOTE – The indicator for this parameter provides how well the SP has performed in complete and correct provisioning at the first attempt.

Service alteration

[Parameter 17] Time for alteration of service: Time elapsed from the instant alteration notification is received by the user to the instant the alteration is completed. This parameter measured as "*Time.*"

[Parameter 18] Successful service alteration within specified period: Ratio (percentage) of the number of contracts (or services) with successful service alteration to the total number of contracts (or services) with announced service alteration within the contractual specified period of time. This parameter measured as "*Ratio or Percentage*."

[Parameter 19] Completeness of fulfillment of contractual specification in the alteration of a service: The ratio of all contracts where all specifications related to the service alteration contractually agreed are met or completed to the total number of contracts where alteration has been requested. This parameter measured as "*Ratio or Percentage*."

[Parameter 20] Punctuality of appointments for service alteration: Time difference between the actual service alteration and the scheduled alteration time announced by the service provider. This parameter measured as "*Time*."

[Parameter 21] Punctuality of equipment delivery for service alteration: Time difference between the actual equipment delivery and the scheduled delivery announced by the service provider. This parameter measured as "*Time*."

[Parameter 22] Service alteration not complete and correct first time: Ratio(percentage) of service alterations that were either not completely or not correctly carried out in the first attempt, to the total number of contracts where alterations have been requested. This parameter measured as "*Ratio or Percentage*."

[Parameter 23] Conformity and success of service alteration: Ratio of the number of contracts where service alterations were not according to specification and therefore requiring reworking or further service alteration, to the total number of contracts where alteration was requested. This parameter measured as "*Ratio or Percentage*."

[Parameter 24] Technical reliability of service within an agreed period after alteration: Number of observation phases after service alteration without any limitation to the total number of service alterations carried out. This parameter measured as "*Ratio or Percentage*."

[Parameter 25] Organizational efficiency of service provider to carry out service alteration: Organizational and hardware resource availability to carry out service alterations to meet the needs of the customer and/or to meet contractual promises. This parameter measured as "*Opinion rating*."

Technical upgrade of ICT services

[Parameter 26] Time for technical upgrade of a service: Time elapsed from the instant the technical upgrade period was announced to the user to the instant the technical upgrade was carried out. This parameter measured as "*Time*."

[Parameter 27] Successful technical upgrade within a specified period of time: Ratio of successful service technical upgrades carried out in a specified time-out interval to the total number of technical upgrades carried out within the same period. This parameter measured as "*Ratio or Percentage*."

[Parameter 28] Completeness of fulfillment of specification in the technical upgrade of a service: Ratio of the number of successful upgrades where all specification requirements were met to the total number of contracts with such upgrades scheduled in a specified period. This parameter measured as "*Ratio or Percentage*."

[Parameter 29] Punctuality of appointments for technical upgrade: Time difference between the actual technical upgrade and the scheduled upgrade time announced by the service provider. This parameter measured as "*Time*."

[Parameter 30] Outage time due to technical upgrade: Duration when the service in part or in full is unavailable to the customer for use due to the technical upgrade process. This parameter measured as "*Time*."

[Parameter 31] Technical upgrade not complete and correct first time: Ratio (percentage) of the number of contracts not completely carried out or not correctly carried out in the first attempt to the total number of contracts. This parameter measured as "*Ratio or Percentage*."

NOTE – The indicator for this parameter provides how well the SP has performed in complete and correct technical upgrade at the first attempt.

[Parameter 32] Conformity and success of technical upgrade: Ratio of technical upgrade not according to specification and therefore requiring reworking or further service upgrade processes and resources to get it right, to the total number of contracts upgraded. This parameter measured as "*Ratio or Percentage*."

[Parameter 33] Technical reliability of service within an agreed period after technical upgrade: Ratio of the upgrades that perform satisfactorily for a specified period after the upgrade to the total number of upgrades carried out. This parameter measured as "*Ratio or Percentage*."

[Parameter 34] Organizational efficiency of service provider to carry out technical upgrade: Organizational and hardware resource availability on the part of the SP to carry out technical upgrades to meet the needs of the customer and/or to meet contractual promises. This parameter measured as "*Opinion rating*."

[Parameter 35] Competence and preparedness of service provider for technical upgrade: Degree of ability (competence) and willingness (preparedness) to incorporate technical upgrade relevant to the service for the benefit of users. This parameter measured as "*Opinion rating*."

Documentation of services (operational instructions)

[Parameter 36] Documentation of delivery time: Time taken from the instant a service is provided to the instant documentation for the commissioning and use of the service is delivered to the customer. This parameter measured as "*Time*."

NOTE – Documentation not delivered before a specified timeout will be considered as not delivered in time.

[Parameter 37] Availability of documentation within specified period of time: Number of contracts where documentation was supplied within a specified period of time to the total number of contracts where documentation was expected. This parameter measured as "*Ratio or Percentage*."

[Parameter 38] Integrity (correctness and completeness) of documentation: Correctness, completeness and user friendliness of pertinent information associated with the use of all features of a service and its maintenance. This parameter measured as "*Opinion rating*."

[Parameter 39] Modes of documentation: Number of modes in which documentation is made available to the customer or user of a service. This parameter measured as "*Number*."

[Parameter 40] Legibility of documentation: Visual clarity, language, understandability and layout of the information in the medium in which it is presented. This parameter measured as "*Opinion rating*."

[Parameter 41] Overall reliability of documentation services: Consistent availability, integrity and speed of provisioning of the documentation and associated support activities provided by the SP for a given service. This parameter measured as "*Opinion rating*"

Technical support provided by service provider

[Parameter 42] Accessibility to technical support: Ratio of the number of successful attempts to technical support to the total number of attempts to reach this support. This parameter measured as "*Ratio or Percentage*."

[Parameter 43] Technical solutions achieved within a specified period: Ratio of the number of contracts with successful technical solutions applied, to the total number of contracts where solutions were sought and applied within the specified period. This parameter measured as "*Ratio or Percentage*."

[Parameter 44] Number of attempts before successful solutions: Number of attempts before the technical request was successfully resolved. This parameter measured as "*Number*."

[Parameter 45] Integrity of technical solutions: Proportion of successful solutions with respect to the total number of requests within a specified period of time. This parameter measured as "*Opinion rating*."

[Parameter 46] Reliability of technical solutions achieved: Ratio of number of services that were trouble-free for a specified period of time after the technical solution was resolved, to the total number of services where the technical support was requested and implemented. This parameter measured as "*Ratio or Percentage*."

[Parameter 47] Modes of technical support: Number of modes in which technical support is available to the customer or user of a service. This parameter measured as "*Number*."

Commercial support provided by service provider

[Parameter 48] Accessibility of the commercial support: Ratio of the number of successful access attempts to the commercial support to the total number of attempts to reach this support. This parameter measured as "*Ratio or Percentage*."

[Parameter 49] Commercial solution delivery time: Time elapsed from the instant the customer raised a problem with commercial support to the instant a solution was achieved. This parameter measured as "*Time*."

[Parameter 50] Commercial solutions achieved within a specified period of time: Ratio of the number of contracts with successful commercial solutions achieved, to the total number of contracts where solutions were sought within a specified period. This parameter measured as "*Ratio or Percentage*."

[Parameter 51] Integrity of commercial solutions achieved by service provider: Ratio of successful solutions achieved within the specified period of time to the total number of commercial support requests. This parameter measured as "*Opinion rating*."

[Parameter 52] Modes of commercial support: Number of modes in which commercial support is available to the customer or user of a service. This parameter measured as "*Number*."

[Parameter 53] Organizational efficiency of commercial support: Availability of organizational resource to fulfill customer needs on commercial support. This parameter measured as "*Opinion rating*."

Complaint management

[Parameter 54] Accessibility of the complaint management: Ratio of the number of successful attempts to the total number of attempts to reach complaint management (CM) in a specified period. This parameter measured as "*Ratio or Percentage*."

[Parameter 55] Recognition of the customer complaints: Ratio of the customer claims recognized by the SP as genuine complaints to the total number of potential complaints. This parameter measured as "*Ratio or Percentage*."

[Parameter 56] Complaint solutions not complete and correct first time: Ratio of the number of complaints not successfully resolved at the first attempt to the total number of complaints received by the service provider. This parameter measured as "*Ratio or Percentage*."

NOTE – The indicator for this parameter provides how well the SP has performed in the complete and correct handling of the customer complaint at the first attempt.

[Parameter 57] Integrity of complaint resolution: Ratio of the number of complete and professional resolutions of the contributory causes of a complaint, to the total number of user complaints accepted. This parameter measured as "*Ratio or Percentage*."

[Parameter 58] Customer perception of the complaint management: The service provider's exhibition of the combination of assurance, empathy and responsiveness in dealing with complaint(s) from reporting to satisfactory resolution. This parameter measured as "*Opinion rating*."

[Parameter 59] Overall quality of the complaint management process: The combined effect of accessibility of the complaint management service: correct solutions at the first attempt, speed of resolution and the organizational capability to carry out these services. This parameter measured as "*Opinion rating*."

[Parameter 60] Organizational efficiency of complaint management system: The availability and deployment of organizational and hardware resources on the part of the service provider to resolve user complaints. This parameter measured as "*Opinion rating.*"

Repair services

[Parameter 61] Accessibility of repair services: Availability of hardware, software and staff resources necessary to restore a service (and its features) to its specified level of performance. This parameter measured as "*Ratio or Percentage*."

[Parameter 62] Successful repairs carried out within a specified period of time: Ratio of the number of repairs successfully carried out to the total number of repair requests accepted by the SP within a specified period. This parameter measured as "*Ratio or Percentage*."

[Parameter 63] Repairs not complete and correct first time: Ratio of the number of repairs which were not successfully carried out at the first (and only) attempt to the total number of repairs carried out during the specified period. This parameter measured as "*Ratio or Percentage*."

[Parameter 64] Punctuality of appointments for repairs: Record of attendance of a SP agent to carry out repair at the specified time (allowing, if necessary, a grace period for lateness). It may also be expressed as an opinion rating of customers. This parameter measured as "*Opinion rating and/or Time*."

[Parameter 65] Efficiency of the repair services: "Efficiency of the repair service" (mainly technical) of a service provider is characterized by the combined performances of:

- accessibility,
- the number of repairs in a specified period of time,
- repairs carried out successfully the first time,
- punctuality.

This parameter measured as "*Opinion rating*."

[Parameter 66] Organizational efficiency of repair services: "Organizational (or operational) efficiency of repair service" is characterized by the combined performances of:

- punctuality,
- time to repair,
- provision of resources (human, hardware and software),
- the organizational logistics to provide an effective repair service.

This parameter measured as "*Opinion rating*."

[Parameter 67] Notification of root cause of outage: Ratio of the number of repairs, the root causes of which were shared with the customer, to the total number of repairs carried out. This parameter measured as "*Ratio or Percentage*."

Charging and billing

[Parameter 68] Accessibility of tariff information: Ratio of the number of successful attempts to the total number of attempts to reach this facility located as indicated in the contract or regulations (access details to this facility to be provided by the service provider). This parameter measured as "*Ratio or Percentage*."

[Parameter 69] Successful notification of exceeding billing budget: Ratio of the number of successful notifications by the SP of exceeding the customer's billing budget to the total number of exceeding customer's billing budget events. This parameter measured as "*Ratio or Percentage*."

[Parameter 70] Notification time (delay) of exceeding billing budget: Time from the instant of billing budget overrun to the instant of the reception by the customer of this notification from the service provider. This parameter measured as "*Time*."

[Parameter 71] Accessibility of the account management: Ratio of the number of successful attempts to the total number of attempts to reach the account management. This parameter measured as "*Ratio or Percentage*."

[Parameter 72] Time to update charging information: The time between the use of service and the instant the related charging information is available on the account. This parameter measured as "*Time*."

[Parameter 73] Timeliness of bill delivery: The ratio of the number of bills delivered within the bill expectation period divided by the number of bills expected within the observation period. This parameter measured as "*Ratio or Percentage*."

[Parameter 74] Bill delivery delay: The delay between the expected time of bill and its receipt by the customer. This parameter measured as "*Time*."

[Parameter 75] Late notification of amount due: The ratio of the number of bills whose "Direct Debit" amount was not advised to the customers before payment was taken from their account to the total number of "Direct Debit" payment arrangements in place. This parameter measured as "*Ratio or Percentage*."

[Parameter 76] Modes of billing information transfer: The number of modes offered by the SP to communicate the billing information to the customers. This parameter measured as "*Number*."

[Parameter 77] Organizational efficiency of the billing service: "Organizational efficiency of the billing service" of a SP is described and measured by the organizational and hardware resource availability to carry out the billing service. This parameter measured as "*Opinion rating*."

Network/Service management by customer

[Parameter 78] Outage duration: The total time a network/service management facility was not accessible to the customer during a specified reporting period. This parameter measured as "*Time*."

[Parameter 79] Frequency of outages: The number of times access to the network/service management facility was not available to the customer during a specified period divided by the duration of this period. This parameter measured as "*Number*."

[Parameter 80] Response time for reply to requests: The time elapsed from the instant customer requests access to the network/service management facility to the instant such a request was carried out. This parameter measured as "*Time*."

[Parameter 81] Successful request response: The ratio of the number of requests made by the customer successfully handled (within a specified time-out period) to the total number of requests made over the observation period. This parameter measured as "*Ratio or Percentage*."

[Parameter 82] Overall reliability of network/service management service: The consistent combined performance of availability, response times, response rates, correctness and completeness in the processing and fulfillment of customer requests for network/service management facilities. This parameter measured as "*Opinion rating*."

[Parameter 83] Organizational efficiency of the network/service management service: Described and characterized by the combined effects of human, network and other pertinent resources made available by the SP to process and fulfill any volume of customer requests to the network/service management facility on a 24/7 basis. This parameter measured as "*Opinion rating*."

[Parameter 84] Reliability of planned outage notification: Ratio of the number of advanced notification of planned outage to customers by a service provider to the total number of planned outage carried out. This parameter measured as "*Ratio or Percentage*."

Cessation of service

[Parameter 85] Cessation acknowledgement time: The time elapsed from the instant of sending the cessation request to the instant of receipt by the customer of the acknowledgment from the service provider. This parameter measured as "*Time.*"

[Parameter 86] Cessation request acknowledgement: The ratio (percentage) of the number of cessation requests that were acknowledged to the number of such requests made in a specified period. This parameter measured as "*Ratio or Percentage*."

[Parameter 87] Accessibility of the cessation facility: The ratio (percentage) of the number of successful attempts to the total number of attempts to reach the cessation facility. This parameter measured as "*Ratio or Percentage*."

[Parameter 88] Contractual cessations achieved: The ratio (percentage) of the number of contractual cessations requested to the total number of such requests made within a specified period. This parameter measured as "*Ratio or Percentage*."

5 Building blocks and mechanisms for QoS

QoS is about supporting the characteristics and properties of specific applications which different applications may have quite different requirements. For example, for telemedicine, the accuracy of the delivery is more important than overall delay or packet delay variation (i.e., jitter), while for IP telephony, jitter and delay are key and must be minimized. The architectural framework for QoS support is the focus of certain mechanisms within the network delivers required network performance. [b- ITU-T Y.1291]

5.1 QoS building blocks

Certain network mechanisms related with QoS can be specific to a network element or for signaling between network elements, or for controlling and administering traffic across a network. Key focus on architectural framework for QoS which forms of QoS building blocks identifies a set of generic network mechanisms for controlling the network service response to a service request. A QoS building block may be specific to a network node (as exemplified by buffer management) or applicable to a network segment (as exemplified by QoS routing) which requires signaling between network nodes, whether they are part of a network segment (e.g., end to end, end to edge, edge to edge, or network to network).

Applying the general reference architecture model of telecommunication, QoS building blocks are organized into three planes as shown in Figure 10 and summarized following as explained in [b-ITU-T Y.1291]:

- Control Plane: contains mechanisms dealing with the pathways through which user traffic travels (e.g., admission control, QoS routing, and resource reservation);
- Data Plane: contains mechanisms dealing with the user traffic directly (e.g., buffer management, congestion avoidance, packet marking, queuing and scheduling, traffic classification, traffic policing and traffic shaping) and;
- Management Plane: contains mechanisms dealing with the operation, administration, management aspects of the network (e.g. SLA, traffic restoration, metering and recording, and policy).

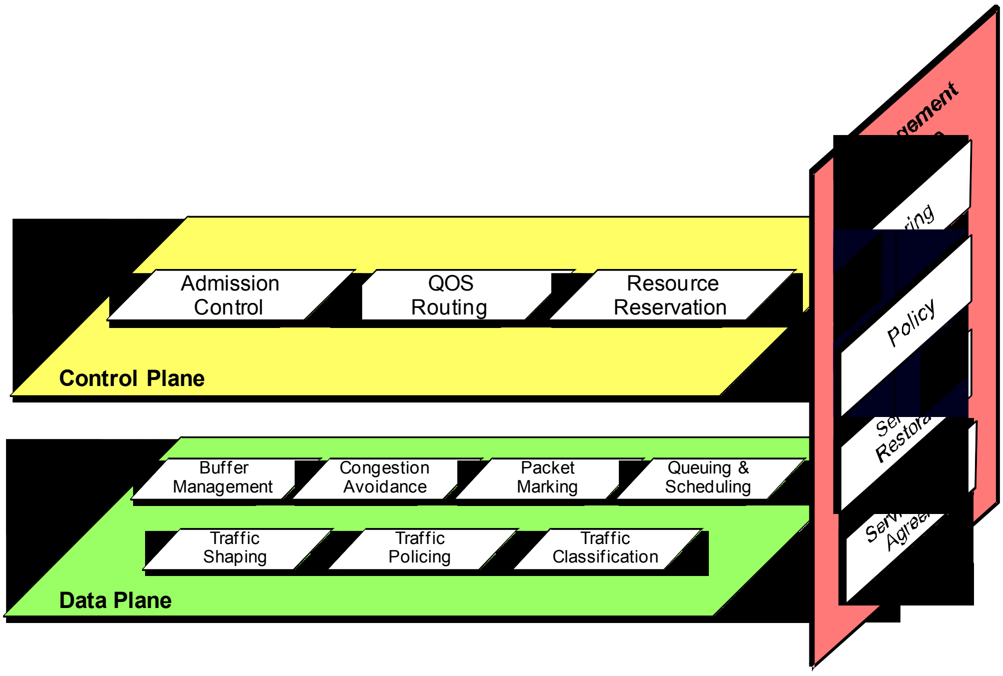


Figure 10 – Architectural views of QoS building blocks

5.2 Control-plane mechanisms

Key role of control-plane mechanisms is for controlling the network service response and traffic. There are three mechanisms as following.

Admission control

This is a mechanism control the traffic to be admitted into the network, whether traffic is admitted depends on an a priori service level agreement. The decision on admission can depend on availability of network resources so that newly admitted traffic does not overload the network and degrade service to ongoing traffic. Service provider expects to admit maximal traffic while the same level of QoS is maintained.

This mechanism can also be used for service reliability/availability requested as a priority level for admission control over a specified period as in the SLA. Admission control priority is a way of giving preference to admit higher priority of traffic streams (e.g., for emergency communications) ahead of lower priority under conditions of congestion.

QoS routing

This is a mechanism providing a capability to select of a path (not the traditional shortest path) satisfying the QoS requirements of a flow. Proper selection of the path becomes an issue as the network size grows depending on the specifics and the number of QoS metrics involved. Thus practical QoS routing schemes consider mainly cases for a single QoS metric (e.g., bandwidth or delay) or for dual QoS metrics (e.g., cost-delay, cost-bandwidth, and bandwidth-delay). QoS routing provides a means to determine only a path possibly proper for the requested performance, not guaranteed. To guarantee performance, it is required to reserve necessary network resources along the path which identified by the QoS routing.

The path selection process involves the knowledge (supported by signalling protocols) of the flow's QoS requirements and characteristics as well as information on the availability of network resources (available bandwidth and delay). The knowledge is typically obtained and distributed with the aid of signaling protocols. There are three strategies for reduction of computation complexity of path selection according to how the state information is maintained and how the search of feasible paths is carried out: source routing, distributed routing, and hierarchical routing. In addition, there are two other strategies according to how multiple QoS metrics are handled: metric ordering and sequential filtering [IETF RFC 2386].

Resource reservation

This is a mechanism set aside required network resources on demand for delivering desired network performance. In general, a reservation request is granted by the admission control when has sufficient resources. A resource reservation depends on network performance requirements including the specific network approach to satisfying them and it is important for service providers to be able to charge for the use of reserved resources. Therefore, resource reservation needs support of authentication, authorization, and accounting and settlement between different service providers.

Resource reservation is typically done with a purpose-designed protocol such as RSVP [IETF RFC 2205]. Functionality for resource reservation can be distributed or centralized. However a major issue is the discrepancy of resource availability between actual and the predicted and care should be given to use the most current information, making the node, link and other resources available for the requesting application.

5.3 Data-plane mechanisms

Key role of data-plane mechanisms deals with operation and management. There are several mechanisms as following.

Queue (or Buffer) management

This is a mechanism deals with packets (waiting for transmission) to store or drop aiming for minimize the steady-state queue size. This mechanism is useful to avoid underutilizing link and the lock-out phenomenon where a single connection or flow monopolizes the queue space. Schemes for queue management differ mainly in the criteria for dropping packets and what packets drop. General common for dropping packets is reaching the maximum size of the queue, that is, the queue is full. The use of multiple queues introduces further variation in the schemes, for example, packets are distributed among the queues.

What packets drop depends on the drop disciplines; "Tail drop" rejects the newly arriving packet (the most common strategy), "Front drop" expense of the packet at the front of the queue (while keeps the newly arriving packet) and "Random drop" keeps the newly arriving packet at the expense of a randomly-selected packet from the queue.

Congestion avoidance

This is a mechanism for keeping the load of the network under its capacity thus network can operate at an acceptable performance level, not experiencing congestion. Congestion occurs when the traffic exceeds or nears the capacity of the network because of lack of resources (e.g., link bandwidth and buffer space). A sign of congestion, for example, is that the router (or switch) queues are always full and routers start dropping packets. Packet dropping induces retransmission resulting in more traffic and worsens congestion. The chain reaction could grind the network to a halt with zero throughputs.

Reducing the amount of entering traffic to the network is a typical way for congestion avoidance upon an indication of network congestion (or about to occur). Normally packet loss or timer expiration is regarded as an implicit indication of network congestion unless use of explicit indication such as ECN (Explicit Congestion Notification for IP and TCP: IETF RFC 3168).

Queuing and scheduling

This is a mechanism to control which packets to select for transmission on an outgoing link. Incoming traffic is held in a queuing system, which is made of multiple queues and a scheduler. Governing the queuing system is the queuing and scheduling discipline it employs. There are several key approaches:

- First-in, first-out queuing: Packets are placed into a single queue and served in the same order as they arrive in the queue;
- Fair queuing (or per-flow or flow-based queuing): Packets are classified into flows and assigned to queues dedicated to respective flows. Queues are then serviced in round robin;
- Priority queuing: Packets are first classified and then placed into different priority queues. Within each of the priority queues, packets are scheduled in first-in, first-out order;
- Weighted fair queuing: Packets are classified into flows and assigned to queues dedicated to respective flows. A queue is assigned a percentage of output bandwidth according to the bandwidth need of the corresponding flow. By distinguishing variable-length packets, this approach also prevents flows with larger packets from being allocated more bandwidth than those with smaller packets;
- Class-based queuing: Packets are classified into various service classes and then assigned to queues assigned to the service classes, respectively. Each queue can be assigned a different percentage of the output bandwidth and is serviced in round robin. Empty queues are skipped.

Packet marking

This is a mechanism for marking of packets, typically performed by an edge node according to the specific service classes that they will receive in the network on a per-packet basis. Packet marking involves assigning a value to a designated header field of a packet in a standard way. (For example, the type of service in the IP header or the EXP bits of the MPLS shim header [IETF RFC 3032] is used to codify externally observable behaviours of routers in the DiffServ [IETF RFC 2474] or MPLS-DiffServ [IETF RFC 3270] approach.) If done by a host, the mark should be checked and may be changed when necessary by an edge node. Sometimes, special values may be used to mark non-conformant packets, which may be dropped later due to congestion. Whether done by a host or an edge node, the criteria for packet marking need to be provisioned or configured dynamically.

Traffic classification

This mechanism is for classification of traffic at the flow or packet level. At the edge of the network, the entity responsible for traffic classification typically looks at multi-fields (such as the five tuples associated with an IP flow) of a packet and determines the aggregate to which the packet belongs and the respective service level agreement.

Traffic policing

This mechanism deals with policing for the determination of whether the traffic being presented is on a hop-by-hop basis compliant with pre-negotiated policies or contracts. Typically non-conformant packets are dropped. The senders may be notified of the dropped packets and causes determined and future compliance enforced by SLAs.

Traffic shaping

This is a mechanism for controlling the rate and volume of traffic entering the network, for example, buffers non-conformant packets until it brings the respective aggregate in compliance with the traffic. Thus the resulted traffic is not as bursty as the original and is more predictable. Shaping often needs to be performed between the egress and ingress nodes. There are two key methods for traffic shaping: leaky bucket and token bucket.

The leaky bucket employs a leaky bucket to regulate the rate of the traffic leaving a node. Regardless of the rate of the inflow, the leaky bucket keeps the outflow at a constant rate. Any excessive packets overflowing the bucket are discarded. The token bucket allows packets to go out as fast as they come in provided that there are enough tokens which are generated at a certain rate and deposited into the token bucket till it is full. At the expense of a token, certain volume of traffic is allowed to leave the node. No packets can be transmitted if there are no tokens in the bucket. Yet multiple tokens can be consumed at once to allow bursts to go through. The leaky and token bucket methods can be used together. In particular, traffic can be shaped first with the token bucket method and then the leaky bucket method to remove the unwanted busts.

5.4 Management-plane mechanisms

Key role of management-plane mechanisms deals with the user traffic directly. There are four mechanisms as following.

Service level agreement

This mechanism deals with the agreement between a customer and a provider of a service that specifies the level of availability, serviceability, performance, operation or other attributes of the service, called SLA.

Traffic metering

This mechanism deals with monitoring the temporal properties (e.g., rate) of a traffic stream against the agreed traffic profile. This mechanism involves observing traffic characteristics at a given network point, collecting and storing the traffic information for analysis and further action. Depending on the conformance level, a meter can invoke necessary treatment (e.g., dropping or shaping) for the packet stream.

Traffic restoration

This mechanism deals with the mitigating response from a network under failure conditions. This should be considered at multiple layers, for example, optical networks with dynamic ring and mesh protection including restoration functionality at the wavelength level and SONET/SDH layer with Automatic Protection Switching (APS) as well as self-healing ring and mesh architectures. As in the case of admission control, certain traffic streams related to critical services may require higher restoration priority than others. Thus a service provider needs to plan for adequate levels of spare resources such that QoS SLAs are in compliance under conditions of restoration. Typical parameters for measuring service restorability are time-to-restore and the percentage of service restorability.

Policy

This mechanism deals with policies for administering, managing and controlling access to network resources. This mechanism can be specific to the needs of the service provider or reflect the agreement between the customer and service provider, which may include reliability and availability requirements over a period of time and other QoS requirements. Some examples of the mechanism are policy routing (directing packet flow to a destination port without a routing table), packet filtering policies (marking or dropping packets based on a classifier policy), packet logging (allowing users to log specified packet flows) and security-related policies.

6 Measurement and Assessment

6.1 QoS Measurement

Considerations for QoS measurement [b-ITU-T E.800-Sup.8]

In general, the QoS measurement methodology, protocol including reporting should be capable of estimating at least the set of QoS metrics of packets transmitted between specified measurement points. And QoS measurement should be possible on-demand or on a periodic, ongoing basis.

It is important to know that the characteristics of connectionless service such as IP and NGN that delivers user payloads in the form of packets/bytes in each direction. In this case, outbound and inbound traffic routes may differ, so the targets and measurements for all QoS attributes are practically one-way to reflect the connectionless nature of the service. Thus measurements should also be made one-way. Because this raises some practical challenges (e.g., clock synchronization), there may be occasions where two-way measurements will be made (and one-way metrics may be estimated from the two-way measurements). If this is the case, it should be noted and reported.

Measurement probe packets should traverse as much as possible the same path as customer packets having the same QoS service class and the same QoS mechanisms in routers along the path, implying value of probe packets should be appropriately set for the QoS class to be measured.

The measurement methodology should not require that providers provide access to measurement points nor exchange measurement data. However, the protocols should support access to measurement points or measurement data between consenting providers for authorized requestors. It should ideally be possible to make PE-PE: Provider Edge router) or CE -CE measurements (CE: Customer Edge router), even when the PEs or CEs are contained in, or attached to, the networks of different providers. The measurement methodology should specify how the errors in measurements are treated, and how results are processed in terms of any statistical treatment of data. Finally, the measurement methods and protocol must provide means to limit and detect attempts to tamper with or alter the QoS metric estimates.

Measurement of QoS along the path between end-to-end customers should be the essential part for the monitoring and troubleshooting of SLAs.

For end-end QoS measurement which may involve several different providers, it is required to identify the inter-provider measurement. Therefore it is required that measurements made across the networks of multiple providers could be compared and combined to create meaningful and reasonably accurate end-to-end measurements. To facilitate of this, some agreement on common approaches to measurement with co-operating providers who involved in that end-end path will simplify the tasks of service monitoring and troubleshooting. ITU-T introduces objectives, considerations, methodology and protocols for QoS measurement in the case of inter-provider through ITU-T Recommendation Y.1543 [13] and supplement 8 to ITU-T E.800 series of recommendations [b-ITU-T E.800-Sup.8].

A provider may also designate an MPoP (Measurement Point of Presence) as a location that has specific capabilities for measurement. In these cases, service providers should agree on the volume of the test traffic that they will generate into each others' networks. Service providers should publish enough information about the location of measurement devices that are available for customers and/or other service providers to enable customers or other service providers to make rational choices of where to direct their measurement traffic. Co-operating providers should agree on the clock accuracy they will support. In order to support diagnostics and service conformance tracking, each provider should retain QoS measurement data for some agreed-upon period.

End user aspects of QoS including QoE should be a set of QoS and performance measurements as shown in Figure 11. [2] Therefore the measurements will be taken from each of the segments of the measurement network model and may be combined to form multi-segment, site-to-site, edge-to-edge or terminal-to-terminal metrics. A subset of these metrics will be used for reports for the offered services.

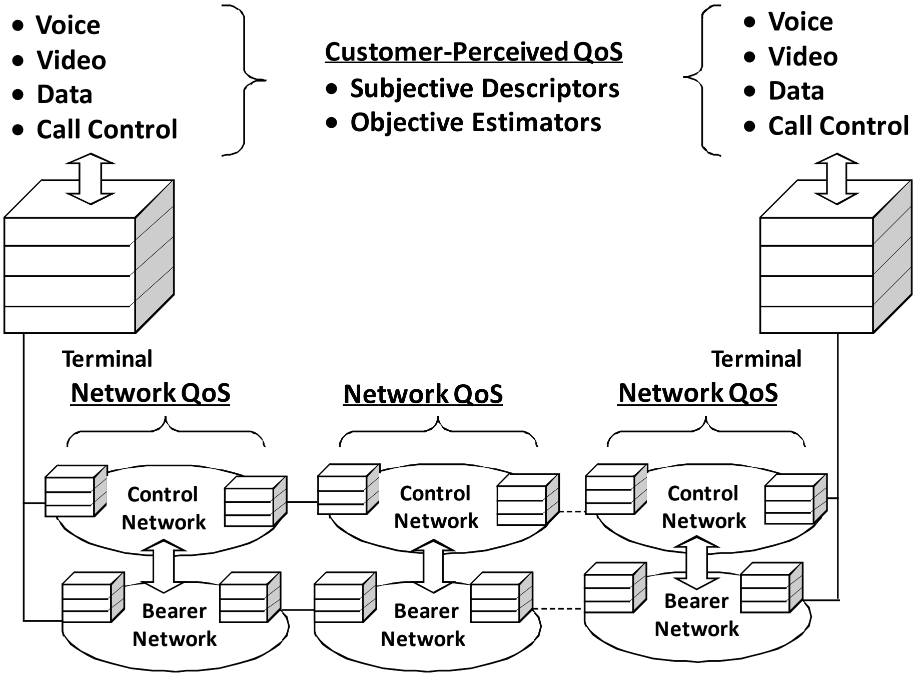


Figure 11 – Customer perceived QoS

QoS measurement objectives

Enhancement of QoS is contributed to increase the level of confidence in the expected service characteristics of the networks. Increased confidence will enable new applications, services and revenue streams. An integral part of achieving this confidence is the continuous measurement of QoS and performance. Thus the objective of QoS measurements is to provide information for customers, potential customers and service providers, and includes [13]:

- For customers and potential customers:
  - Reports to customers of what service has been delivered;
  - Reports to potential customers to support marketing claims on service characteristics.
- For service providers and third party delivery assurance entities:
  - Reports to design service offerings;
  - Reports for troubleshooting;
  - Data for marketing collateral;
  - Reports to enable capacity planning and service development.

Therefore the QoS measurement system and the statistics should provide followings:

- be well defined (non-ambiguous) and be easily understood by SPs and customers;
- be relevant to customers' applications;
- enable service providers to diagnose issues and anticipate capacity requirements;
- be independently repeatable (multiple SP measurements over the same time get the same result);
- be independently verifiable by customers (customer measurements should be close to SP estimates);

- be widely applicable (traffic type, link size, load independent, any IP network);
- be appropriately sensitive to distance and path;
- not significantly impact the forwarding of other data;
- be sufficiently scalable to support enough (e.g., million) customer sites;
- be sufficiently reliable and accurate to enable SLAs with financial penalties to be administered.

QoS measurement methodologies

If possible, the measurement methodology would be common among providers as much as possible (however, this may not be practical). There are two methodologies for measuring QoS: Passive measurement (using test packet), and Active measurement.[22] It may be noted, however, that any combination of these techniques may be used in a particular performance measurement tool.

Passive measurement:

As a general method, test packets are sent from management systems, and performance metrics such as delay, jitter, and packet loss are measured along the way. The results of these measurements are used as a proxy for the performance of real traffic. This method is also often used for troubleshooting. While this is a very simple technique, care needs to be taken when interpreting the results. It is not desirable to send test packets during busy hours since this will unnecessarily load the network with management traffic. On the other hand, unless testing is performed during peak usage hours, the measurements will not truly reflect user experience at the most important time

Active measurement:

In this method, probes in the form of software agents or network appliances are deployed on network elements and user devices (for the software agent case). Measurements based on these probes provide a very accurate status of the devices at any time. Furthermore, in the case of software agents, true user experience can be measured unobtrusively since measurements are obtained directly from user devices. The main drawback of this measurement is that it doesn’t scale for large networks. While it is very useful for fault isolation and root cause analysis, this measurement cannot be used for monitoring large networks with millions of user devices and network elements.

The sources and sinks of probes may be either dedicated measurement devices, routers that are dedicated to measurement tasks or routers that support both data traffic and measurement probes.

The measurements may be reported as point-to-point measurements between two measurement points or a matrix of such measurements among various points. It is also possible to report average measurements or other statistics computed over a number of different point-to-point measurements. To enable measurement of QoS parameters across multiple provider networks, one of the following methods could be used:

- Each provider agrees to use a common measurement protocol and to make probe points available to other providers, enabling measurements to be made along the end-to-end path;
- Each provider network uses its own methods and probe devices to collect measurements on a per-provider basis, with these measurements being combined to estimate the concatenated end-to-end performance. It is noted that this requires co-operation among interconnected service providers in terms of the protocol and availability of probe points to measure the QoS parameters of the inter-provider links.

QoS measurement protocols

There are various protocols developed by standard organizations, especially IETF to use QoS measurement including vendor-proprietary measurement protocols used by some providers and end customers. ICMP (Internet Control Message Protocol)-based PING measurement of TWPD (Two-Way Packet Delay), TWPL (Two-Way Packet Loss), and instantaneous bidirectional connectivity have historically been used by a number of providers when monitoring networks to deliver QoS-oriented SLAs.

The IPPM (IP Performance Metrics) protocol OWAMP (One-Way Active Measurement Protocol) [IETF RFC 4656] (or a protocol compatible with it) should be used for one-way measurements, with TWAMP (Two-Way Active Measurement Protocol) [IETF RFC 5357] as an alternative if two-way measurements are to be used. (Note all measurements should be one-way but measurements may be two-way as long as the distinction is reported.)

The use of ICMP-based PING as a measurement protocol is not recommended as a reliable protocol for measurement of customer IP path performance. IP network elements often treat ICMP messages quite differently to end-customer traffic, particularly under higher network traffic conditions. Other lower layer OAM protocols, such as Ethernet OAM, with its performance measurement parameters defined in [ITU-T Y.1731], may be suitable for deducing IP delay and loss performance where Ethernet maintenance entities exist at or near representative IP network segment measurement points. Care needs to be taken to ensure that the measured performance of Ethernet OAM frames is truly representative of IP customer traffic performance over the same network path. Use of Ethernet OAM may be a valid protocol across Ethernet access segments where achieving a large enough sample of paths makes deployment of dedicated IP measurement probes otherwise uneconomic.

6.2 Basic network model for measurement

Ideally, measurements would be taken between the same endpoints as each customer's traffic. Whether these endpoints are the customer's terminal (TE), customer edge router (CE) or provider edge router (PE), the number of measurements would be so great as to make this impracticable. Therefore, as for a practical solution, it is required that segmenting the network into a measurement network model. The greater the leverage of a single measurement produced by a segment probe, the fewer probes will be needed. If fewer segment measurements may be used in the calculations of thousands of concatenated estimates, then there will be lower total probe overhead.

Providers offer assured delivery services between different endpoints as following cases:

- edge-edge: extend to the edge of a providers' network;
- site-site: extend to the edge of a customer's premises (also called end-to-end);
- TE-TE for a managed customer network service: extending to a customer's terminal.

The network model is partitioned into segments, each being monitored independently and should be support for these three cases of services. Typically, the network is considered to consist of ingress and egress access segments, and a transit segment. It is assumed that one regional service provider will provide an access network that supports both ingress and egress segments for a specific site. There may be a backbone service provider(s) providing transit services between the regional service providers. A specific service provider may act as either or both an access provider for some traffic and as a transit provider for some traffic. Taking into account all these considerations, following Figure 12 shows a basic network model for QoS measurements.

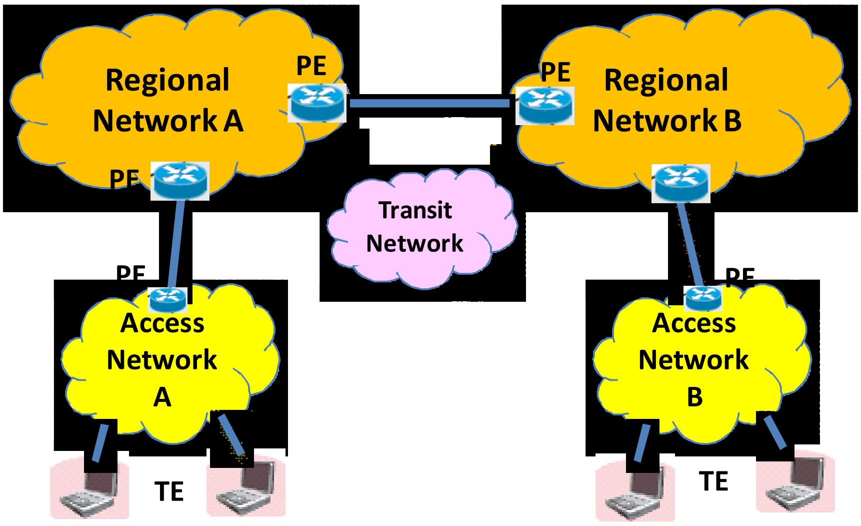


Figure 12 – Basic network model

In addition to this, Figure 13 shows a network model of the access portion of the IP network. In the downstream direction, from the core to the customer premises, a series of network elements and wires are connected: edge router, DSLAM (or OLT for GPON), DSL modem (or ONT for GPON), firewall, and router. This model is bidirectional, so upstream traffic traverses the same elements in reverse order.

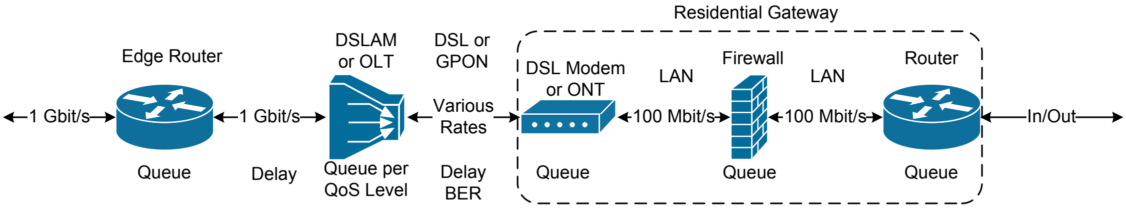


Figure 13 – Access network portion

6.3 Quality Assessment methodology

There are two methodologies for quality assessment: Subjective and Objective assessment. Subjective quality assessment is a method that the quality of audio and visual media should be evaluated in subjective terms. It is a psychoacoustic/visual experiment, which is the most fundamental and reliable way to quantify users' QoE. However, subjective quality assessment, in which human subjects evaluate the quality of various testing conditions, is time-consuming and expensive. In addition, special assessment facilities such as professional audio-visual devices and soundproof chambers are required. Therefore, it would be desirable to develop a way to estimate subjective quality solely from the physical characteristics of a system under test. This is called objective quality assessment. Thus objective quality assessment is defined as a means for estimating subjective quality solely from objective quality measurement or indices. [3]

In principle, assessment of QoE must be performed using subjective tests with metrics such as mean opinion score (MOS). However, it is also possible and sometimes more convenient to estimate QoE based on objective testing and associated quality estimation models. For evaluation of QoE, the most important thing is to identify the objective parameters affecting QoE which should be measured and calculated through different quality estimation models. While subjective testing needs more resources and efforts (because it requires human subjects), objective measurement and automatic calculation using appropriate quality estimation models is generally much faster and cheaper. [b- ITU-T G.1011]

There are three modes of objective test for evaluation of QoE: intrusive mode, non-intrusive mode, or planning mode. "Intrusive mode" injects a signal into the system under test in order to generate a degraded output signal, so the channel must be taken out of service for normal traffic. Conversely, for "non-intrusive mode", the quality assessment system can be used whilst live traffic is carried by the channel, without the need for any active test signals. "Planning mode" is not used in a real-time environment, but as a tool for the design of systems, and hence does not require any real-time inputs. [b- ITU-T G.1011]

7 QoS/QoE in IP-based platform and NGN

7.1 IP service performance model

Services in telecommunication including IP-based platform and NGN should be performed in two ways in general: vertical and horizontal aspects. Understanding the performance features of these two different aspects through the model is important for identifying the service performance. ITU-T recommendation Y.1540 deals with these issues.

Layered (Vertical) model

Layered model of IP service performance is simple, benefiting from the key feature of IP called “everything over IP” and “IP over everything.” As shown in Figure 14, the layered model consists with three layers: lower layers, IP layer and higher layers. Accordingly the performance provided to IP service users depends on the performance of other layers:

- Lower layers: provide (via "links") connection-oriented or connectionless transport supporting the IP layer. Links are terminated at points where IP packets are forwarded (i.e., "routers", "source host" and "destination host") and thus have no end-to-end significance. Links may involve different types of technologies, for example, ATM, frame relay, SDH, PDH, ISDN and leased lines;
- The IP layer: provides connectionless transport of IP and has end-to-end significance for a given pair of source and destination IP addresses;
- Higher layers: supported by IP, that further enables end-to-end communications, using by, for example, TCP, UDP, FTP, RTP and HTTP. The higher layers will modify and may enhance the end-to-end performance provided at the IP layer.

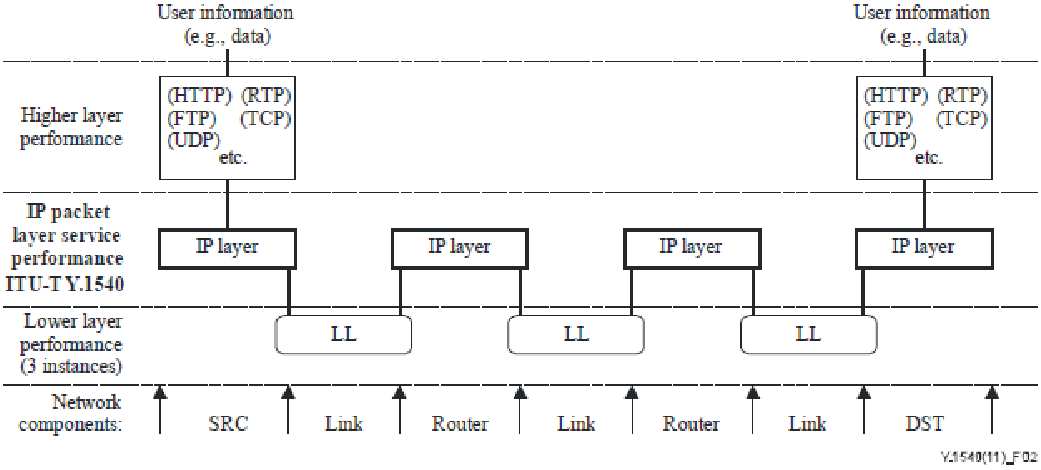


Figure 14 – Layered model for IP service performance

Generic (Horizontal) model

A generic model for IP service performance deals with horizontal configuration of IP-based networks and is primarily composed of two types of sections: the exchange link and the network section. Each of the performance parameters can be applied to the unidirectional transfer of IP packets on a section or a concatenated set of sections.

Figure 15 shows an model of IP network components and their features are following:

- Host: A computer that communicates using the IP. A host implements routing functions and may implement additional functions including higher layer protocols and lower layer protocols;
- Router: A host that enables communication between other hosts by forwarding IP packets based on the content of their IP destination address field;
- Source host (SRC): A host and a complete IP address where end-to-end IP packets originate. In general, a host may have more than one IP address; however, a source host is a unique association with a single IP address. Source hosts also originate higher layer protocols (e.g., TCP) when such protocols are implemented;
- Destination host (DST): A host and a complete IP address where end-to-end IP packets are terminated. In general, a host may have more than one IP address; however, a destination host is a unique association with a single IP address. Destination hosts also terminate higher layer protocols (e.g., TCP) when such protocols are implemented;
- Link: A point-to-point (physical or virtual) connection used for transporting IP packets between a pair of hosts. It operates below the IP layer.

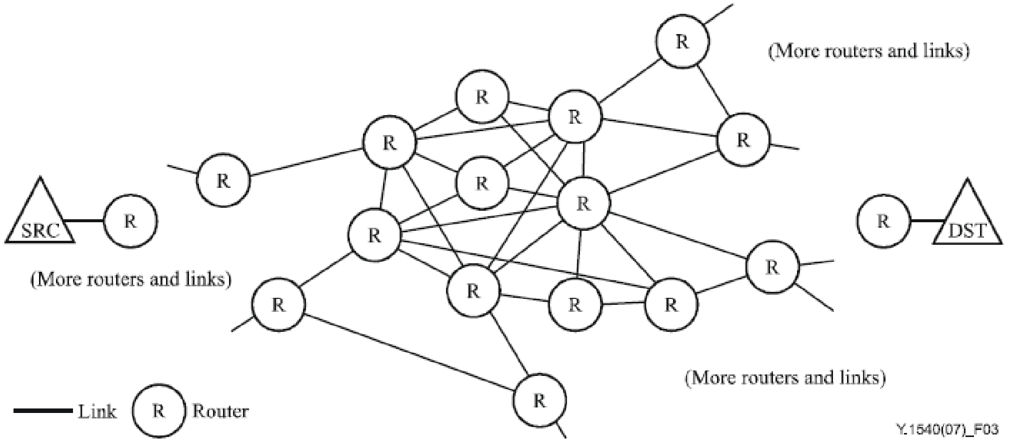


Figure 15 – IP network components

Figure 16 illustrates the network connectivity relevant to IP service between a SRC and a DST. At the edges of each NS (Network Section), gateway routers receive and send packets across exchange links. Features of each component are following:

- Exchange link (EL): The link connecting:
  - 1) a source or destination host to its adjacent host (e.g., router) possibly in another jurisdiction, sometimes referred to as an access link, ingress link or egress link; or
  - 2) a router in one network section with a router in another network section.
- Network section (NS): A set of hosts together with all of their interconnecting links providing a part of the IP service between a SRC and a DST, and are under a single (or collaborative) jurisdictional responsibility. Source NS and destination NS are particular cases of network sections. Pairs of network sections are connected by exchange links or via Transit NS. For the purpose of IP performance allocation, it will be relevant to focus on the set of hosts and links under a single (or collaborative) jurisdictional responsibility (such as an ISP or an NSP).
  - Source NS (A in Figure 16): The NS that includes the SRC within its jurisdictional responsibility. In some cases, the SRC is the only host within the source NS.

- Destination NS (G in Figure 16): The NS that includes the DST within its jurisdictional responsibility. In some cases, the DST is the only host within the destination NS.
- Transit NS (C in Figure 16): The NS that included the hosts for transit of traffics between SRC and DST within its jurisdictional responsibility.

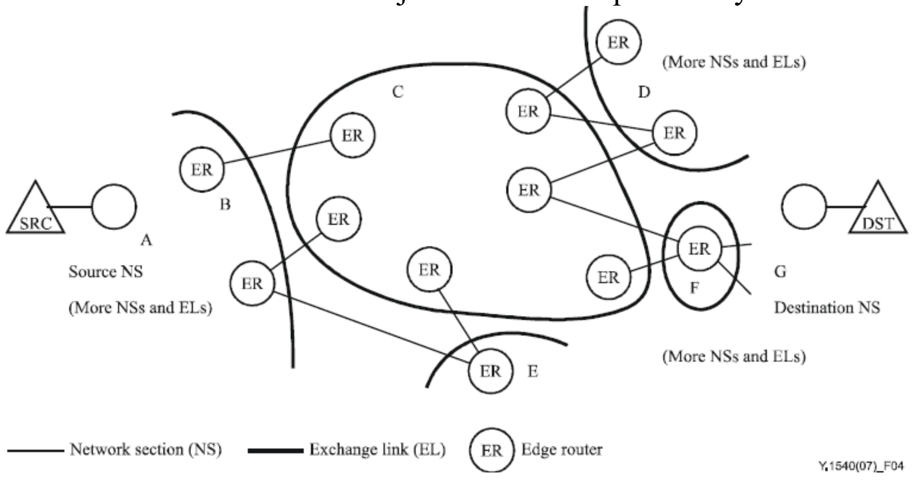


Figure 16 – IP network connectivity

7.2 IP packet transfer performance parameters

A set of IP packet information transfer performance parameters using the IP packet transfer are identified in ITU-T Recommendation Y.1540 and summarized following:

- ● Packet qualifications: This uses for qualifying the applicability of performance parameters to sets of packets. There are two parameters as following:
  - Populations of interest: the total set of packets being sent from SRC to DST, for the end-to-end case;
  - Packet flow: the most common example of a population of interest and the set of packets associated with a given connection or connectionless stream having the same source host address (SRC), destination host address (DST), class of service, and session identification (e.g., port numbers from a higher-layer protocol).
- ● IP packet transfer delay (IPTD): this is the time, (t2 – t1) between the occurrence of two corresponding IP packet reference events, ingress event IPRE1 at time t1 and egress event IPRE2 at time t2, where (t2 > t1) and (t2 – t1) ≤Tmax. IPTD is defined for all successful and errored packet outcomes across a basic section or an NSE:
  - Mean IP packet transfer delay: the arithmetic average of IP packet transfer delays for a population of interest;
  - Minimum IP packet transfer delay: the smallest value of IP packet transfer delay among all IP packet transfer delays of a population of interest including propagation delay and queuing delays common to all packets;
  - Median IP packet transfer delay: the 50th percentile of the frequency distribution of IP packet transfer delays from a population of interest. The median is the middle value once the transfer delays have been rank-ordered. To obtain this middle value when the population contains an even number of values, then the mean of the two central values is used;
  - End-to-end 2-point IP packet delay variation (PDV): Streaming applications might use information about the total range of IP delay variation to avoid buffer underflow and

overflow. Extreme variations in IP delay will cause TCP retransmission timer thresholds to grow and may also cause packet retransmissions to be delayed or cause packets to be retransmitted unnecessarily. End-to-end 2-point IP PDV is defined based on the observations of corresponding IP packet arrivals at ingress and egress MP.

- IP packet error ratio (IPER): IP packet error ratio is the ratio of total errored IP packet outcomes to the total of successful IP packet transfer outcomes plus errored IP packet outcomes in a population of interest.
- IP packet loss ratio (IPLR): the ratio of total lost IP packet outcomes to total transmitted IP packets in a population of interest;
- Spurious IP packet rate: Spurious IP packet rate at an egress MP is the total number of spurious IP packets observed at that egress MP during a specified time interval divided by the time interval duration (equivalently, the number of spurious IP packets per service-second);
- IP packet reordered ratio (IPRR): the ratio of the total reordered packet outcomes to the total of successful IP packet transfer outcomes in a population of interest;
- IP packet severe loss block ratio (IPSLBR): the ratio of the IP packet severe loss block outcomes to total blocks in a population of interest;
- IP packet duplicate ratio (IPDR): the ratio of total duplicate IP packet outcomes to the total of successful IP packet transfer outcomes minus the duplicate IP packet outcomes in a population of interest;
- Replicated IP packet ratio (RIPR): the ratio of total replicated IP packet outcomes to the total of successful IP packet transfer outcomes minus the duplicate IP packet outcomes in a population of interest;
- Stream repair parameters: the probability that a given packet interval (or information block, b) will contain more than x impaired packets;
  - IP packet impaired interval ratio (IPIIR): the ratio of the IP packet impaired interval outcomes to total non-overlapping intervals in a population of interest;
  - IP packet impaired block ratio (IPIBR): the ratio of the IP packet impaired block outcomes to total non-overlapping.
- Capacity parameters: An end-to-end IP packet transfer service traverses an ordered sequence of basic sections from a source host, to a destination host. The capacity parameters define properties for basic sections in terms of their ability to carry IP traffic, and corresponding properties for network section ensembles (NSE), also referred to as "paths". It is important to note that a basic section as well as a sequence of basic sections is associated with a direction. The direction is significant, as the properties of a sequence of sections in the forward direction need not be the same as in the reverse direction.
- Flow-related parameters: characterize performance in terms of flow or throughput-related parameters that evaluate the ability of IP networks or sections to carry quantities of IP packets. It should be noted that a parameter intended to characterize the throughput of an IP application would not be equal to the amount of resources available to that application, because the higher layer protocols over IP (e.g., TCP) also influence the throughput experienced.

7.3 Measurement network model

To perform measurement of quality for IP based platforms including NGNs, it is required to set up proper models to apply possible cases. For this purpose, a basic network model for measurement introduced in Figure 12 would be based. Measurement network models should be cover following three cases which are offered by the providers to assure delivery services between different endpoints. [13]

- edge-edge: extend to the edge of a providers' network;

- site-site: extend to the edge of a customer's premises (also called end-to-end);
- TE-TE for a managed customer network service: extending to a customer's terminal.

Edge-Edge Model

Figure 17 shows an edge-edge model which service delivery is assured to a PE nearest a customer. In this model, service between customer terminals or CE to the PE is not assured. The assured performance characteristics of the network are comprised of the aggregate of the performance characteristics of the ingress, transit and egress segments.

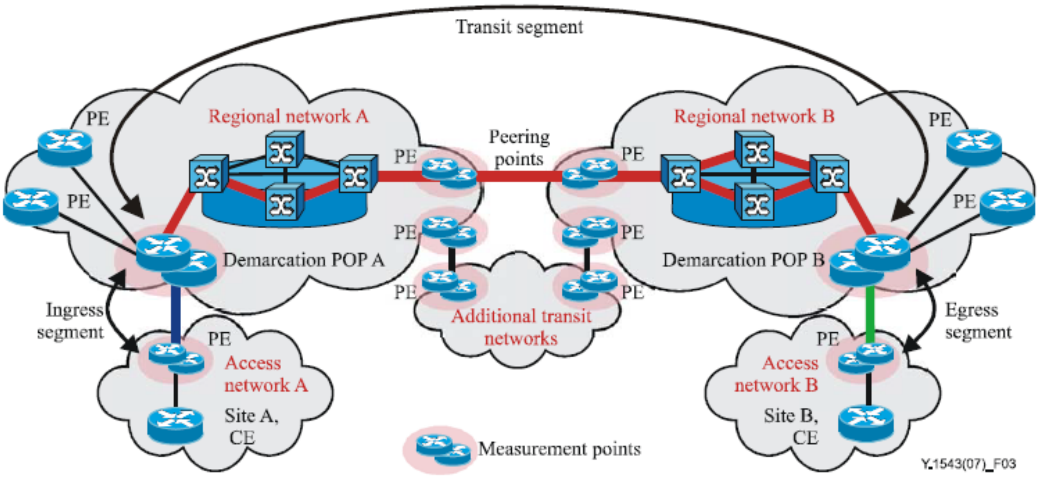


Figure 17 – Edge-Edge model

The ingress and egress segments do include the provider edge router as well as regional switching and transport, but do not include the CE-PE link. The transit segment is measured from demarcation POP of the ingress regional service provider to the demarcation POP of the egress regional service provider. Geometrically, the transit segment may span a city, country/state, continent or multiple continents therefore this segment may or may not include separate backbone service providers according to the coverage. The transit segment may include parts of the ingress and egress regional networks, interconnects between the regional and backbone providers, and transit service across any backbone networks. In this case, the transit service of the backbone network should be form of a sub-segment of the entire transit service. Partitioning of measurement responsibility may follow network boundaries however, measurement responsibilities may cross boundaries in any configuration to complete measurement coverage (e.g., two or more networks may be covered by a single measurement system's measurement points.)

Inter-domain QoS relies on the ability to collect inter-service provider statistics on a continuous basis and for service providers to resolve the causes of performance targets not being met. Service providers should support followings to support this monitoring and troubleshooting requirement:

- Each participating provider must provide measurement points that act as performance characteristic test points for their use, and possibly for restricted use by other SPs;
- Measurement points must be located at any participating service providers' major interconnection peering POP;
- Regional providers should provide a measurement point (demarcation POP) supporting for each participating customer site;
- A service-dependent measurement point at PE, and possibly at CE and/or customer TE if this scope of performance assurance is supported.

Site-Site Model

Figure 18 shows a site-site model which delivery is assured to customer CE.

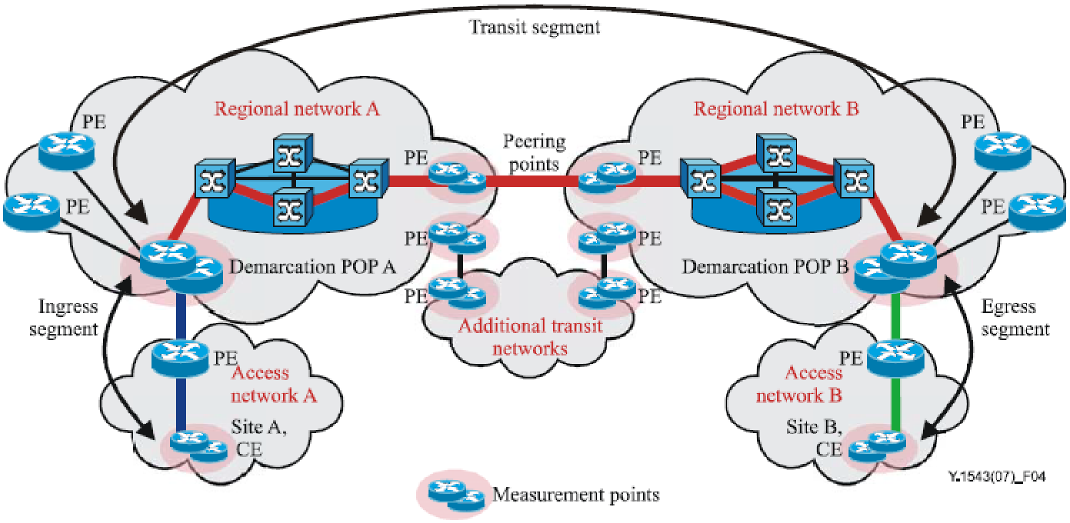


Figure 18 – Site-Site model

In the site-site model, service between customer terminals to the CE is not assured by the service provider and this is the responsibility of the customer. The assured performance characteristics of the network are comprised of the aggregate of the performance characteristics of the ingress, transit and egress segments.

The ingress and egress segments include an access segment (DSL, cable, SONET/SDH, Ethernet, etc.) including the customers edge (CE) router as well as regional switching and transport.

TE-TE Model

Figure 19 shows a TE-TE model which the assured performance characteristics of the network are comprised of the aggregate of the performance characteristics of the ingress, transit egress and customer segments. In this model, the customer segment includes the network between a CE and a customer's TE and this may include home networking arrangements to company LANs, computers and appliances.

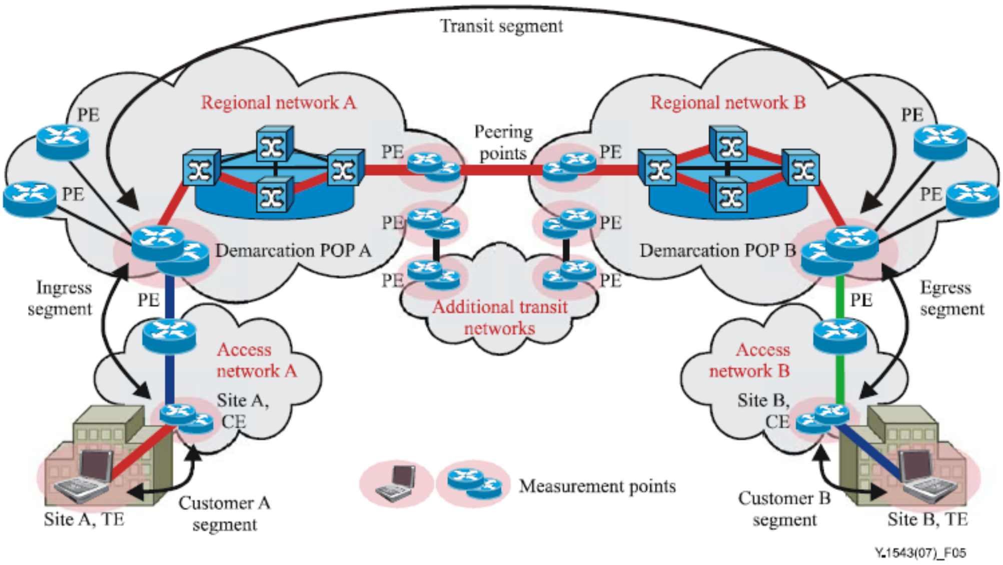


Figure 19 – TE-TE model

Selection of the customer's TEs to be used for measurements including following considerations:

- Stability:
  - static address or directory lookup;
  - stationary rather than mobile;
  - always online.
- Performance: probe response not impacted by other programs;
- Clock synchronization: required for one-way delay and delay percentile measurements;
- Representativeness of many other TEs: analysis or measurement may show that measurements between a CE and a particular TE is representative of many other TEs, called "landmark" TEs;
- Number of TEs probed:
  - to minimize the number of probes, a minimum number of landmark TEs should be used;
  - to minimize the complexity of data handling and reporting, a minimum number of landmark TEs should be used.

Communication from a CE to a TE may require NAT traversal. Therefore pre-provisioning or NAT traversal protocols may need to be used. Alternatively, the NAT device may be used as a measurement point as a proxy for TEs. There will be cases when there will be very little performance variation in the customer's network, in these cases, instead of the use of operating measurements, fixed impairment values may be used.

7.4 Quality control in NGN

NGN uses IP as a key transport technology, in this sense, NGN is one of IP based platform. However NGN has been adopted many of new features which were not supported by the traditional IP based network such as Internet. Those new features are addressed on QoS, Mobility management and Security.

Features and functions of NGN

ITU-T identified a definition of the NGN through ITU-T Recommendation Y.2001 as following: “A packet-based network able to provide telecommunication services and able to make use of multiple broadband, QoS-enabled transport technologies and in which service-related functions are independent from underlying transport-related technologies. It enables unfettered access for users to networks and to competing service providers and/or services of their choice. It supports generalized mobility which will allow consistent and ubiquitous provision of services to users.”

NGN definition clearly indicated that the NGN should be a packet-based network over the broadband infrastructures (both over fixed and mobile) with separation between service and transport. Because of these given nature of the NGN, fundamental characteristics of the NGN are summarized as following by the ITU-T Recommendation Y.2001:

- packet-based transfer;
- separation of control functions among bearer capabilities, call/session, and application/ service;
- decoupling of service provision from transport, and provision of open interfaces;
- support for a wide range of services, applications and mechanisms based on service building blocks (including real time/ streaming/ non-real time and multimedia services);
- broadband capabilities with end-to-end QoS;
- interworking with legacy networks via open interfaces;
- generalized mobility;
- unfettered access by users to different service providers;
- a variety of identification schemes;
- unified service characteristics for the same service as perceived by the user;
- converged services between fixed/mobile;
- independence of service-related functions from underlying transport technologies;

- support of multiple last mile technologies, and;
- compliant with all regulatory requirements, for example concerning emergency communications, security, privacy, lawful interception, etc.

Various aspects of requirements for the NGN have been identified following the characteristics of the NGN. Most significant points of characteristics having great impacts to the NGN requirements are followings:

- “decoupling of service provision from transport” which means separation of service functions from the underline transport functions;
- “packet-based transfer but support for a wide range of services, applications and mechanisms based on service building blocks (including real time/streaming/non-real time and multimedia services)” which means providing integrated services using packet based transport means;
- “broadband capabilities with end-to-end QoS” which requires support of QoS from one end to other end;
- “generalized mobility and converged services between fixed and mobile” which available services crossover the of fixed and/or mobile access environments in any directions.

One of the biggest challenges of the NGN is the separation between services from underline transport technologies. The reference architecture of the NGN is shown in Figure 20 (ITU-T Recommendation Y.2011) [b- ITU-T Y.2011].

In general, any and all types of network technologies may be deployed in the transport stratum indicated as "NGN transport", including connection-oriented circuit-switched (CO-CS), connection-oriented packet-switched (CO-PS) and connectionless packet-switched (CLPS) layer technologies according to ITU-T Recommendations G.805 and G.809. Until today it is considered that IP is the preferred transport protocol used to support NGN services as well as supporting legacy services. The "NGN services" provide the user services, such as a telephone service even and Web services and others. Therefore "NGN service" may involve a complex set of geographically distributed services platforms or in the simple case just the service functions in two end-user sites.

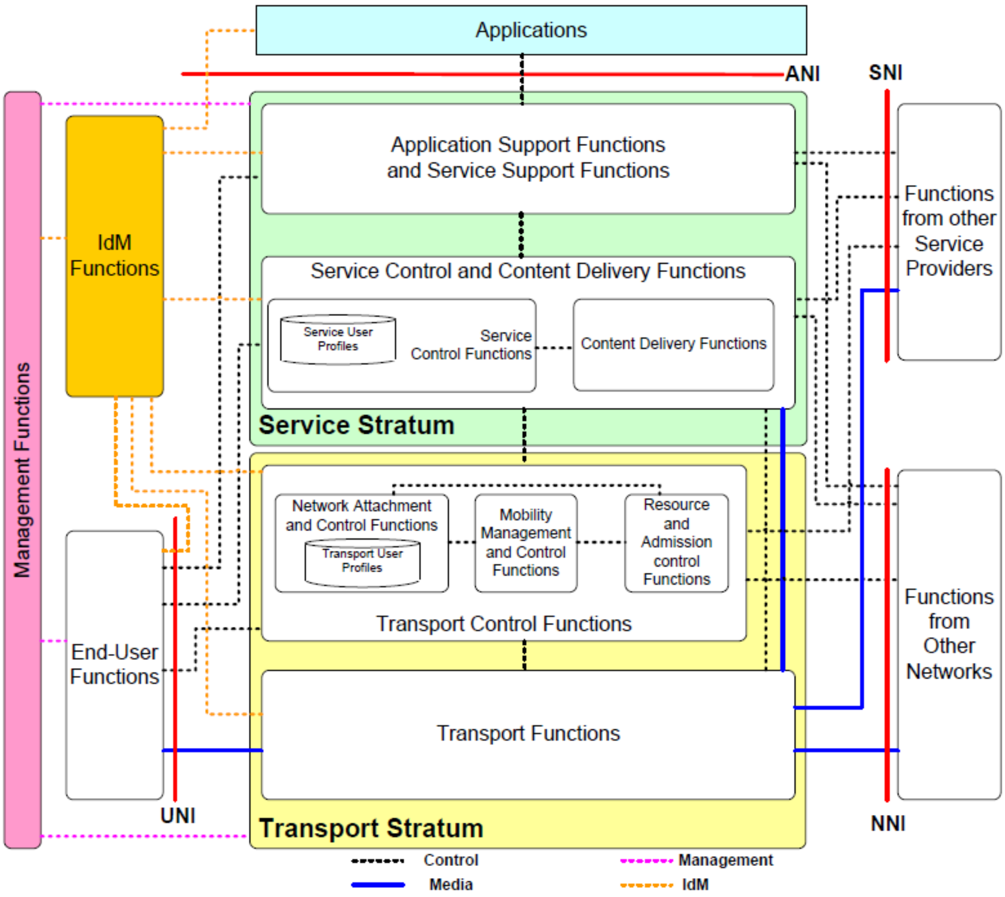

Applications
ANI
SNI
Management Functions
IdM Functions
Application Support Functions and Service Support Functions
Service Control and Content Delivery Functions
Service User Profiles
Service Control Functions
Content Delivery Functions
Service Stratum
Functions from other Service Providers
Network Attachment and Control Functions
Transport User Profiles
Mobility Management and Control Functions
Resource and Admission control Functions
Transport Control Functions
End-User Functions
Transport Functions
Functions from Other Networks
UNI
Transport Stratum
NNI
Control
Media
Management
IdM

Figure 20 – NGN Reference Architecture

Transport Stratum Functions

The transport stratum of NGN provides IP connectivity for both end-user equipment outside the NGN and controllers and enablers, which usually reside on the servers inside the NGN. The transport stratum is responsible for providing end-to-end QoS and is divided into access networks and the core network, with a function linking the two transport network portions.

- Transport functions: provide the connectivity for all components and physically separated functions within the NGN. These functions provide support for the transfer of media information, as well as the transfer of control and management information. Transport functions include access network functions, edge functions, core transport functions, and gateway functions;
- Transport control functions: The transport control functions include Resource and Admission Control Functions, Network Attachment Control Functions and Mobility management and Control Functions;
  - Network attachment control functions (NACF): provide registration at the access level and initialization of end-user functions for accessing NGN services. The functions provide network level identification/authentication, manage the IP address space of the access network, and authenticate access sessions. The functions also announce the contact point of the NGN Service/Application functions to the end user. That is, the functions assist end-user equipment to register and start the use of the NGN;
  - Resource and Admission Control Functions (RACF): provides QoS control (including resource reservation, admission control and gate control), NAPT and/or FW traversal

control functions over access and core transport networks. Admission control involves checking authorization based on user profiles, SLAs, operator specific policy rules, service priority, and resource availability within access and core transport. The RACF act as the arbitrator for resource negotiation and allocation between Service Control Functions and Transport Functions;

- Transport User Profile functions: These functions take the form of a functional database representing the combination of a user's information and other control data into a single "user profile" function in the transport stratum;
- Mobility Management and Control Functions (MMCF): provide functions for the support of IP based mobility in the transport stratum. These functions allow the support of mobility of a single device.

Service Stratum Functions

The service stratum functions provide session-based and non session-based services including subscribe/notify for presence information and the message method for instant message exchange.

- ● Service control and content delivery functions (SC&CDF): The SC&CDF includes service control functions and content delivery functions
  - Service Control Functions (SCF): The SCF includes resource control, registration, and authentication and authorization functions at the service level for both mediated and non-mediated services. They can also include functions for controlling media resources, i.e., specialized resources and gateways at the service-signalling level;
  - Service user profile functions: The service user profile functions represent the combination of user information and other control data into a single user profile function in the service stratum, in the form of a functional database;
  - Content Delivery Functions (CDF): The CDF receives content from the application support functions and service support functions, store, process, and deliver it to the end-user functions using the capabilities of the transport functions, under control of the service control functions.
- ● Application/Service support functions: The application/service support functions include functions such as the gateway, registration, authentication and authorization functions at the application level. These functions are available to the "Third-Party Applications" and "End-User" functional groups.

End User Functions

No assumptions are made about the diverse end-user interfaces and end-user networks that may be connected to the NGN access network. Different categories of end-user equipment are supported in the NGN, from single-line legacy telephones to complex corporate networks. End-user equipment may be either mobile or fixed.

Management Functions

Support for management is fundamental to the operation of the NGN. These functions provide the ability to manage the NGN in order to provide NGN services with the expected quality, security, and reliability. These functions are allocated in a distributed manner to each functional entity (FE), and they interact with network element (NE) management, network management, and service management FEs. Further details of the management functions, including their division into administrative domains, can be found in ITU-T recommendation M.3060. The accounting management functions also include charging and billing functions (CBF). These interact with each other in the NGN to collect accounting information, in order to provide the NGN service provider with appropriate resource utilization data, enabling the service provider to properly bill the users of the system.

NGN QoS control mechanism

The NGN determines the requirements to support QoS across multiple heterogeneous service providers. Existing standards specify several metrics and measurement methods for point to point performance. Notable are ITU-T Recommendations, Y.1540 and Y.1541 standards and the IETF IP Performance Metrics (IPPM) Working Group standards. The NGN considers the options and parameters left unspecified, taking into account the concatenation of performance over multiple network segments, allocation of impairment budgets, mapping between IP and non-IP metrics, accuracy, and data handling. Thus NGN defines the relationship among individual networks' performance which may be observed at physical interfaces between a specific network and associated terminal equipment, and at physical interfaces between specific networks.

NGN identifies the functional entities to facilitate interworking with the QoS functionality in the core network as well as that specific to each type of access networks. Functional requirements and architecture for resource and admission control in NGN, called RACF are developed to provide high-level requirements, scenarios and functional architecture. The decomposition to functional entities is specified to provide reference points and interfaces for the control of Quality of Service (QoS), Network Address and Port Translator (NAPT) and/or Firewall (FW) traversal are described. Following Figure 21 shows an overall functional configuration model of RACF which defined in ITU-T Recommendation Y.2111.

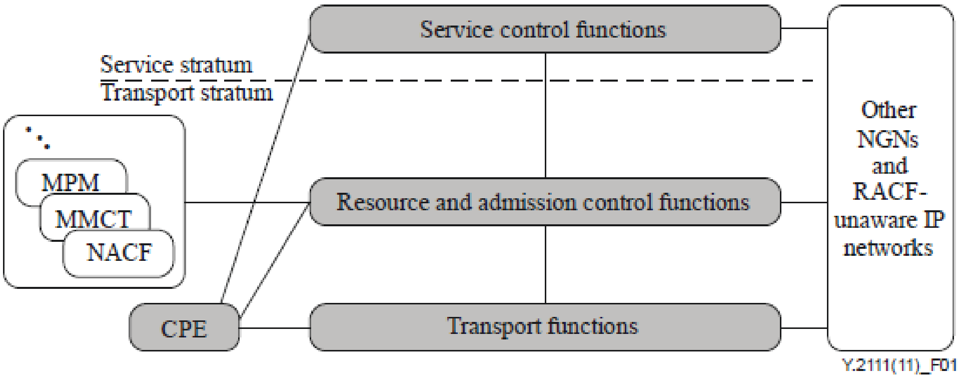

Service control functions
Service stratum
Transport stratum
MPM
MMCT
NACF
Resource and admission control functions
Other NGNs and RACF-unaware IP networks
CPE
Transport functions
Y.2111(11)_F01

Figure 21 – RACF within NGN Architecture

As shown in Figure 21, RACF act as the arbitrator between service control functions (SCF) and transport functions for QoS related transport resource control within access and core networks. The policy decisions made by the RACF are based on transport subscription information, service level agreements (SLAs), network policy rules, service priority, and transport resource status and utilization information.

The RACF provides an abstract view of transport network infrastructure to the SCF and decouples the provision of services from the details of transport facilities such as network topology, connectivity, resource utilization and QoS mechanisms/technology. The RACF interacts with the SCF, other transport control functions (e.g., mobility management and control function (MMCF), Management of Performance Measurement), and transport functions for a variety of applications (e.g., SIP-based call or video streaming) that require the control of NGN transport resources, including QoS control, NAPT/firewall control and NAT traversal. The RACF executes policy-based transport resource control upon the request of the SCF or other transport control functions; it determines transport resource availability, makes admission decisions, and applies controls to transport functions for enforcing the policy decisions. The RACF interacts with transport functions for the purpose of controlling one or more of the following functions in the transport stratum: bandwidth reservation and allocation, packet filtering; traffic classification, marking, policing, and priority handling; network address and port translation; firewall.

The RACF takes into account the capabilities of transport networks and associated transport subscription information for subscribers in support of the transport resource control. The RACF interacts with network attachment control functions (NACF), including network access registration, authentication and authorization, parameter configuration, etc., for checking transport subscription information. For delivery of those services across multiple providers or operators, SCF, RACF and transport functions may interact with the corresponding functions in other NGNs or RACF-unaware IP networks.

According to the capability of QoS negotiation, NGN identifies the CPE with following three different types:

- Type 1 – CPE without QoS negotiation capability (e.g., vanilla soft phone, gaming consoles)
  The CPE does not have any QoS negotiation capability at either the transport or the service stratum. It can communicate with the SCF for service initiation and negotiation, but cannot request QoS resources directly.
- Type 2 – CPE with QoS negotiation capability at the service stratum (e.g. SIP phone with SDP/SIP QoS extensions)
  The CPE can perform service QoS negotiation (such as bandwidth) through service signalling, but is unaware of QoS attributes specific to the transport. The service QoS concerns characteristics pertinent to the application.
- Type 3 – CPE with QoS negotiation capability at the transport stratum (e.g. UMTS UE)
  The CPE supports RSVP-like or other transport signalling (e.g. GPRS session management signalling, ATM PNNI/Q.931). It is able to directly perform transport QoS negotiation throughout the transport facilities (e.g. DSLAM, CMTS, SGSN/GGSN).

It is important to note that the SCF shall be able to invoke the resource control process for all types of CPE. In order to handle such different types of CPE and transport QoS capabilities, the RACF shall support two different modes of QoS resource control as part of its handling of a resource request from the SCF.

First is the Push Mode. In this mode, the RACF makes the authorization and resource control decision based on policy rules and autonomously instructs the transport functions to enforce the policy decision. Figure 22 shows a flow diagram for this mode and detailed procedures are following [b- ITU-T Y.2011]:

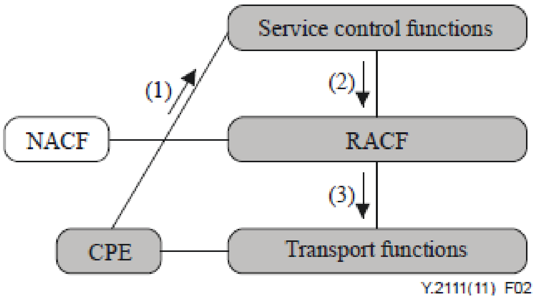


Figure 22 – Flow diagram of Push Mode

1)The CPE requests an application-specific service by sending a service request (e.g., SIP Invite, HTTP Get) to the SCF and may also send a dedicated application layer QoS signalling request. The service request may or may not contain any explicit service QoS requirement parameters;

2)The SCF extracts or derive the service QoS requirement parameters (e.g., bandwidth) of the requested service, and then requests QoS resource authorization and reservation from the RACF by sending a request for resource reservation which contains the explicit QoS requirement parameters;

3)The RACF performs authorization and admission control based on policy rules, resource admission decision and transport subscription profile stored in the NACF. If the request is granted, the RACF pushes the gate control, packet marking and bandwidth allocation decisions to the transport functions.

The other is Pull Mode. In this mode, the RACF makes the authorization decision based on policy rules and, upon the request of the transport functions, re-authorizes the resource request and responds with the final policy decision for enforcement. Figure 23 shows a flow diagram for this mode and detailed procedures are following [b- ITU-T Y.2011]:

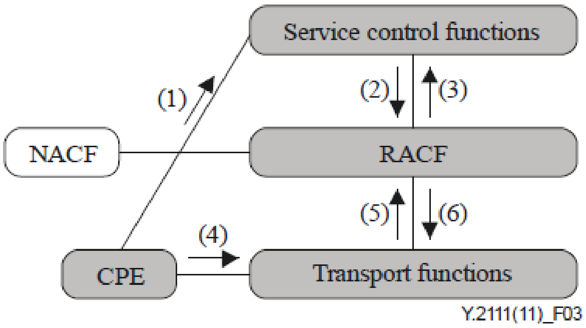


Figure 23 – Flow diagram of Pull Mode

1)The CPE requests an application-specific service by sending a service request (e.g., SIP Invite, HTTP Get) to the SCF. The service request may or may not contain any explicit (application) service QoS requirements;

2)The SCF extracts or derives service QoS requirements (e.g., bandwidth) for the requested service, and sends an authorization request to the RACF that contains explicit QoS requirements;

3)The RACF checks authorization based on network policy rules. If the resources are authorized, an authorization token (use of this is optional) is assigned to this service and informed to the CPE. It is also possible to perform authorization without the use of a token;

4)The CPE may initiate a request with QoS information for resource reservation directly to the transport functions through a dedicated path-coupled transport QoS signalling. This QoS request contains the explicit transport QoS requirement parameters for an application specific service. It may also contain an authorization token assigned at the first phase. Alternatively, the CPE may initiate a request without explicit QoS information;

5)On receipt of the QoS request, the transport functions at the network edge send a request to the RACF for resource reservation and admission control that may contain the authorization token as an option. When the CPE sends a request without explicit QoS information, the transport functions send the QoS request to RACF on behalf of the CPE;

6)The RACF makes a reservation and admission control decision based on transport subscription profile held in the NACF, service information, network policy rules and resource availability. If the request is granted, the RACF provides the gate control, packet marking and bandwidth allocation decisions to the transport functions.

7.5 Management of NGN performance measurement

ITU-T Recommendation Y.2173 identified management aspects of performance measurements of NGN covering requirements, architecture and procedure.

Requirements for performance measurement

Requirements on management of performance measurement for the NGN are identified as three different views such as high level aspects, functional aspects and non-functional view. Following Table 2 is the summary of these requirements.

Table 2 – Requirements for NGN performance measurement

Category	Requirements	Status
High Level	● Support inter-domain performance measurement; ● Enable inter-domain performance measurement functional entities to discover each other; ● Enable distribution of measured performance data from one domain to another; ● Enable aggregation of measured performance data from different domains; ● Enable management of performance measurement in a NAT/NAPT environment; ● Enable management of performance measurement in a NAT-PT environment; ● Enable management of performance measurement in an ECMP environment; ● Enable identification of performance degradation problems based on performance measurements in a multi-domain environment.	Mandatory
Functional	● Be distributed; and the distributed architecture may be federated, hierarchical, or cascading neighbor; ● Enable the partitioning of networks to provide accurate end-to-end measurement, and support comparison of measurement with provider impairment targets; ● Support a common information model for storing the measured data for the management of inter-domain performance measurement; ● Support a reference point for exchanging the measured data for the management of inter-domain performance measurement; ● Support configuration and monitoring measurement entities; ● Support control measurement entities; ● Enable collection, aggregation and storage of measurements; ● Support management of the collection of measurement data; ● Support derivation of performance metrics from measured data; ● Support of exchange of performance metrics among domains; ● Support provision of performance data to the RACF or the like for resource and admission control purposes; ● Support provision of performance data to a management application for resolving inter-domain performance degradation problems; ● Support management of active measurements, passive measurements and spatial measurements.	Mandatory
Non-Functional	● Meet the performance objectives for the latency of transfer of performance data among providers and performance data registries; ● Take into account events that have performance impacts (e.g., policing events); ● Enable performance management information to be exchanged efficiently, reliably, securely, and with scalability; ● Be consistent with a common information model that is protocol neutral, extensible, and flexible.	Mandatory

Architecture for the management of performance measurement

The three functions in NGN such as transport, transport control and service control are related with performance management in various aspects. Transport functions provide support for the transfer of media information including control and management information. Thus performance measurement and management of these functions are needed.

Transport control traffic measurement is important for ensuring the performance of NGN transport control functionality which is consisted with network access control and RACF. Especially, RACFs require reasonably accurate real-time network resource performance and usage data to enable effective resource-based admission control decisions. NGN performance management functions can provide such information to the RACF.

Service control functions include resource control, registration, and authentication and authorization functions at the service level for both mediated and non-mediated services. Service control traffic measurement is essential to ensure the quality of services offered and to account for their usage.

Taking consideration of above rationale, MPM (Management of Performance Measurement) should be belonging to one of the NGN management functions as shown in Figure 24. MPM interacts with various NGN functional entities to collect and analyze performance of NGN networks and services. It can interact with transport functions for performance measurement of NGN transport services as well as with transport control functions for NGN transport control traffic. It can also interact with service control and application support functions for performance measurement of NGN application and service control traffic. The results of such measurements can be provided to MPM applications or MPM of other NGN providers for inter-domain NGN performance measurement.

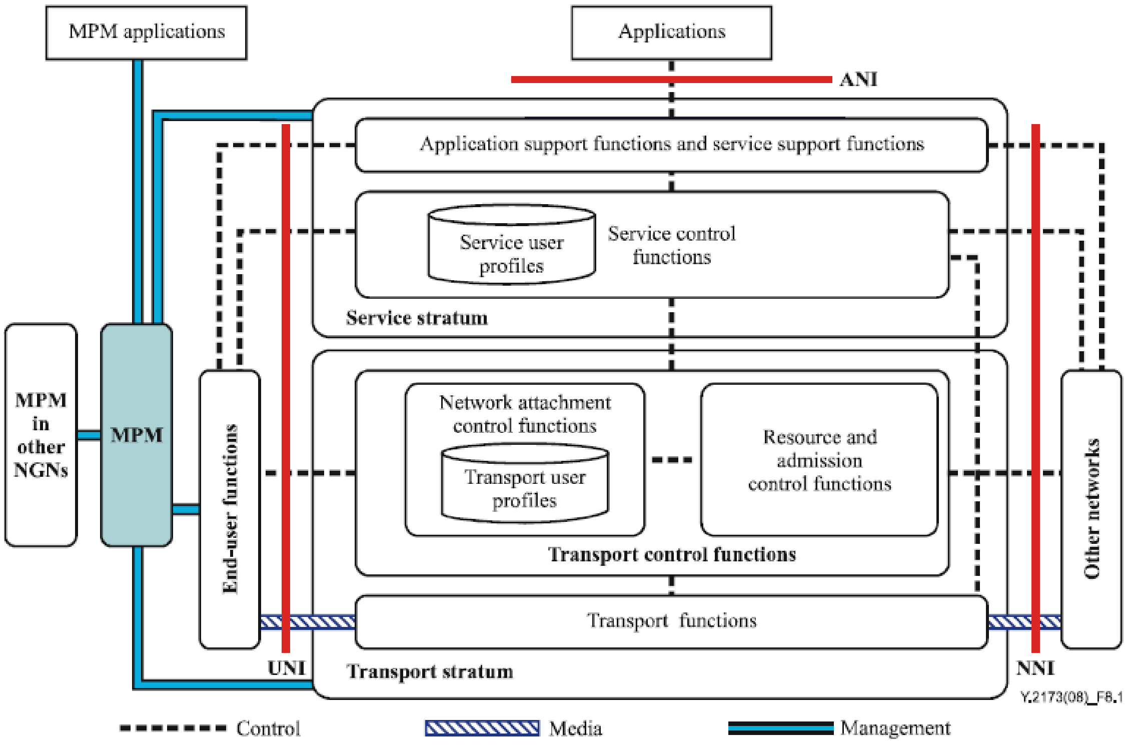


Figure 24 – Overall Architecture for NGN Performance Management

The functional architecture shown in Figure 25 is the detailed view of the functional architecture based on the requirements for management of NGN performance measurement.

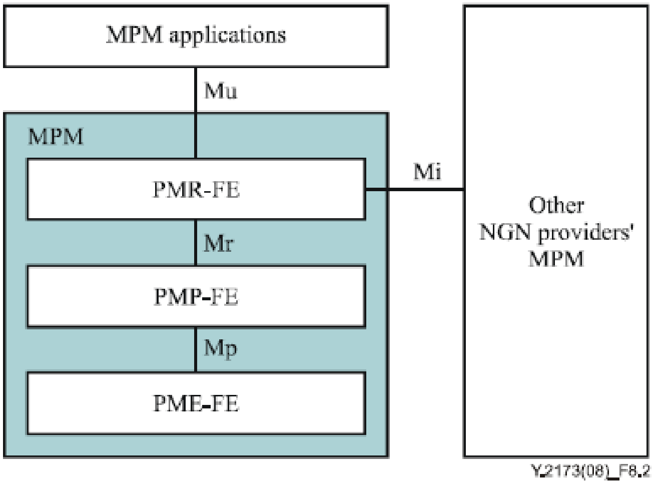


Figure 25 – Functional Architecture for NGN MPM

PME-FE (Performance Measurement Execution Functional Entity) is responsible for three groups of functionality: measurement execution, single measurement processing and measurement device configuration. Measurement execution includes active probe initiation, active probe termination and passive measurement. Single measurement processing includes collection of time-stamped packets and calculation of the single probe delay and loss. Measurement device configuration includes configuration of measurement-related policies received from PMP-FE. More specifically, the PME-FE is responsible for:

- Receiving configuration information from the PMP-FE via the Mp reference point;
- Clock synchronization among functions;
- Performance measurements, by initiating "per-class" active probes with the timestamp;
- Time-stamping the received active probes and packets collected;
- Suspending initiation of the probes and collection of packets and reporting to the PMPFE via the Mp reference point if loss of clock synchronization occurs;
- Processing the time-stamped probes and packets;
- Generating a flow identification and packet identification for each captured packet;
- Collecting time-stamped packets;
- Calculating delay and loss of single probe under measurement based on the time stamped packets;
- Exporting single probe delay and loss results to PMP-FE via the Mp reference point.

PMP-FE (Performance Measurement Processing Functional Entity) is responsible for two groups of functionality: measurement processing and network-wide measurement configuration. The measurement processing function includes measurement report collection, passive measurement report analysis, measurement data aggregation and rollup period analysis. It receives single measurement processing result from the PME-FE via the Mp reference point and sends the results of analysis to the PMR-FE via the Mr reference point. The network-wide measurement configuration function includes the creation of network-wide performance measurement configuration policies, the selection of appropriate measurement points at which to apply them, and the deployment of the policies into the measurement points. More specifically, it performs the following functions:

- Collecting active measurement, passive measurement or spatial measurement reports from the PME-FE via the Mp reference point;
- Performing flow-based passive measurement analysis using (for example) an RTP/RTCP-based scheme;
- Collecting single probe delay and loss results from the PME-FE via the Mp reference point;

- Calculating the rollup metrics including IPTD, IPDV, IPLR and IPUA, etc., based on single probe delay and loss results;
- Exporting rollup metrics to the PMR-FE via the Mr reference point;
- Performing measurement data aggregation to reduce the amount of data to be processed;
- Performing correlation analysis among data received from various PME-FEs.

PMR-FE (Performance Measurement Reporting Functional Entity) collects rollup metrics from the PMP-FE via the Mr reference point and provides reports to MPM applications (e.g., RACF), or other NGN provider's MPM. It performs the following functions:

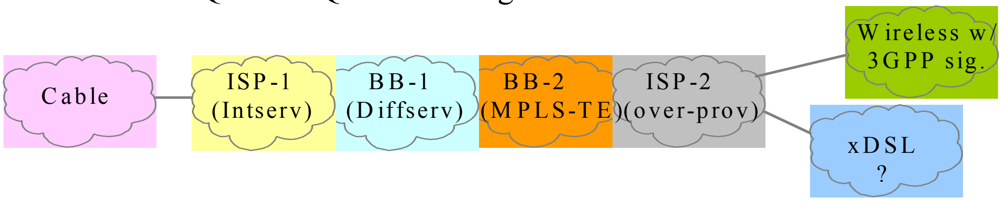
- Collecting rollup metrics from the PMP-FE via the Mr reference point;
- Reporting to MPM applications via the Mu reference point;
- Requesting and responding to requests from other NGN providers' PMR-FE for network performance metrics via the Mi reference point;
- Authenticating requests for initiation of measurements from MPM applications or other NGN providers, and initiating the requested measurements.

8 Considerations for Regional Environment

8.1 Considerations for practical status

IP-platforms and networks

There is a fundamental difficulty in the IP based platforms and networks. That is heterogeneity. Just use of IP as a transport technology does not mean networks and platforms are same even compatible. Figure 26 shows an extreme case of configuration among different networks but using IP as a transport mean. In this case, it is hard to provide services from one end to other end with certain level of quality, because each network has different mechanism and the level of control and provision are also different. Therefore standard documents developed and agreed at the regional level (possibly based on global standards such as ITU-T Recommendations) are required as an essential parts to increase QoS and QoE in that region.



Note: BB ~ Backbone network

Figure 26 – Heterogeneous Network Environments

Configuration in the region

A region should be consisted with several countries whether their belongings to the region identified by geometrics or politics or business or any others. Each country in the region also consisted with various several providers in terms of networks and services with their own policy and regulatory environments. And as a minimum element, each provider composed of various networks to cover their business areas with various different technologies according to the services requirements and features.

Following Figure 27 shows an example configuration of this scenario in the region.

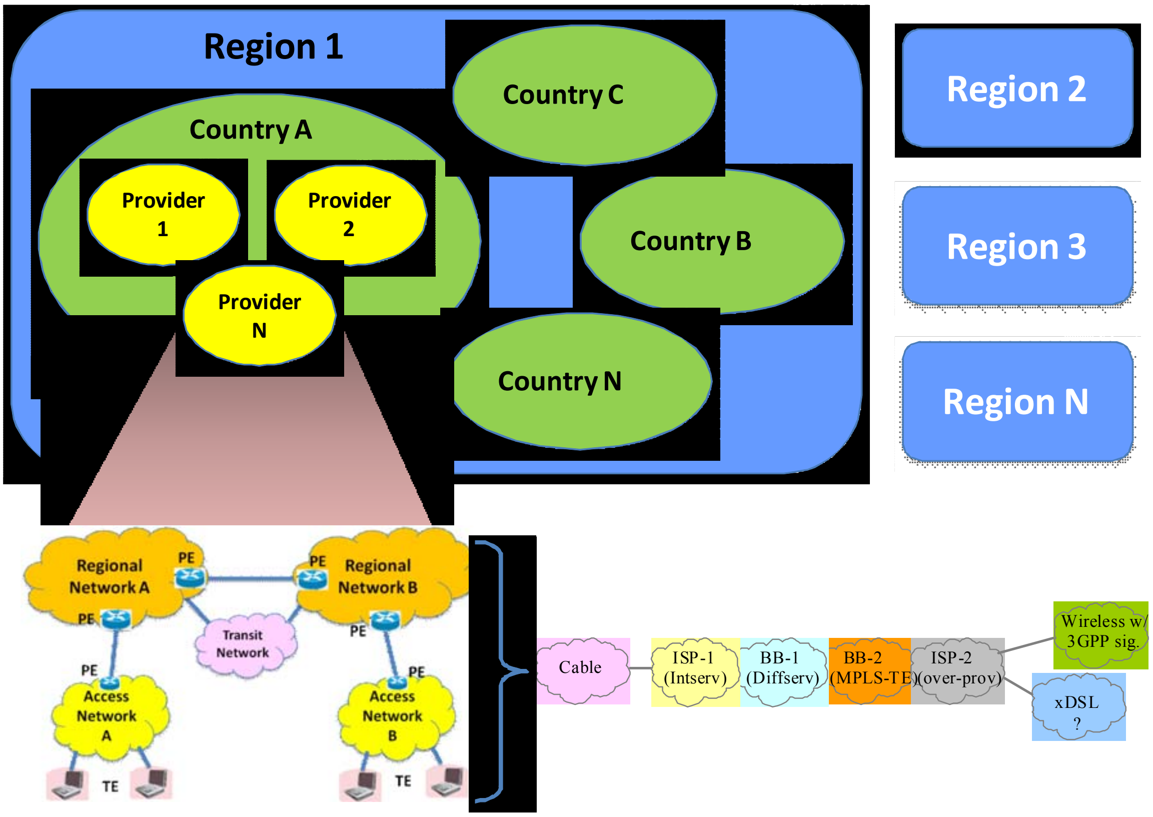


Figure 27 – QoS/QoE environment in the Region and Global

A basic network model used in this Technical Paper (left bottom in Figure 27) should be applied commonly to the providers, but the detailed technologies to compose such network model would be different by providers and countries. In addition to this, there are various differences on regulation environment which are fundamental framework for such as SLAs, QoS and QoE.

Taking into account these, it is noted that common standard applying in the region should be developed. In this sense, followings are recommended to be considered in the region for the development of relevant standards for QoS/QoE.

- Common understanding about the nature of NP, QoS, QoE and SLAs including relationships among them including common terminology;
- Ensuring the SLA for end-end is an important objective but it is quite difficult to realize this practically in the regional level. So it is recommended to develop common parts of parameters identify NP and QoS based on global standards (possibly incorporated specifics if needed in the region specifically);
- Agreed details in QoS building blocks (see Figure 10). More special attention should be needed in the management plane mechanism because this provide a tool for communication between providers and countries;
- Building consensus on the measurement objectives and methodologies are crucial to verify and evaluate the practical operation of the networks and provision of services;
- Develop reference network model of IP based for the development of detailed standards in the region.

8.2 Considerations for performance measurement and its management

There are many of issues to be considered when identify standards possibly applying to the region. Followings are derived from several ITU-T recommendations.

Applied measurements

Measurement purposes fall into three broad categories, operating, supporting and testing.

- Operating measurements are those which are made on an ongoing basis between measurement points to monitor normal operation of the assured segments along customers' data paths, e.g., measurements of ingress, transit and egress segments.
- Supporting measurements, which may be taken continuously, are used to provide information for SPs. These measurements occur in addition to operating measurements and can be between various measurement points, e.g., measurements of each SPs' contribution to the transit segment.
- Testing measurements are made on an exception basis following the detection of abnormal operating measurements for troubleshooting or to test a new path. These measurements occur in addition to operating or supporting measurements and are between measurement points which do not have operating or supporting measurements being taken, e.g., measurement of a particular CE-to-CE path for a prospective customer.

Some measurements may fall into multiple classes. For example, a CE-to-CE measurement may be used for a prospective customer (testing), as a sanity check for providers (supporting), or as a premium (un-scalable) customer service (operating).

Different views of the same measurement data may be useful for different purposes. For example, a provider that collects and analyses ongoing measurements at sub-intervals of RP may evaluate the impact of remedial action upon network performance more quickly than had they waited for the RP before doing so.

The following scenarios show how the various performance measurement techniques may be applied to the measurement network models. The flexibility of the models support more applied measurements than those described previously.

In the following scenarios, the measurement information exchanged among providers every rollup period includes the following:

- minimum delay;
- mean delay;
- high delay percentiles;
- loss ratio;
- unavailability period information;
- miscellaneous information

All measurement scenarios described below are applicable to active, passive and spatial measurement techniques, unless otherwise noted. When measurement results have been obtained, the results should be conveyed from the collection points to management systems with oversight responsibility.

Generic inter-domain management process for measurement systems

Figure 28 shows the general procedures for management of inter-domain performance measurement systems. Measurement point (MP) is a functional entity located in the transport, transport control, or service control networks. In case of active measurement, it is responsible for initiating and receiving probe packets. In case of passive measurement, it is responsible for capturing target packets. Management of performance measurement (MPM) functions includes the interaction with measurement applications and the MPs, configuration of MPs, and exchanging the required configuration and measured information.

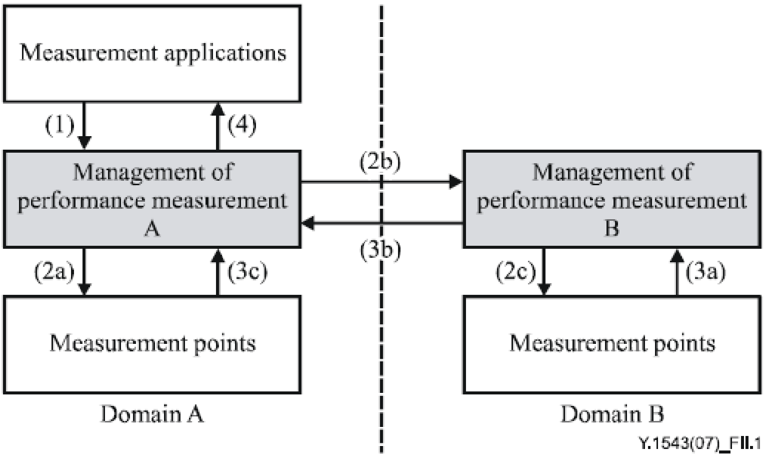


Figure 28 – MPM Procedure for Inter-Domain

(1) The measurement application of SP A initiates a measurement task by sending measurement request to MPM.
(2a) Upon receipt of measurement request, MPM A locates the involved MPs. For the MPs located in domain A, MPM A sends the measurement parameters to MPs.
(2b) Upon receipt of measurement request, MPM A locates the involved MPs. For the MPs located in domain B, MPM A sends the measurement request to MPM B.
(2c) Upon receipt of measurement request, MPM B locates the involved MPs. For the MPs located in domain B, MPM B sends the measurement parameters to MPs.
(3a) MPM B collects the measured data from MPs located in domain B.
(3b) MPM B sends the measurement information to MPM A.
(3c) MPM A collects the measured data from MPs located in domain A.
(4) Based on the received measurement information from domain A and domain B, MPM A sends the response to the measurement applications.

Management consideration for performance measurement in inter-domain

The inter-domain is an essential part for configuring regional level of networks which subject for QoS/QoE enhancement. ITU-T Recommendation Y.2173 deals with the case for supporting inter-domain. [b- ITU-T Y.2173] Inter-domain performance measurement requires close collaboration between different administrative domains such as country or regional regulatory organization. Performance measurement in each domain can be relatively easily achievable. However, when a measurement crosses a domain boundary, the complexity increases dramatically. The main issues involved are the following:

- Who will measure what, and how?
- What is the common data model to store the measured data?
- How should the measured data be exchanged?
- Are PMR-FEs involved in the measurement collaboration?

Depending on the answers to these questions, it can classify the architectures as follows:

- Architectures not involving PMR-FE: In this case, manual model is considered. In the manual model, performance measurement data is stored in a standardized common information object and exchanged among service providers through a standard protocol. Management of measurement processes such as configuration of active/passive probes, collection of performance metrics, conversion of metrics to common information objects, and triggering exchange protocols may be performed in a proprietary way in each domain. This model does not assume that there exists a representative performance measurement management system which coordinates all such processes. The advantage of this model is simplicity and cost effectiveness. However, configuration in each domain requires manual intervention, thus it has limitations in automation. This model may raise security issues if the exchange protocol is not secure.

- Architectures involving PMR-FEs: In this case, one or more PMR-FEs exist in each domain and are responsible for both internal and inter-domain performance measurement management and collaboration.
  - Centralized model: In the centralized model, a single PMR-FE is responsible for the management of all the active and passive measurement over each domain. It is simple to manage but has scalability limitations. Also, it is not easy to have one centralized PMR-FE control domains that fall under different administration responsibility. Scalability issues may arise since all performance measurement data are required to be reported to one centralized PMR-FE;
  - Distributed models: there are several models to support distribution of management as following:

    1) Federated model: In this model, PMR-FEs of the domains are structured into a freely federated process group for the management of active and passive measurement. This model distributes the responsibility of the centralized PMR-FE into a number of PMR-FEs across domains. Thus, it enhances the scalability greatly. Figure 29 illustrates one example model. PMR-FE1, PMR-FE2 and PMR-FE3 are responsible for regional network A, regional network B and a transit network. Each PMR-FE measures performance data for its associated regional/transit network, and exchanges it with other PMR-FEs to collaborate in developing end-to-end performance data;

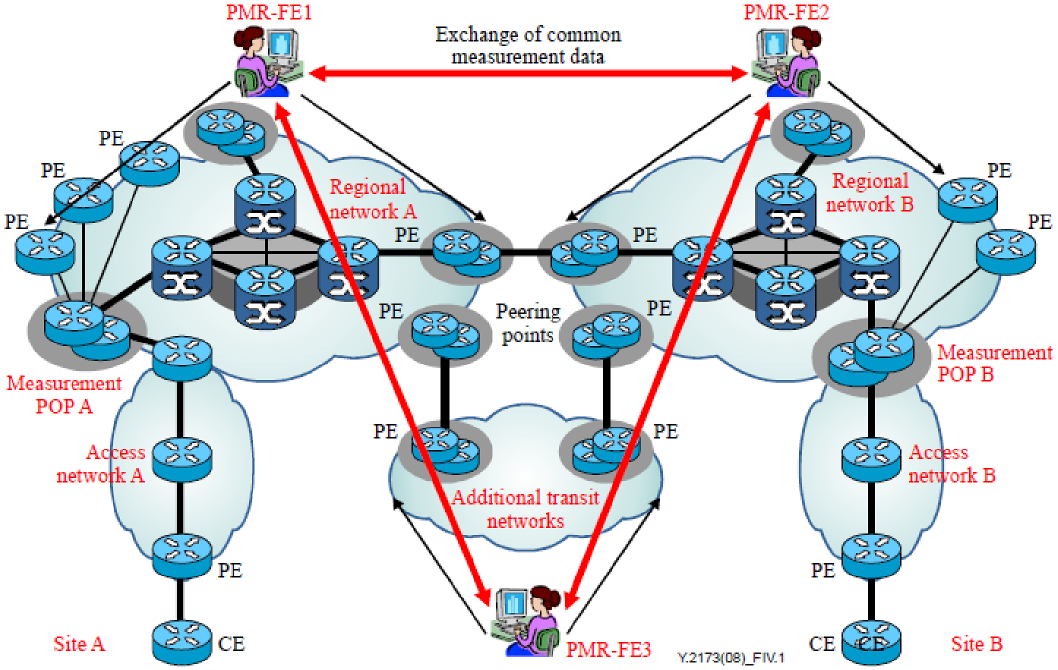


Figure 29 – An example federated inter-domain measurement model

    2) Hierarchical model: In this model, PMR-FEs of the domains are hierarchically stratified into a process group of a well-defined structure for the management of active and passive measurement. This model is similar to the federated model but involves a more rigid relationship and structure among PMR-FEs. A PMR-FE at a certain level can perform specifically defined functions only. Lower level PMR-FEs perform detailed functions and upper level PMR-FEs perform overall functions. For example, one lower-layer PMR-FE measures the access network segment, another measures the backbone segment, and still another measures the transit or peering segment. An upper-layer PMR-FE then correlates the results from lower layer PMR-FEs and exchanges them with peers in other domains. Figure 30 shows an example

hierarchical model. PMR-FE1 manages two PMR-FEs which perform measurement of access and core networks. Similarly, PMR-FE2 manages two PMR-FEs. PMR-FE 4 is the highest level PMR-FE which sits on top of PMR-FE1, 2 and 3;

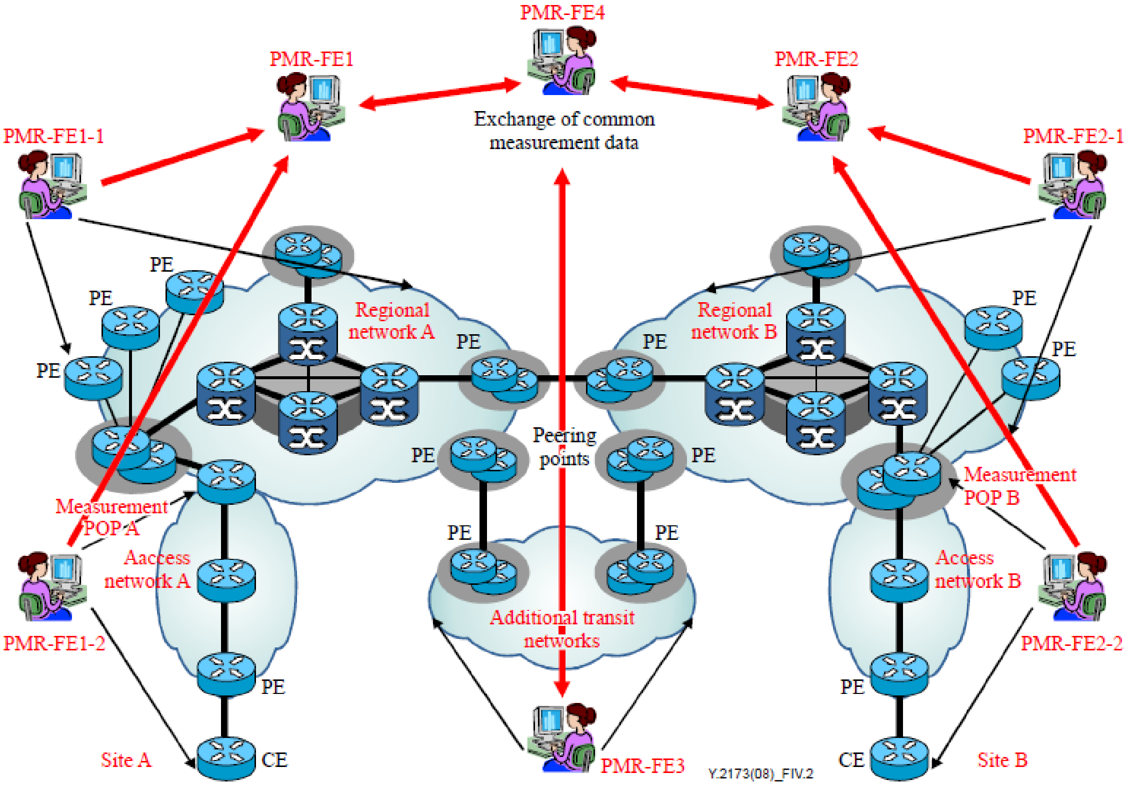


Figure 30 – An example hierarchical inter-domain measurement model

3) Cascading neighbour model: In this model, PMR-FE of the domains are interconnected only with those of the neighbouring domains to create a well-defined structure for the management of active and passive measurement. Figure 31 illustrates a cascading neighbour inter-domain measurement model. Each PMR-FE exchanges measurement data with the neighbouring PMR-FE only. Direct interactions occur only among neighbouring PMR-FEs.

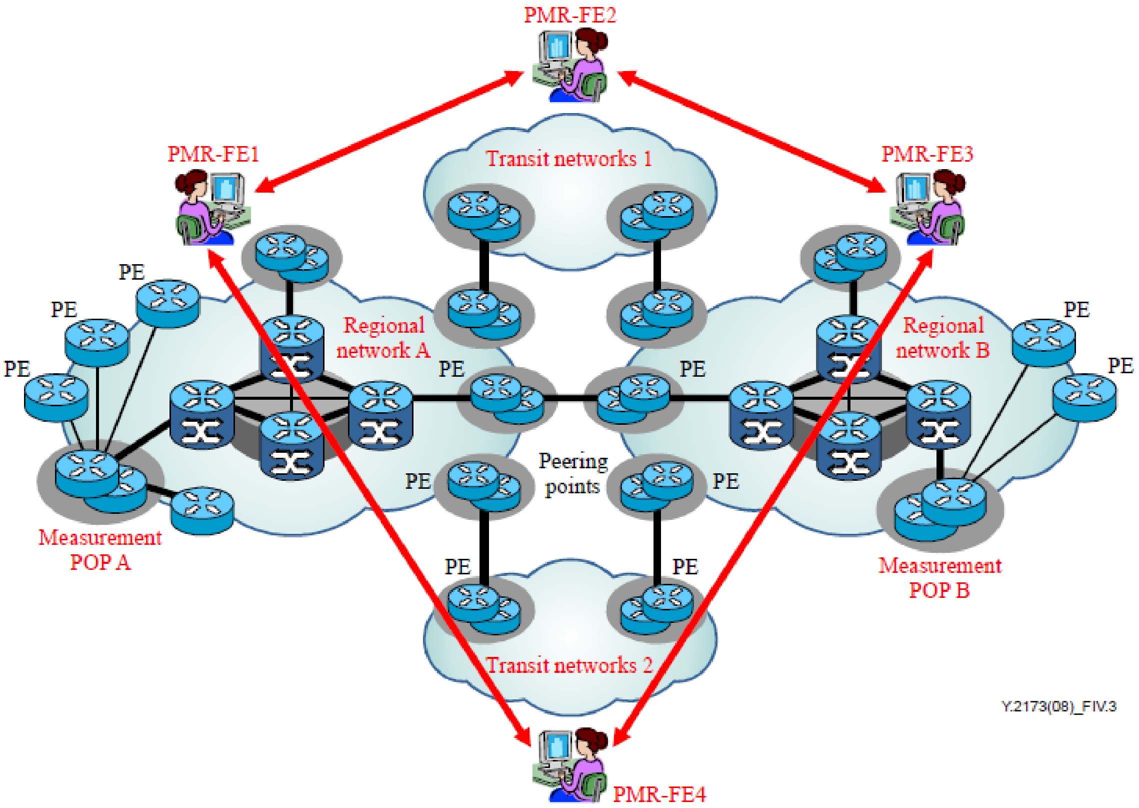


Figure 31 – An example cascading neighbour inter-domain measurement model

Example realization for the management of NGN performance measurement

To provide better understanding of the general architecture, it is useful to look at example of realization. ITU-T Recommendation Y.2713 provides example architecture for the realization of the management of performance measurement in terms of NGN. In this example, the performance reporting system represents the performance measurement reporting functional entity (PMR-FE). Collection platforms are used as an instance of performance measurement processing functional entity (PMP-FE). Measurement points are instances of the performance measurement execution functional entity (PME-FE). Configuration, measurement and reporting are accomplished by a group of devices having different functions which are arranged in a hierarchy. The hierarchy and the communications among components are shown in Figure 32.

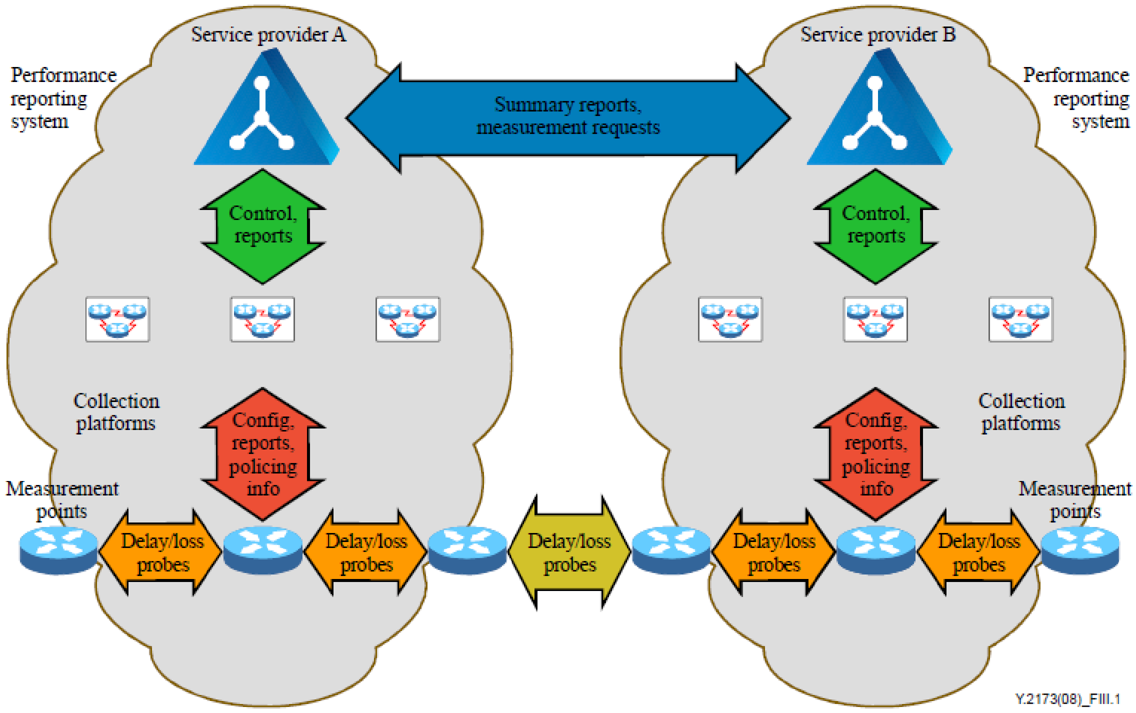


Figure 32 – Active measurement architecture inter-providers

- Performance reporting systems: At the top of the hierarchy is the performance reporting system (PRS). While this may be implemented using multiple computers, it will appear as a single system to other SPs. The PRS is the primary configuration tool. The PRS pushes probe initiation commands to a collection platform associated with the measurement point that initiates probing. That collection platform then configures the measurement points.
- Collection platforms: At the second level of the hierarchy are the collection platforms. Collection platforms communicate "up" to the PRS and "down" to the measurement points.
- Measurement points: At the third level of the hierarchy are the measurement points that send and receive probes. Measurement points communicate "up" with associated collection platform(s), and "horizontally" with other measurement points, which may belong to other SPs.
- Customer terminal equipment: The functions of customer "landmark" terminal equipment are to be defined. When TEs are included in a measurement scheme, the functionality of customer equipment is likely to be expanded.

References and Bibliography

[b-ITU-T E.800] Recommendation ITU-T E.800 (2008), *Definitions of terms related to quality of service*.

[b-ITU-T G.1000] Recommendation ITU-T G.1000 (2001), *Communications Quality of Service: A framework and definitions.*

[b-ITU-T P.10] Recommendation ITU-T P.10 (2006), *Vocabulary for performance and quality of service.*

[b- ITU-T G.1010] Recommendation ITU-T G.1010 (2001), *End-user multimedia QoS categories.*

[b- ITU-T G.1011] Recommendation ITU-T G.1011 (2010), *Reference guide to quality of experience assessment methodologies.*

[b- ITU-T G.1030] Recommendation ITU-T G.1030 (2005), *Estimating end-to-end performance in IP networks for data applications.*

[b- ITU-T G.1050] Recommendation ITU-T G.1050 (2011), *Network model for evaluating multimedia transmission performance over Internet Protocol.*

[b- ITU-T E.860] Recommendation ITU-T E.860 (2002), *Framework of a service level agreement.*

[b- ITU-T E.803] Recommendation ITU-T E.803 (2011), *Quality of service parameters for supporting service aspects*

[b- ITU-T Y.1291] Recommendation ITU-T Y.1291 (2004), *An architectural framework for support of Quality of Service in packet networks.*

[b- ITU-T Y.1543] Recommendation ITU-T Y.1543 (2007), *Measurements in IP networks for inter-domain performance assessment.*

[b-ITU-T E.800-Sup.8] ITU-T E.800-series Recommendations − Supplement 8 (2009), *Guidelines for inter-provider quality of service*.

[b- ITU-T Y.1540] Recommendation ITU-T Y.1540 (2011), *Internet protocol data communication service – IP packet transfer and availability performance parameters.*

[b- ITU-T Y.2111] Recommendation ITU-T Y.2111 (2011), *Resource and admission control functions in next generation networks.*

[b- ITU-T Y.2173] Recommendation ITU-T Y.2173 (2008), *Management of performance measurement for NGN.*

[b- ITU-T Y.2001] Recommendation ITU-T Y.2001 (2004), *General overview of NGN.*

[b- ITU-T Y.2011] Recommendation ITU-T Y.2011 (2004), *General principles and general reference model for Next Generation Networks*.

[1] ITU-T SG12 Workshop, “End to End QoS control in VoIP systems”, Mike Buckley, Telchemy

[2] ITU-T Workshop on End-End Quality of Service, Neal Seitz, Geneva 1-3 October 2003

[3] “Framework and Standardization of Quality of Experience (QoE) Design and Management for Audiovisual Communication Services”, Akira Takahashi, NTT Technical Review, Vol. 7 No.4 Apr. 2009