INTERNATIONAL INTERCONNECTION FORUM FOR SERVICES OVER IP

(i3Forum)
www.i3forum.org

Workstream "Services Aspects"

Service value and process of measuring QoS KPIs

(Release 1.0) May 2010

This document further analyzes and updates the i3Forum guidelines on QoS given in "IP international interconnections for voice and other related services - (V 1.0) June 2009"

Executive Summary

The main objective of i3Forum, representing a large portion of the international voice wholesale industry, is to continue to deliver the current standard for high quality to the market while introducing new technology and services over IP. In the i3Forum document "IP international interconnections for voice and other related services - (V 1.0) June 2009", a list of relevant quality parameters for voice services has been presented and discussed. The scope of this document is to further analyze both values and limits of measuring and controlling relevant quality parameters for Voice traffic transported over IP networks, in order to share with the industry a common understanding on what, when, where and how to use those parameters. An i3Forum position on a set of relevant KPI is also presented, aiming to clarify the implication on the reach of responsibility of the wholesale carriers, and to give guidance in the form of guidelines on managing these parameters. A pragmatic approach to introduce end-to-end quality monitoring is presented.

This document is structured to first provide a short description of the process of measuring each relevant key indicator in order to highlight the reach of responsibility and to explain the value for carriers/service providers of measuring the parameter, together with some limitations that are specific for the parameter. Finally, the document proposes some recommendations for end-to-end measurement and responsibility in line with the current technical implementation.

With relation to voice parameters (i.e. NER, ASR, MOS, …), it has to be highlighted that some parameters by definition include the end user behavior (e.g., ASR) and should not be considered for a service level agreement (SLA), given end user behavior is not dependent on activities which are under the control or manageable by carriers. Some others service parameters (e.g. NER) by definition include the performances of all operators in the chain, carriers and retail service operators, and a commitment on the "end-to-end" result for these indicators should involve all operators in the chain. It has to be noted however that a carrier commercially committed to a SLA/SLO for quality parameters can face potential circumstances where, even though the performance has been controlled and has resulted in a value above the threshold, the specific value for a single customer may be below the committed value. This leads to the conclusion that a carrier is always taking a commercial risk and the SLA responsibility can not be fully cascaded downstream in the delivery chain.

Additional care has to be devoted for measuring parameters impacted by release causes, because in some specific interworking scenarios the reported nature of the release can change and this may lead to a misleading measurement of the parameter. An analysis of impacts and some guidelines are provided in Annex 2.

With relation to IP parameters (i.e. Packet Loss, Round Trip Delay and Jitter), it is important to remember that IP layer quality is not alone representative of the voice quality. Considering the current alternatives to compute these parameters, i3Forum endorses the measurement point with the Border Function that can assure values are related to the real voice traffic, and can allow end-to-end monitoring at these network edge points. i3Forum recognizes the current network implementation and IP migration might limit a wide use of this approach but suggests this method should be used in carrier-to-carrier interconnections.

Table of Contents

1. Scope of the document

The scope of this document is to present the value and limits of measuring and controlling relevant quality parameters for Voice traffic transported over IP networks. The main objective of the voice wholesale industry is to continue to provide to the market the high level of quality standard for premium routes, while introducing new technology and services over IP. A common recommendation on a set of relevant KPIs is presented together with a detailed explanation on the wholesale industry responsibility for managing these parameters and a pragmatic approach to introduce end-to-end quality monitoring.

With reference to the list of QoS KPIs described and recommended in previous i3Forum documents [1], [2], the document aims to:

a) describe the process to measure key quality indicators (KPI);
b) explain the service value of measuring these indicators for SPs/carriers and/or customers;
c) recommend some guidelines for end-to-end measurement and range of responsibility.

2. Acronyms

ALOC	Average Length of Call
ASR	Answer Seizure Ratio
CDR	Call Detail Record
CLI	Calling Line Identification
KPI	Key Performance Indicator
MOS	Mean Opinion Score
MTRS	Maximum Time to Restore the Service
NER	Network Efficiency Ratio
PGRD	Post Gateway Ringing Delay
SIP	Session Initiation Protocol
SIP-I	SIP with encapsulated ISUP
SLA	Service Level Agreement
SLO	Service Level Objective
SP	Service Provider

3. Document References

[1] i3Forum, “Technical Interconnection Model for International Voice Services”, Release 3.0, May 2010

[2] i3Forum, “IP international interconnections for Voice and other related services”, June 2009

[3] i3Forum, “White Paper – Mapping of signalling protocols: from ISUP to SIP, SIP-I” Release 2, May 2010

[4] ITU-T Recommendation Y.1540 “Internet Protocol Data Communications Services - IP Packet Transfer and availability performance parameters”, November 2007

[5] IETF RFC 3393 “IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)”, November 2002

[6] ITU-T Recommendation P.10 “Vocabulary of terms on telephone transmission quality and telephone sets”, December 1998

[7] ITU-T Recommendation G.107 “The E model, a computational model for use in transmission planning”, March 2005

[8] ITU-T Recommendation E.437 “Comparative metrics for network performance management”, May 1999

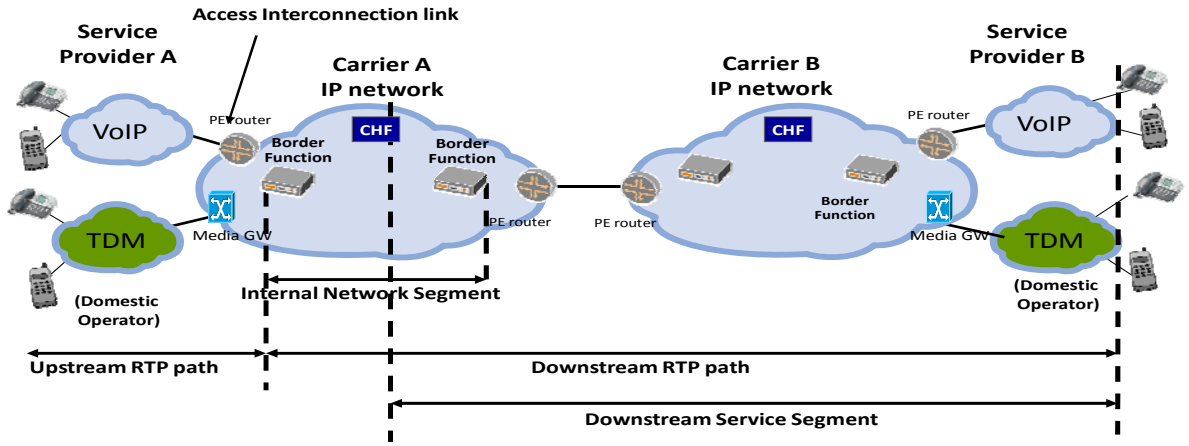
[9] ITU-T Recommendation E. 411 “International Network Management – Operational guidance”, March 2000

[10] ITU-T Recommendation E.425 “Network Management – Checking the quality of the international telephone service. Internal automatic observations”, March 2002

4. Reference Schemes

The reference diagram for international interconnection in figure 1 [1] will be used in order to describe all following QoS parameters in this document. In this reference diagram only two Carriers are involved in the delivery chain between Service Providers (SP), but all considerations and recommendations may also be used for a scenario where more than two Carriers may be involved in the service provisioning, or alternative providers are used for the same origin/destination SP at the same time.

This reference diagram figure 1 should help identifying the correct responsibility of the networks in the chain in case of a dispute. It should not be used to determine the process of measuring QoS, which will be explained later.



CHF: Call Handling Functions
Note: it is possible that more than two Carriers can be involved in the Service Provider-to-Service Provider communication.
If more than two Carriers are involved, Carrier B is meant to be the last in the path,
i.e. the Carrier interconnecting to Service Provider B. Consequently, Carrier A and Carrier B may not have a direct relationship.

Figure1 – Reference configuration for the Carrier-to-Service Provider relationship

Considering both the objectives to determine the end-to-end measurement required by customers, and to identify the responsibility in case of dispute, the document will describe the measurement process for each operator in the chain in order to compute the quality parameters from its points of measurement down to the end destination, and to evaluate its contribution to the quality degradation. This approach excludes the need to share and aggregate each individual Network segment quality measurements in order to compute the end-to-end value, avoiding complex and more expensive solutions (e.g. exchange and post-elaboration of each internal network segment measure or a public database containing partial quality contributions from each operator for an end-to-end measurement).

In previous i3Forum documents, there is a description of the general back-to-back SLA/SLO principle where a party in the chain can cascade the responsibility downstream to the next party in the chain, assuming the first party is able to determine whether the quality degradation is (or is not) its own responsibility. This document further studies the definition of technical and operational standard procedures in the measurement and implementation of the back-to-back SLA/SLO principle.

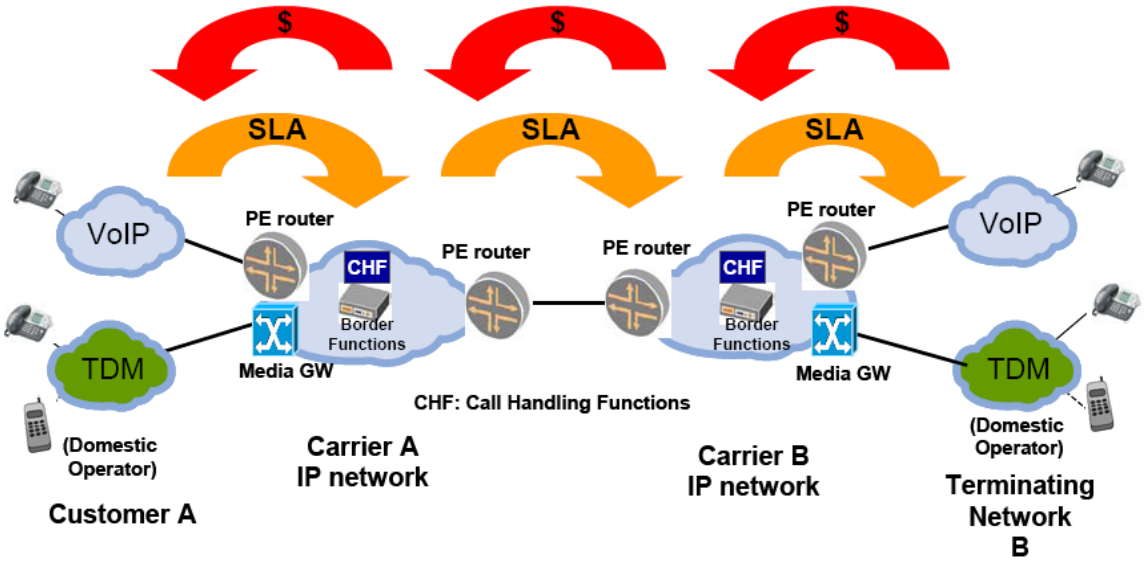


Figure 2 - Reference Scheme with Back to Back SLA/SLO among Carriers

The theoretical principle of back-to-back SLA/SLO is quite easy to describe. However, in the current wholesale context, where multiple providers are often used and the aggregation of traffic from different sources toward the same destination is commonly used, there are possible situations where the cascading responsibility principle can not be assured.

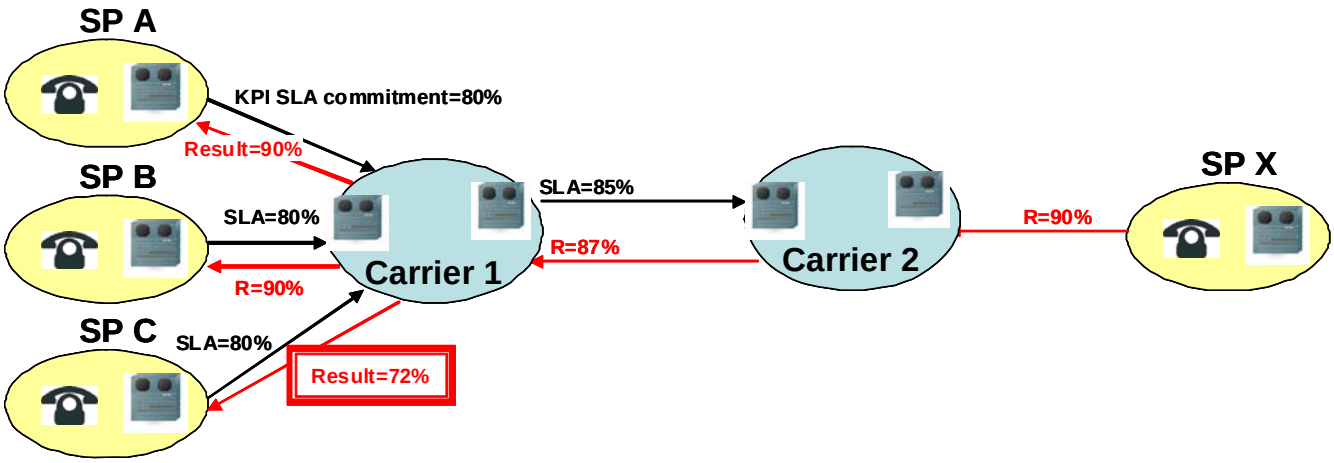


Figure 3 – Example of limitation of the cascading back to back SLA principle

As described above in the figure 3 for a generic KPI (e.g. NER), the Carrier 1 with a SLA commitment with different SPs for the same destination SP X can be subject to penalties even if the cascading

mechanism has produced a result with Carrier 2 that is in line with the committed SLA. In this example, Carrier 1 is not meeting the SLAs agreed with its customer SP C. However, the provider for Carrier 1 (Carriers 2) is meeting its SLAs towards Carrier 1. As a consequence, Carrier 1 is liable to SP C, but Carrier 1 cannot cascade its responsibility and penalties to its downstream providers (Carrier 2). The reason this problem can occur is due to the fact that while the SLAs are monitored individually with SPs, the traffic is all mixed and switched simultaneously by Carrier 1, and then delivered to the downstream providers.

This example shows that when a Carrier is commercially committed to SPs to assure a SLA/SLO for quality parameters it is always taking a commercial risk even in the case when the Carrier is requiring and controlling adequate SLA/SLO commitments from its downstream (Carrier 2) provider(s).

5. Relevant QoS parameters

According to other i3Forum documents ([1] - [2]), the following parameters are considered as the most relevant for the voice Quality of Service (QoS). The list indicates which KPIs are not affected by the end-user behavior and hence are fully under the control of carriers and service providers. Therefore, those KPIs under the control of Carriers and SPs could be included in a SLA/SLO agreement, with the assumption that all involved players are committed. The KPIs affected by end-user behavior should be used for "information" only and not included in a SLA/SLO agreement. Any commercial agreement associated with parameters to be considered for SLA/SLO and/or QoS reporting is subject to agreement between parties, and is outside the scope of this document.

	Dependent on Carriers and SPs
▪ **Service parameters**	
▸ NER 2002	fully
▸ ASR	partially
▸ ALOC	partially
▸ PGRD	fully
▸ MOS_{CQE} / R-factor	fully
▪ **Call attributes**	
▸ CLI Transparency	fully
▪ **Transport parameters**	
▸ Round-Trip Delay	fully
▸ Jitter	fully
▸ Packet Loss	fully

It has to be noted that the relative measurement for all parameters includes the terminating Service Provider network performance. For a coherent end-to-end SLA commitment, the measurement points in the terminating SPs network need to also be included in the back-to-back SLA mechanism. In the case when the SP is not committed, the measurement of the KPIs can still be achieved and reported, but the performance is not under the full control of the involved carriers.

In addition to this set of parameters defined by i3Forum in previous documents, the following other optional KPI is considered important to be measured:

- **Service parameters**
  - MTRS - Maximum Time to Restore the Service (for basic Voice telephony)

For the definitions of parameters, please refer to **Annex 1** which reflect the description included in [1].

6. List of Service parameters

6.1 NER - Network Efficiency Ratio

6.1.1 Process of Measurement

- **NER** is measured by means of an analysis of the CDR's generated by the **Call Handling Function** element of the network which considers the number of sent INVITE and the Release Causes (RC) for all delivered calls in a given period of time.

- **NER** measures the effectiveness of the networks to establish calls successfully in the downstream direction, i.e., in the **downstream service segment**: from the Call Handling Function (CHF) of the measuring carrier down to the terminating SP's network, which contributes to the measure. The NER measures the capability of all downstream networks to set up calls to a given destination, not the audio quality of the calls themselves.

It has to be understood that in some specific interworking scenarios, the meaning of Releases Causes can change due to some translations in mixed TDM/VoIP environments and may lead to a meaningless measurement of the parameter. For this reason, i3Forum has proposed some technical recommendations to significantly limit the impact of such eventuality (see [3] for details). For the purpose of this document, i3Forum assumes the issue of coherency in **SIP Release Causes** to be already solved through specific recommendations and their practical implementation, so that the consistency of the measured value is assured. More details describing the Release Cause issues can be found in Annex 2.

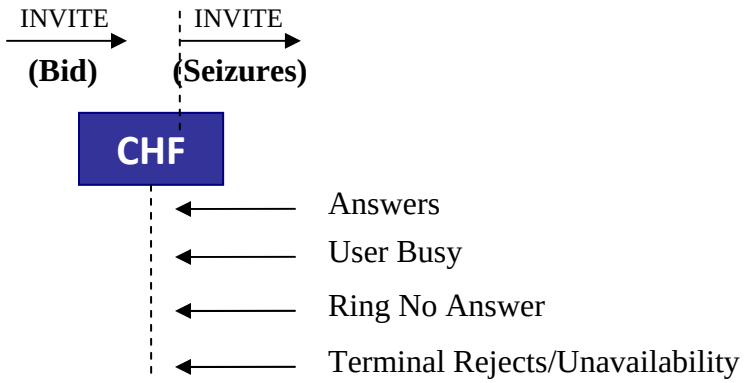
6.1.2 Service Value of measuring NER

- NER, as well as all other Voice Parameters with the exclusion of MOS, **refers only to the call set-up process**. NER considers the effectiveness of the networks in establishing calls successfully and does not provide any information regarding quality degradation or failures on the RTP flows (i.e., audio/video) when the call is already established.

- NER is evaluated **after** the Call Handling (Switching) Functions and may or may not include some degradation determined from the transport layer, according to the specific network architecture. In most cases, the transport degradation from the end user handover to the SP up to the Call Handling equipment within the SP's network can be considered negligible.

- NER is calculated on **Seizures,** meaning the number of INVITEs sent by the Switching Functions (not INVITEs received from the previous actor in the chain, usually referred to as Bid). As a consequence, the NER value measured by a specific Carrier A does not take into account the failures due to its own switching element (Bid requests coming from SP A which do not become Seizures for Carrier A because it is blocked by some switching element of Carrier A). Thus, every Carrier, by measuring NER, actually measures the effectiveness of the downstream network up to the far end destination, including the contribution of the destination SP, but excluding its own performance.

NER Evaluation



Note: NER is calculated on Seizures delivered to the far end terminal, which include Answers + User Busy + Ring No Answer + Terminal Rejects/Unavailability.

- The formula to compute the NER is included in Annex 1 (see also [1]). It has to be reminded that the measurement can be affected by the correct Release Causes mapping which is still under review (see Annex 2).
- Considering the adopted reference diagram in figure 1:
  - **Service Provider A**, measuring NER in its network actually measures the effectiveness of **Carrier A** in establishing calls successfully all the way to the destination SP B.
  - **Carrier A**, measuring NER in its network actually measures the effectiveness of **Carrier B** (or an aggregation of Carrier B and down steam providers) in establishing calls successfully all the way to the destination SP B.
  - **Carrier B**, measuring NER in its network actually measures the effectiveness of **Service Provider B** in setting-up calls successfully.

- Each Carrier is also able, using its own network statistics, to measure the network degradation introduced by its own network. Then, each Carrier can estimate the NER values measured by the previous player in the chain, for example, by considering the successfully delivered calls versus the total received INVITE (Bid), and not the total sent INVITE (Seizures).
- As a result, **each player** in the delivery chain has the ability to calculate for its customer (previous element in the chain) the aggregate measurement of NER per destination as a result of its own and its provider's performances. It has also the capability to measure the specific performance of each of its providers per destination.

- In case of NER parameter required for an end-to-end SLA, from the measurement point of view, a Carrier (or Service Provider) can compute on its own the proper NER measure to monitor the SLA with the downstream Carriers. While end-to-end measures are available for each operator involved in the delivery chain, there is no need for complex, costly and time consuming methods of exchanging measurements among the operators (including Service Providers).

- From a commercial point of view, it has to be considered that the NER commitment with the previous operator is normally managed downstream using different providers, which might deliver different values of downstream NER, and not back-to-back with a single Carrier. The Carrier usually takes a commitment on an average NER value, considering all the traffic managed for a specific SP and for a specific destination. Each Carrier usually controls the downstream quality for each downstream provider, and has to assure the average value on the total traffic to the SP.

- The back-to-back mechanism by itself does not guarantee a Carrier committed to a SLA against possible penalties, but in any case carriers have to manage the commercial risk related to the combined effect of aggregating traffic from multiple sources toward a given destination, and using multiple providers for the same route for efficiency reasons. As described before, there are potential circumstances where, even though all the downstream SLAs are met and may even be above the thresholds, their average value, for a specific SP, may be under the upstream threshold for which the Carrier is committed.

6.1.3 i3Forum recommendations on NER

➔ The measure of NER does always include the destination SP network degradation, but it does not include any failure related to the end-user behavior. A commitment on this parameter should include a commitment from the terminating Service Provider.
➔ A Service Provider/Carrier can measure the NER delivered by its providers, having all information needed to measure the downstream NER delivered by the following parties in the chain.
➔ In addition to the current NER measure, each Carrier can use internal information (network statistics and CDRs) in order to elaborate its own measurement for the NER commitment with the previous player in the chain.
➔ The measurement of NER has to be computed as an Average on a given period of time (e.g., monthly), which has to be agreed between the parties, but for a long enough period of time to avoid results based on normal variability of network performance.
➔ Where a SLA/SLO commitment is agreed between parties (outside the scope of this document) it has to be recognized that there is always a commercial risk to deal with, as the cascade mechanism does not prevent the committed provider from infringing on the commitment to a customer even if it and its downstream providers have performed to the overall commitment

6.2 ASR - Answer Seizures Ratio

6.1.1 Process of Measurement

- ASR is measured by analyzing information (CDRs) generated by the Call Handling Function.
- The ASR calculation formula utilizes the INVITE Seizures and the number of calls effectively answered in a given period of time.
- The parameter measures the downstream performance, including the destination Service Provider network performance. In addiction, the ASR is also affected by the end-user behavior who ultimately determines the number of answered calls.

6.1.2 Service Value of measuring ASR

- ASR is one of the most common parameter measured, monitored and managed for quality purpose using historical data.
- Because an ASR statistic measures both users behavior and destination SP network performance, which are elements not controlled by carriers, this KPI should be considered only for information use only and not for a SLA.
- Each carrier and service provider has all information needed in order to measure the parameter from its network element downstream to the destination end, and to evaluate its own contribution to ASR degradation.

6.1.3 i3Forum recommendations on ASR

➔ ASR is an important KPI to measure and monitor and has to be evaluated using historical behavior. The computation has to be done on average for all the calls in a given period of time that has to be agreed between the parties and should last enough to avoid considering temporary end user behavior (e.g. monthly). SP and carriers must remember that an ASR toward "corporate numbers" will be very low during the weekend and vacation periods. An ASR towards "mobile customers" who activate voicemail will be very good (high), while the same ASR towards customers that do not activate voicemail will be very poor (low). This user behavior is independent of carrier and Service Providers' network quality. Therefore, the use of ASR results must be done by historical trending analysis for a given destination, rather than by an absolute value within a short period of time.

➔ It has to be recognized that while ASR can not be fully managed by carriers considering it is impacted by end-user behavior and terminating SP's network, this parameter is suitable for information use and should not be considered for SLA.

➔ Considering that each carrier can easily calculate the downstream ASR delivered by the following party in the chain, and can evaluate the ASR it is delivering to the upstream SP/Carrier, if commercially agreed between the parties, a pragmatic approach for a cascading

mechanism can be realized by defining and measuring the KPI for the complete downstream segment, leaving the next carrier to transfer that requirement to the following party in the chain. In this case, the carrier subject to the commitment always takes a commercial risk for the reasons explained above.

6.3 ALOC - Average Length of Call

6.3.1 Process of Measurement

- The information for the length of a call is available for every call in CDRs generated by the Call Handling Function. The ALOC is measured by analyzing CDRs produced for a given period of time and for a specific source-destination route. It can be computed as an average, according to a commercial agreement between the parties (e.g., monthly).
- The measurement of the ALOC parameter is affected by end-user behavior so that a variation of measured values can be related to a change in the user behavior.
- Each player involved in the delivery chain can collect all the information needed to compute, monitor and control the ALOC parameter.

6.3.2 Service Value of measuring ALOC

- ALOC is a parameter traditionally measured and monitored for quality purpose, despite it being affected by end-user behavior. The measure has to be done using the length of all calls in a given period of time and computing the average value. The value has to be interpreted using historical data collected for the same route. In order to avoid misinterpretation, the period of measurement should last enough to absorb short misalignments due to temporarily end-user activities.
- Because the ALOC measure includes also users behavior and destination SP network performance, which are elements not controlled by carriers, this KPI should be considered only for information use and not for a SLA.

6.3.3 i3Forum recommendations on ALOC

➔ ALOC is a traditional parameter used to monitor the voice traffic and should be continued to be monitored and compared using historical statistics for the same route and over an adequate given period of time in order to avoid misinterpretation due to end-user activities.

➔ It is important to emphasize that carriers should not be considered fully responsible for misalignments of ALOC due to the fact they do not completely manage all elements impacting the ALOC value, which is indeed also affected by the end-user behavior and the terminating SP's network.

➔ The ALOC parameter can be monitored and reported for information and it should not be used for a SLA commitment.

➔ ALOC can be measured by all Carriers/SPs involved in the voice provisioning chain and thus there is no need under normal conditions to exchange ALOC values among Carriers and Service Providers, unless there is a misalignment with the expected value

6.4 PGRD - Post Gateway Ringing Delay

6.4.1 Process of Measurement

- The delay is available in CDRs generated by the Call Handling Function for every call. The KPI is measured by analyzing the available CDRs.
- The PGRD can be computed as an average, in a given period of time (e.g., monthly), is interpreted using historical data for that specific destination and subject to an agreement between the parties.
- The measurement includes the terminating SP's network performance but it is not affected by end-user activities

6.4.2 Service Value of measuring PGRD

- Every Carrier/SP involved in the traffic delivery chain can measure the Post Gateway Ringing Delay parameter, having received via signaling all information needed to compute the KPI.
- A Carrier should not be considered fully responsible for the PGRD performance, while the resulting value considers also the terminating SP's contribution, unless this entity is also included into a back-to-back SLA/SLO commitment.

6.4.3 i3Forum recommendations on PGRD

➔ PGRD parameter can be computed by all Carrier/SPs involved and it measures the downstream contribution to the delay.
➔ This KPI has to be evaluated in a given period of time and controlled using historical measurements
➔ A carrier should not be considered fully responsible for the resulting value, unless the terminating SP is also committed to the end-to-end result.

6.5 MOSCQE / R-factor - Mean Opinion Score

6.5.1 Process of Measurement

- In the IP environment, MOS is measured by analyzing CDRs generated by the Session Border Controllers and/or Call Handling Function from the measuring equipment to the terminating RTP/RTCP end-point in both directions
- MOS is defined as a subjective parameter which assigns a score between 0 and 5 to the quality performance of a telephone call, representing an estimation of the voice quality perceived from the end-user.

- Operators may obtain an objective evaluation of this parameter through a transmission rating model (E-model) which represents voice quality as an R-Factor, accounting for transmission impairments including lost packets, delay impairments and codec. The MOS is an estimation of the voice quality and can vary from the user perception.
- Carrier may obtain a MOS measurement for every call, using information generated by Session Border Controllers and/or Call Handling Function when the RTCP is activated; these measurements can be statistically represented with reference to a defined period of time, source-destination path, and can be specific for a given provider
- The measurement could be widely affected by the availability of RTCP along the RTP flow and by the location of the SIP end-point (e.g. a media gateway, an end user SIP terminal, a codec translation) which can not be dynamically determined. This implies the measurement should be considered meaningful only in a complete Voice over IP environment (no TDM-IP translation) and in a tested and verified voice path. It is also to be noted that the voice path is determined by the location and routing decision of the voice equipment: it has to be considered normal that the voice path can be different and with a longer distance than the IP best effort path. It is to be noted that it is not possible to detect on a per call basis if a carrier blocks the RTCP flow, or when the call is trans-coded or passed onto TDM to reach its destination. Therefore, it is important to use the MOS on routes that have been initially tested between partners in a trusted commercial environment.

6.5.2 Service Value of exchanging MOS

- MOS is the only parameter able to measure the RTP voice quality and the measurement can be performed also in an IP context using specific elements of the IP network architecture.
- Meanwhile, it has to be emphasized this is possible only if a number of specific conditions are met: RTCP active along the path with all involved players committed to not block the protocol; Border Function located close to the network borders; no trans-coding involved and end-to-end IP transmission.
- Considering the above pre-requisites, the MOS measure could include the performance of the terminating SP network, and also of the end user terminal, which are out of the Carriers range of responsibility.

6.5.3 i3Forum recommendations on MOS

➔ i3Forum supports the effort of introducing MOS measurement in a Voice over IP environment, but strongly recommends the measurement to be restricted to a trusted, tested and all IP path in order to avoid misinterpretation of the resulting statistics.

➔ Special care should be taken in considering the network architectures, trans-coding and TDM break-outs, the availability of the required RTCP along the entire path.

➔ While some limited use of the measurement is to be expected, i3Forum suggests, for the time being, the MOS parameter not yet to be used for SLA, and only in a direct interconnection or in a trusted and tested multi-carrier configuration. As soon as the conditions for a more reliable measure exist across the industry, i3Forum will reconsider this position.

➔ The suggested metric for MOS should be percentile, meaning the percentage of calls in a given period of time the MOS per call/per destination/per customer is equal or better to an agreed

value between the parties. For instance 80% percentile for MOS=4 would mean that 80% of calls to destination Z over a month has the MOS per call equal or better than 4.

7. List of Transport Parameters

7.1 Round-Trip Delay, Jitter, Packet Loss

For definitions of these parameters, please refer to the Annex 1 or the relevant i3Forum technical documentation [2].

The following parameters are usually considered relevant for transport at the IP level. Theoretically, these parameters influence the RTP voice quality in both directions, and they are used in the MOS calculation. However, the control of the IP transport quality is necessary but alone is not sufficient to achieve a desirable level of voice service quality. In any case, an acceptable level of these parameters is a pre-requisite for an acceptable level of the overall voice quality.

7.1.1 Process of Measurement

Two possible processes are available to measure IP transport parameters according to the involved network elements: using IP Routers or CDR produced by Border Function.

7.1.1.1 Measurements through IP Routers

This type of measurement is usually done by injecting synthetic traffic (probes) at IP Edge Routers, for every route and every class of service. Using voice traffic and an adequate Class of Service (conversational) it is possible to measure the IP parameters for Voice traffic. Every Carrier involved can measure its standard contribution from edge to edge (taking into consideration the specific interconnection points). This static measure (even when done on a regular basis) represents performances that are not directly related with the voice traffic transported in real conditions, or with the real voice path of the IP transport, but only refers to the synthetic traffic of the probes.

For instance, in the case of a voice customer that interconnects to carrier 1 using IP in New York and buys voice termination to Ireland. It is possible at the IP layer to measure the IP quality directly from New York to Ireland. The path that will be measured will depend on the IP Border Gateway Protocol (BGP) routing, which in most of the cases is the shortest available path from NY to Ireland. However, the actual voice path will depend on the location of the voice network elements involved in the actual call. This voice path of an actual call can be different and will generally be longer than the shortest BGP IP path. Consequently, it is important to remember that IP layer quality is not alone representative of the voice quality if it is not measured on the voice path.

This type of measurements also reflect IP network performances in standard conditions (i.e. no congestions), while the performances in real conditions may be very different. In addition this measurement can be used only to determine the standard contribution of each single network. Thus in order to estimate an end to end performance further collaboration (post-measurement) is required and could be very complex especially in a multi-provider environment. Consequently, taking into account all mentioned aspects and the nature of this type of measurement, they can be shared, when commercially agreed, for information use only and it is not recommended for use for a SLA.

7.1.1.2 Measurement through Border Function

In this case, the measurement refers to the real voice traffic and is realized with an analysis of CDRs produced by the involved Border Function and referred to the real voice traffic under consideration.
The measurement can be made for every RTP flow related to any call and refers either to the downstream or to the upstream segments. Every measure is observed from the Border Function up/down to the SIP end point, assuming the RTCP is allowed, including any other network segment involved, and not limited to the specific network segment where the measure is realized.

All Carriers/SPs in the chain can then measure IP parameters for the complete downstream/upstream path with a sufficient level of confidence to measure and manage end-to-end values, if the SIP end points are located at the origin and destination SP borders (i.e., media gateways). This method requires: the RTCP protocol to be active in the entire chain; no IP/TDM conversions and no trans-coding in the entire call path.

While the RTCP information is collected by the Border Function, and its position in the network depends on the specific network architecture (centralized or distributed at edge), which may be very different from Carrier to Carrier, it can happen that a single downstream or upstream measure does not consider a portion of the transport network in the network segment where the measure is realized.

7.1.2 Service Value of measuring IP Transport parameters

With reference to the previous methods of measurement the IP parameters, i3Forum strongly recommends the approach of measuring IP transport parameters using Border Function, and considers this the only reliable solution for voice traffic. The value of this method is:

- The measure refers to the real voice traffic and not to synthetic injected traffic.
- The measure refers to the real route of both RTP flows related to the voice traffic and not to a standard BGP IP route.
- All parties can compute the end-to-end value, monitor the downstream provider, and control the performance provided to the upstream customer

7.1.3 i3Forum recommendations on measuring IP Transport parameters with Border Function

- ➔ The i3Forum recognizes the importance of measuring and controlling IP Transport parameters (RTD, Packet Loss and Jitter) for voice traffic.
- ➔ The parameters have to be computed in a given period of time, agreed to between the parties, and provide an average or percentile value for RTD, Packet Loss and Jitter.
- ➔ Considering current alternatives, i3Forum endorses the use of Border Function for measurement able to assure values related to the real voice traffic. This can also allow end-to-end monitoring.
- ➔ i3Forum recognizes the current network implementation and the IP migration might limit a widely use of this approach, but suggests this method should be used in carrier-to-carrier interconnections. i3Forum strongly supports the industry acceptance of this measurement method and will continue to monitor its introduction.

8. Call Attributes

8.1 CLI Transparency

8.1.1 Process of Measurement

- A Carrier can transport CLI across its network only if the upstream network has presented a CLI. Therefore, a carrier can not guarantee that all the calls will have a CLI (given it is possible the CLI was not received) but can only guarantee to deliver the same CLI when a CLI was presented.
- In this case, measuring CLI Transparency means to compute the percentage of calls with CLIs received from upstream network and delivered unaltered (transparently) to the downstream network
- This measurement can only be done for all calls in a given period of time, and for a specific source-destination route, in a specific internal network segment, excluding the possibility to measure an average value down to the end of the delivery path.
- When end-to-end CLI Transparency monitoring is required, CLI probing is a solution that can be used to check CLI delivery from the measuring point down to (and including) the terminating Service Provider. It is usually done using automatically generated traffic and it is used usually to monitor the CLI delivery through other networks down to the destination.
- For this reason, the result of CLI probing can only be an indicative value of the CLI delivery ratio, because it does not take into consideration the holistic reality of the terminating service provider's network, e.g. roaming probes are stationary and cover only a limited part of the network.

8.1.2 Service Value of measuring CLI Transparency

- CLI transparency is usually associated with premium route traffic with higher quality requirements. For this reason, the internal network segment measurement is required to monitor and guarantee its own network performance, which is under its full control.
- From a commercial point of view, a customer would require an end-to-end transparency delivery, which is for the time being impossible to assure, considering other players are involved and the performance is not fully manageable.
- In this case, the use of a CLI probing solution for high quality routes can be required in order to control, on a sampling measurement, the end-to-end quality.

8.1.3 i3Forum recommendations on CLI Transparency

➔ CLI transparency in a single network segment and for specific route should be measured by carriers, when commercially required, providing the percentage of delivered calls with CLI unaltered. The given period of time used for the measure should be agreed between the parties.

➔ CLI probing can be used to control end-to-end delivery, but considering the limitation of the measurement method and the limited reach of responsibility.

9. Other Parameters

9.1 MTRS: Maximum Time to Restore the Service for simple telephony

9.1.1 Process of Measurement

The KPI measure the Maximum Time to Restore the Service (MTRS) for simple voice telephony is defined as the ability to dial an E.164 number towards a purchased destination and receive a normal result which can be a pickup answer, a voice mail, a busy answer or no answer (parties should find a testing number that results in a pick-up or voicemail answer, this can usually be the automated clock service in the country of destination). The actual audio quality of the call is out of scope. MTRS is the maximum timeframe between the moment a trouble ticket in Carrier's trouble ticketing system is opened and the moment when the problem is resolved.

The process of measurements is based on the analysis of time elapsed between the opening and closing of the trouble ticket.

9.1.2 Service Value of measuring MTRS

- MTRS allows monitoring of the operational efficiency of a Carrier in solving fault efficiently.
- It is mostly related to the back office performance to restore a fault and to the network and voice service ability to find backup routes and have sufficient volumes to avoid congestion.
- The operational implementation and identification of traceable problems should be agreed between the parties in order to achieve a uniform and standard categorization of faults, and an undisputable attribution of responsibility.

9.1.3 i3Forum recommendations on MTRS

➔ MTRS monitor and reporting is suggested to be used in carrier-to-carrier and carrier-to-service provider relationships.

➔ Commercial agreement is needed to define case-by-case the faults to be monitored and the attribution of responsibility.

ANNEX 1 – Definition of quality parameters

An extract of the i3Forum document [1] "Technical Interconnection Model for International Voice Services - (Release 2.0) May 2009" is included hereafter for some of the parameters relevant for QoS from the service point of view.

Parameters relevant to the transport layer

Round Trip Delay

Round Trip Delay is defined as the time it takes for a packet to go from one point to another and return [4].

Jitter

Jitter is the absolute value of differences between the delay of consecutive packets [4], [5]**.**

Packet loss

Packet loss is the ratio between the total lost packets and the total sent packets over a given time period [4].

Parameters relevant to the service layer

For the following parameters en-bloc signaling is assumed. The case of overlap signaling is out-of-scope.

MOS_{CQE} / R-factor for voice calls

MOS (Mean Opinion Score) is a subjective parameter defined in ITU-T Rec. P.10 [6] as follows: "*The mean of opinion scores, i.e. of the values on a predefined scale that subjects assign to their opinion of the performance of the telephone transmission system used either for conversation or for listening to spoken material*."

ITU-T Rec. G.107 [7] defines an objective transmission rating model (the E-model) for representing voice quality as an R-Factor, accounting for transmission impairments including lost packets, delay impairments and codecs. The impairment factors of the E-model are additive, thus impairments from different network segments may be added to obtain an end-to-end value.

The R-Factor may be converted into an estimated MOS which is called MOS Communication Quality Estimated or MOS_{CQE} (as defined in ITU-T Rec. P.10 [6]) using formula in ITU-T Rec G 107 Annex B [7]. As a result, MOS is thus an actual user opinion score, and all measurements done by equipment (including R-Factor and MOS_{CQE}) are estimates, and may differ from what actual customers would perceive.

ALOC

Average Length of Conversation (ALOC) expresses the average time in seconds of conversations for all the calls successfully setup in a given period of time. In a TDM environment ALOC has been defined in ITU-T Recc.E.437 [8]:

$$ALOC = \frac{\Sigma \text{ time periods between sending answer and release messages}}{\text{--}}$$

Total number of answers

In a Voice over IP environment, and for the purpose of this document, ALOC is defined as follows:

- SIP protocol: ALOC is measured from the time of SIP 200 OK (in response to an INVITE initiating a dialog) to the time of call release (SIP BYE).

- SIP-I protocol: ALOC is measured from the time of a SIP 200 OK with an encapsulated ANM to the time of receiving a BYE message with encapsulated REL.

ALOC depends on the user behaviour.

ASR

Answer Seizures Ratio (ASR) expresses the ratio of the number of calls effectively answered in a given period of time against the number of call session requests in that time. In a TDM environment, ASR has been defined in ITU-T Rec. E.411 [9] with the following formula:

$$ASR = \frac{\text{Seizures resulting in answer signal}}{\text{Total Seizures}}$$

In a Voice over IP environment, and for the purpose of this document, ASR is defined as follows:

- SIP protocol: ASR is the ratio between the number of received 200 OK (in response to an INVITE initiating a dialog) and the number of sent INVITE initiating a dialog.
- SIP-I protocol: ASR is the ratio of the number of received 200 OK with an encapsulated ANM (in response to an INVITE with an encapsulated IAM initiating a dialog) to the number of INVITE sent with an encapsulated IAM.

ASR depends on the user behaviour.

NER

Network Efficiency Ratio (NER) expresses the ability of a network to deliver a call without taking into account user interferences (measure of network performance) in a given period of time. In a TDM environment, NER has been defined in ITU-T E.425 [10] released in 2002 with the following formula:

$$NER = \frac{\text{Answer message or user failure}}{\text{Total Seizures}}$$

Note: user failure includes caller abandonment

In a VoIP environment, and for the purpose of this document, NER is defined as follows:

- SIP protocol: NER is the ratio of the number of received responses amongst the following responses, with the number of sent INVITE initiating a dialog:
  - a response 200 OK INVITE or
  - a BYE response or

- a 3xx response or
- a 404, 406, 410, 480, 484, 486, 488, response or
- a 6xx response
- a CANCEL message (in forward direction i.e. from the calling party)

- SIP-I protocol: NER is the ratio of the number of received responses amongst the following responses, to the number of sent INVITE with an encapsulated IAM:
  - a response 200 OK INVITE with an ANM encapsulated or
  - a '410 GONE' with REL encapsulated and cause value 22 or
  - a BYE response or message type '486 Busy Here' or message type '600 Busy everywhere' with REL encapsulated and cause release 17 or
  - a BYE response or message type '480 Temporarily unavailable' with REL encapsulated with cause value 18 or 19 or 20 or 21 or 31, or
  - a BYE response or message type '484 Address Incomplete' with REL encapsulated with cause value 28 or
  - a BYE response or message type '404 Not Found' or message type '604 Does not exist anywhere' with REL encapsulated with cause value 1 or
  - a BYE response or message type 500 'Server Internal Error' with REL encapsulated with cause value 50 or 55 or 57 or 87 or 88 or 90.
  - a CANCEL message (in forward direction i.e. from the calling party)

Note: it is recognised that cause value 53 (outgoing calls barred within CUG) has to be considered as a user failure. Being the scope of this document limited to international interconnection it is assumed that no SIP message related to this cause value 53 will be received.

PGRD

Post Gateway Ringing Delay (PGRD) expresses the time elapsed between a request for a call setup and the alerting signal for that call. In a VoIP environment, and for the purpose of this document, PGRD is defined as follows:

- SIP protocol: PGRD is the average time between sending an INVITE initiating a dialog and the first received 18X message;

- SIP-I protocol: PGRD is the average time between sending an INVITE initiating a dialog with an encapsulated IAM and the first received 18X message with an encapsulated ACM.

Note: only INVITEs initiating a dialog for which an alerting response is received are taken into account.

ANNEX 2 - Impacts on Release Causes when moving from TDM to VoIP

Definition and use of Release causes

A release cause is a network signaling (ISUP, SIP, SIP-I, …) message send between carriers and service providers when a call (successful or unsuccessful) is terminated / released to indicate to the other party the reason / cause of the release.
Typical release causes are:

- Normal termination when the phone is put on hook at the end of the conversation by the called or calling party
- Ringing but no answer
- User busy
- Number dialed unknown or not assigned
- Service / codec not supported
- Network technical issues like saturation, switch failure, etc…
- Etc…

These release causes are very important for operators and carriers as they are used for quality management, routing optimization and for communication/signaling towards the called / calling party.

Below are some examples:

- When a User busy RC is received the carrier won't try to find another route but send the "user busy" signal/message to the calling party
- When a "Number dialed unknown or not assigned" RC is received the operator will indicate this information to the calling party by means of a specific tone or message
- When a saturation RC is received this could be a trigger for a carrier to try to setup the call via an alternative route. This can be done dynamic (multiple route option implemented in the switch) or static (change the routing table of the switch). This later solution has strong commercial implications as a customer stops sending traffic towards a provider based on the perception of quality given by a release cause.
- Etc…

Release causes are also used for the calculation of quality of service (QoS) performance indicators l(e.g. NER). These indicators are used by carriers and operators to monitor the overall QoS performance of their suppliers towards certain destinations. For example in the computation of the NER, every RC indicating a network problem will lower the NER value, where a RC indicating a user behavior "no answer" or "user busy" won't lower the NER value.

Due to different mapping in a variety of protocols that are used today, it can happen that RCs could be misunderstood by carriers & operators. In this case, it can result in a misleading perception of the operational performance of the global voice infrastructure. For instance, a "no answer" situation that would be communicated or/and understood as a "network saturation" problem would target carriers to try to find other routes to set-up the call. This behavior would be of no use of course and will actually cause the propagation of the problem to multiple networks. Consequently, the RC's have to be

correctly used by all parties generating signaling and they have to be transparent and coherent in their meaning if multiple networks are in the chain of a call-flow.

Generation and treatment of RCs

In a normal situation all RC's are generated by the calling and/or called party. The transit networks in between the originating and terminating network are passing the RC's in a transparent way.

This can happen differently in the following situations:

- A network in the chain between originating and terminating network stops the call-flow for saturation reasons or because it can not interpret the dialed number, etc. As such this transit network becomes the terminating network of that call. In that case, this network will generate the RC indicating the problem about why the call couldn't be transited and send the information back down the chain to the originating network.

- The signaling used at the ingress site of the network is different from the egress site of the network. This is the case when the RC is received via ISUP signaling (TDM) but must be forwarded via SIP signaling (VoIP). This situation is often the case and will be more prevalent as the industry is migrating from TDM to VoIP and will be globally in a hybrid situation for many years to come. In that technology interworking the network will generate a new RC that should mean the same as the received RC, this is also named RC mapping. When this mapping is done in an incorrect way, the mapped RC will be misinterpreted down the chain and this would result in operational issues as explained before.

Actual SIP-ISUP standards and RC resulting problems

The industry is today using 2 mapping standards to translate ISUP RC's to SIP RC's and vice versa: IETF RFC 3398 and ITU-Q.1912.5. Those mappings are the standard ones implemented by switch vendors in most VoIP switches.

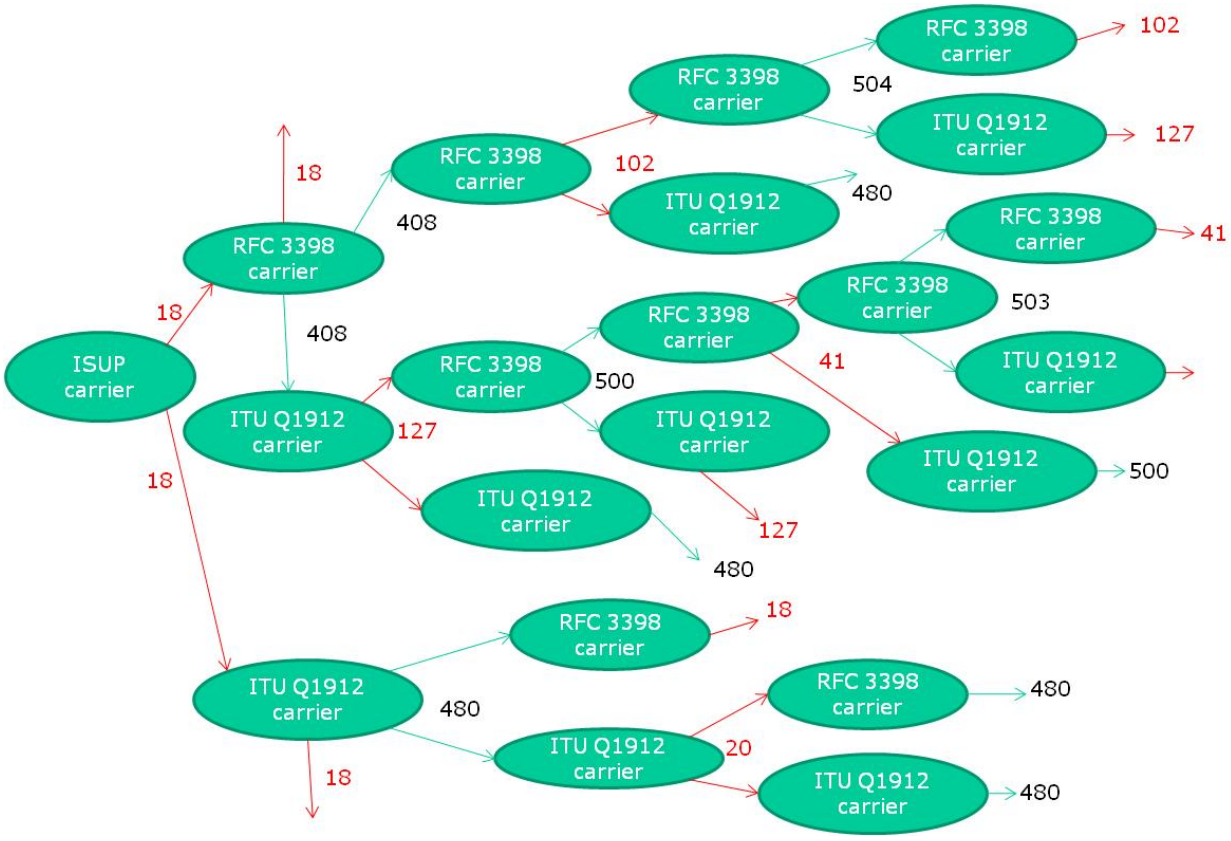
These two mapping standards do however cause issues of different nature:

- Loss of detail: The ITU mapping does map a lot of ISUP RC's to a SIP 500 and a lot of SIP RC's to ISUP 127 making it sometimes difficult for operational departments sitting behind these mappings to identify and resolve certain issues because a 500 or 127 can as such mean a lot of different things.

- Transfer of wrong information: This is more critical as it will drive wrong QoS reporting and inefficient operational actions. It is mainly happening when multiple carriers are in the chain executing a mapping. But it can also happen and cause issues on a peer-to-peer basis if the 2 parties are interpreting the RC-values in a different way which is in especially possible for SIP where the operational use and description of each RC-value is not 100% clear and interpreted the same way for everyone.

The illustration below shows an example of what can happen with multiple parties in the chain using the 2 standards mapping; This example describe what can happen when the terminating party sends an

ISUP RC 18 meaning “No answer” which is a perfectly acceptable situation, and should normally have no impact on NER and shouldn’t drive any operational actions.



As shown in the drawing, where an 18 value should be transparently be passed down the chain as an 18 towards ISUP carriers or a 408/480 towards SIP carriers, it is possible due to a mix of the two mapping standards to end-up with:

- Towards ISUP carriers: 18, 20, 41, 102, 127
- Towards SIP carriers: 408, 480, 500, 503, 504

The RCs 18 and 20 for ISUP and 408 and 480 for SIP are acceptable as they still would be interpreted as a “no answer” cause. But RCs 41, 102 and 127 for ISUP and 500, 503 and 504 for SIP will never be interpreted as “no answer” and will instead be interpreted as a network issue lowering the NER values and even driving operational decisions to reroute the traffic where this of course should never happen for a “no answer” situation.

Many similar examples can occur. If the industry implements randomly one of the two standards mapping, it could create a situation where a number of RCs both in the ISUP and the SIP domain would become unreliable together with the QoS indicators derived from release causes (e.g. NER, ASR).

Carriers and Service Providers requirements

- "Unique" ISUP & SIP RC-values clearly indicating a specific operational / service cause. Unique means each RC is indicating only one and well defined cause and not more than one, which makes the RC-value unspecified & meaningless for operational management.

- Going over multiple carriers in ISUP, SIP or SIP-I the initial cause defined must stay identical. For example, a "No answer" must be at all times and independently from the signaling used (ISUP, SIP-I or SIP) remain as a "No answer".

- For voice quality management a limited number of operational causes need to be defined by carriers for voice quality management. Some of the relevant causes that as a minimum have to be defined by an unique RC values in both SIP & ISUP are:

    - No answer from end-user
    - User Busy
    - Call failure due to end-user or its operator (device out of order, reject, decline, CUG, …)
    - Invalid or incomplete number
    - Unallocated, unassigned number
    - Number changed, redirected to new destination, gone
    - Unsupported media type or service, codec not supported
    - Call rejected by Border Function based on I/C specific configuration parameters e.g. max. sessions reached
    - Call failure due to switch / Border Function / network configuration and/or saturation
    - Unspecified causes
    - Server timeout causes

Possible solutions to tackle the identified issues and make things work

By simply implementing the current ITU/IETF mapping standards, the requirement of operators and carriers to generate and receive correct and useful release causes would not be fulfilled. Consequently, alternative solutions need to be worked-out on and implemented.

The possible solutions identified and recommended by the i3Forum are:

1. **Ignore the SIP RCs and continue to use the ISUP RC values via SIP-I or via SIP with the "Reason Header" enabled**

   The majority of calls are still originated or terminated in TDM/ISUP and correct use of native ISUP RC is perfectly efficient and under control by most carriers within the existing OSS/BSS systems. Therefore, instead of using the new SIP RC-values that can create problems due to the actual implemented mapping standards, the original ISUP RC value can be kept and sent inside the SIP signaling. This is in nature the case for SIP-I. When SIP is used the received ISUP RC value can be copied in the SIP "Reason Header" field. When the destination network is pure SIP and not TDM based, this solution is not going to work. In order to benefit from this solution, carriers/operators need to adapt their OSS/BSS in order to use the ISUP information carried inside the SIP/SIP-I

signaling. As the industry will migrate to full SIP in the coming years, the solution to use the "Reason Header" can only be a temporary solution.

2. Adaptation of the industry standards

Initiatives will be taken by i3Forum and its members to make IETF and ITU aware of the issues with their actual standard in order to come to a new standard that address the mapping issues mentioned in this document. This is however a longer term approach and it will not solve the problem in the short or mid term.

It is outside the scope of this document, but it is possible that in the intermediate period, some pragmatic solutions can appear in order to handle the shortcomings of the default implemented mapping standards, based on bilateral agreements between carriers.

These temporary implementations are not going to modify the standard mapping itself, but only the operational interpretation by a carrier down the chain of a received RC-value (SIP and/or ISUP). Carriers could agree between themselves on how to communicate operational issues to each other via the RC values and as such avoiding potential misinterpretation. In this solution the agreed language, which should be customized from the standards, need to be at all time respected and technically enforced in order to be relevant.

At this stage the i3Forum is not proposing a customized implementation of the mapping standards, but recognizes that besides the use of the "Release Cause" field, carriers can also solve the problem by agreeing on a customer-provider interconnection basis and implementing an ad-hoc mapping.

Regardless of the solution used, it is important that carriers and operators ensure that the network events communicated through the release causes are indeed consistent with the network event.