NICC ND 1433 V1.1.1 (2011-10)

NICC Document

UK-ISUP Automatic Congestion Control Guidelines: Effective Schemes

NICC Standards Limited

Michael Faraday House,
Six Hills Way,
Stevenage
SG1 2AY

Tel.: +44(0) 20 7036 3636

Registered in England and Wales under number 6613589

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to NICC. Pursuant to the NICC IPR Policy, no investigation, including IPR searches, has been carried out by NICC. No guarantee can be given as to the existence of other IPRs which are, or may be, or may become, essential to the present document.

Foreword

This NICC Document (ND) has been produced by NICC Architecture & Requirements (A&R) working group.

1 Scope

The present document provides guidelines which only apply where there is a bilateral agreement between parties to implement the NICC industry standard for adaptive ACC on interconnect routes as documented in ND1007 [1].

This guidance builds on that in section 7 of ND1115 [3] and is based on practical experience of designing, testing, implementing and operating UK-ISUP interconnects with many CPs and switch vendors, with the aim of getting UK-ISUP ACC to work effectively across those interconnects.

These guidelines propose no changes to any existing published documentation.

The guidelines may be of particular use to CPs and switch manufacturers new to the UK market.

2 References

2.1 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For non-specific references, the latest edition of the referenced document (including any amendments) applies.

[1] ISDN User Part (ISUP), ND1007:2007/01.

[2] ITU-T Recommendation Q.764 (Pre-Published 12/99) - Signalling System No. 7 - ISDN User Part Signalling Procedures.

[3] UK-ISUP Overload Controls, ND1115:2001/08.

[4] Flow Control In A Switch Of A Telecommunications Network, World Intellectual Property Organisation, Int. Pub. No. WO 2006/120458 A1, Inventor Philip Williams, Agent Alan Cardus, Priority Date 13.5.2005, International Publication Date 16.11.2006 [A30742 WO publication.pdf].

[5] Communication System, World Intellectual Property Organisation, Int. Pub. No. WO 2006/120462 A1, Inventor Philip Williams, Agent Alan Cardus, Priority Date 13.5.2005, International Publication Date 16.11.2006 [A30774 WO publication.pdf].

[6] Generic Overload Control for use on Interconnect of Next Generation Networks, ND1621 V 1.1.3(2008-08), NICC

3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

ACC	Automatic Congestion Control
ACL	Automatic Congestion Level (UK-ISUP)
CP	Communications Provider
DCS	Deferred Circuit Selection
D&T	Detect and Transmit
GUI	Graphical User Interface
IAM	Initial Address Message (UK-ISUP)
OSS	Operational Sub-System
PEW	Planned Engineering Work
R&R	Receive and Respond
REL	Release Message (UK-ISUP)
RTTM	Real Time Traffic Management
TTB	Temporary Trunk Blocking
UK-ISUP	UK ISDN User Part

4 Introduction

These guidelines have the following four objectives:

1. To focus on the characteristics of effective ACC schemes based on testing and operational experience.
2. To give additional specific guidance on how to implement schemes which will meet the standard.
3. To clarify the recommendations for schemes which, based on testing and operational experience, do not operate effectively.
4. To clarify the alternatives for UK-ISUP interconnect when ACC interconnect requirements are not met.

5 General Description

ISUP ACC as defined in ND1007 [1], consists of 2 parts:

a) Actions taken by an overloaded node (the ‘target node') to inform its neighbours (the ‘source nodes') that it is overloaded, by setting the Automatic Congestion Level (ACL) parameter value in REL messages sent to the source nodes to indicate the severity of overload. ACL = 1 indicates ‘moderate overload’, and ACL = 2 indicates ‘severe overload’. This is known as ACC Detect & Transmit (D&T).

b) Actions taken by a source node to adjust the rate at which it sends calls to the target node when it receives a REL with ACL set to 1 or 2 in response to an IAM. This is known as ACC Receive & Respond (R&R). Note that, for effective control, it may not be necessary for the throttle adaptation algorithm to distinguish between ACL = 1 and ACL = 2.

The adaptive part of the ND1007 [1] ACC requirement relates solely to Receive & Respond. A node which sends an IAM and receives in return a REL message with ACL parameter set to 1 or 2 and with release cause ‘42’ (switching equipment congestion) should measure the rate of such call rejects due to congestion. The measured call reject rate should then be used to adaptively adjust the rate at which each neighbouring node sends calls to the overloaded node. Whilst in overload, the objective is to offer ‘just too much’ traffic, such that the overloaded target will continue to perform close to its maximum effective throughput, whilst rejecting the excess traffic, irrespective of the capacity of the overloaded node and of the number of source nodes causing the overload.

As noted in note 2 to the amendment to section 2.11.2.2.2 of Q.764, specified in section 4.2.2.2 of ND1007 [1], this behaviour can be achieved by a feedback control per route, which adjusts the calling rate a source node offers to a target node so that the measured reject rate due to congestion is brought close to a suitably low target rate. Section 7 of ND1115 [3] and related appendices define these characteristics.

6 Characteristics of Schemes which are effective in an Operational Environment

6.1 Rapid Initial Restriction to a Predefined Configurable Level, Upon Receipt of First ACL

Upon receipt of the first REL + Cause 42 + ACL (1 or 2), the R&R scheme needs to quickly reduce traffic by a significant step. This could be to a predetermined initial traffic rate, or by rejecting a predetermined proportion of calls. This should be applied immediately, in order to facilitate an immediate and sizeable reduction in admitted traffic.

It is not acceptable to wait several seconds until the end of a time interval, before this initial restriction is applied.

The control will normally revise the traffic rate at the end of each defined time interval. Hence in subsequent time intervals, the admitted rate should adapt based on the ACL reject rate in previous time period(s).

As indicated in section 2.11.2.2.1 within 4.2.2.2 of [1], the predetermined initial traffic reduction for ACL2 may be more severe than for ACL1.

6.2 Rapid Convergence to a Small Target Reject Rate (of REL + Cause 42 + ACL (1 or 2) from the Congested Switch)

Adaptive ACC schemes which have been shown to work well in practice, converge to a small (>0) target reject rate (calls/sec) returned from the congested switch, as per the UK spec (section 7 of [3]).

ACC schemes which converge to a target reject proportion are not recommended.

Experience has shown that schemes which converge to a target reject proportion, rather than a target reject rate, do not have a unique solution. Some sources can get starved of capacity, whilst others get a lot.

This target objective, outlined above, should not be confused with the method of restriction. It is possible to use proportional rejection as a restriction algorithm, but converge to a fixed call failure rate.

6.3 When ACC Receive and Respond is Invoked And Rejects Calls, Those Calls Should be Backward Released and Not Re-Routed

Calls subsequently rejected due to overload controls invoking Receive & Respond should be backward released without re-routing (since switch overload is significantly more serious than route congestion). The reason is because re-routing risks multiplying the overload many times and spreading it network-wide with the possible loss of UK network integrity. Collaborative real-time Network Management action may enable specific temporary expansive routing once the nature of the overload is understood.

When calls are rejected by R&R, only a small selection of ISUP release causes are applicable.

These include cause 17 with location TN, or cause 34/42 with location LN, RLN or BI. Release reason 47 is not appropriate for calls rejected by R&R.

6.4 Similar Behaviour for ACL1 and ACL2

It is recommended that the same target reject rate is used for both ACL1 and ACL2. ACL1 and ACL2 might have different initial restriction levels, and possibly different adaptation behaviour, but they should both converge to the same target reject rate.

6.5 Priority Calls Not Rejected

ISUP ACC should not reject a priority call, assuming a low rate of such priority calls.
There may be exceptional circumstances when an overload is due to a very high rate of priority calls. In practice, in such circumstances it may be deemed reasonable to reject some priority calls.

6.6 Gradual Switch Off Of The Control Once The Overload Has Passed

Effective schemes will gradually increase the admitted traffic rate over a period of time when the rate of rejects is lower than the target rate. For example, the admitted rate will be gradually increased if there is a period with no ACLs.
Effective schemes run a timer to determine when control should be turned off completely (after an extended period – multiple time intervals - with no ACLs, and no rejection by the local control).
Under no circumstances should R&R be turned off because a REL is received without an ACL (indeed, normal calls terminating should generate a REL without ACL) as this would prohibit stable control.

7 Specific Information on Implementation of Effective Schemes Which Will Meet the Standard

7.1 How to Realise an Effective Rate Based Leaky Bucket Restrictor, Which Admits Calls in a Stable Fashion

A leaky bucket restrictor has a leak rate (the allowed traffic rate), and a depth (to cope with some traffic variation).
It is important to ensure that the rate based scheme maintains a roughly consistent rate across the whole time period, via use of internal sub-second measurement periods. A scheme would not work well if it admits a spike of traffic in 1 second followed by several seconds of zero traffic. A description of suitable leaky bucket restrictor behaviour is given in Figure 4 of ND1621 [6].

7.2 Recommended Periodicity for Updating the Restriction Level

The control needs to be stable, but responsive. Experience has shown that updating too frequently can introduce instability. It is recommended that adaptation takes place on a time interval of a small number of seconds (typically up to 5 seconds; 1 second is probably too frequent).

7.3 Examples of Adaptation Algorithms Which Have Been Shown to Work Effectively (Including Patent Numbers, Where Applicable)

A wide range of successful adaptive ACC algorithms have been deployed. Most are not subject to patents, and it is possible to comply with the guidance in this document without infringing patents.

There are patents which provide possible ways of implementing the adaptation algorithm.

One example of a public domain scheme is referenced at the end of Section 7 of ND1115 [3].

More recent BT patents which provide ways of implementing such algorithms are given in [4] and [5].

7.4 Scaled Deployment of ACC Receive and Respond

Each separate traffic route to an overloaded switch needs a separate instance of the R&R control. But, for some switch types, it may be possible to define a small subset of restriction profile templates, governed by steps in route size, spanning the full range. The profile template can then be used to define the parameters for all control instances for traffic routes of the appropriate size. This may provide a more scalable and manageable way to deploy ACC.
Data managers will need to check which routing cases are affected and if several traffic routes share a single originating point code, then separate restriction control instances should be set against each traffic route (not just against the point code).

7.5 Operational Observations

Switch alarms should *actively* monitor the UK-ISUP signalling during ACC deployment.

Alarms and analysis of captured data should be put in place so that network management centres can see immediately if ACC is invoked, and the extent of that restriction. It is also important to know when control has ceased.

When ACC is invoked, the affected network management centres should liaise to establish the extent of the problem and to put in place any temporary alternative routing which will safely route around the problem, in the case where it is localised.

ISUP ACC should not normally be invoked by planned engineering works (PEWs). However there have been instances of some switch types which momentarily invoke D&T following a PEW, when the switch is restarted. This in turn momentarily invokes R&R at source switches. ACC activation in such circumstances could be beneficial, if the load on the restarted switch is temporarily high.

In the event of ACC invocation, the network management centres should liaise to establish the root cause of the ACC incident (genuine high traffic or some other reason).

It is recommended that PEWs of this nature are normally performed during the night.

8 Review of the Recommendations for Characteristics Which Experience Has Shown Are Not Effective

8.1 The TTB/DCS Scheme

Experience has shown that schemes based on TTB or DCS are more difficult to configure, and are unlikely to provide stable control. The discussion in section 5.2.1 of ND1115 [3] shows how these schemes take significantly longer to reduce the offered load to the target switch.

On this basis it is recommended that these schemes are avoided, in favour of a scheme which converges to a small target reject rate.

8.2 Sending ACLs with All Release Messages

As soon as the internal overload control on the overloaded switch determines that it needs to reject a call, then that call rejection should be accompanied by an ACL. ACLs should be sent with all calls rejected by the overloaded exchange (i.e. REL + cause 42 + ACL1/2). It is the rate of REL + cause 42 + ACL received which will determine the adjustment to the admitted rate at the source.

In order to avoid confusion, the sending of ACLs with other release messages (not cause 42) is not recommended.

Note that an ACC restriction scheme which conforms to section 2.11.2.2.2 within 4.2.2.2 of ND1007 [1], will ignore such extraneous ACLs.

8.3 The Optional Use of Priorities

The Optional Use of Priorities is described in paragraph 2.11.2.4 in the UK Interconnect Specification ND1007 [1].

The use of such priorities is NOT recommended for use, since it is based on the SCCP traffic limitation mechanism, rather than application level control.

9 Clarification of the Options for Interconnect when Adaptive ACC Interconnect Requirements Are Not Met

9.1 Permanent Route Restriction (e.g. Route Gapping)

If adaptive ACC is not available across a route, then an interim solution is to apply permanent route restriction across the route. The most efficient form of permanent route restriction is a rate based leaky bucket, which gives a steady flow of traffic, even across sub-second time intervals. The rate for permanent route restriction would need to be a margin (typically 30%) above the busy hour rate for the route.

It will be necessary to monitor the permanent route restriction rate as traffic levels change.

This interim solution is not as efficient as adaptive ACC. In order to ensure that a switch is not overloaded it is necessary to partition its capacity amongst all the routes. This may mean that an operator may obtain much less capacity than would be possible with ACC because normally not all routes are simultaneously contributing to an overload.

Permanent route restriction, which has been applied **in lieu of ACC**, must be removed when adaptive ACC is subsequently configured and enabled on those routes.

9.2 Practical Implementation of the Formula for Limiting Route Sizes When Controls Do Not Exist

In the case where neither adaptive ACC nor permanent route restriction are present, section 8.3 of ND1115 [3] provides a formula to calculate how many E1s of unprotected traffic can be safely supported by a switch, before the switch is likely to be pushed so far into overload that it could go into restart.

This limit should then be divided across the all of the unprotected routes connected to the switch.

Experience has shown that an effective way to implement and manage the implications of this formula is to limit the size of every route which does not have either adaptive ISUP ACC or permanent route restriction to two E1s per unprotected route for a switch with capacity of at least 100 calls per second.

Use of route size limits is the least satisfactory solution, and should be viewed as an interim solution.

10 Recommendations for a Scheme to Achieve an Effective Rate Based Adaptive ACC Control

An example of how to develop and deploy an effective adaptive ACC scheme would be to meet the following requirements.

10.1 Functional Description

10.1.1 Detailed Description/Operation

1. Meet the requirements given in section 7 of this document, and 4.2.2.2 of reference ND1007 [1] - specifically, the UK amendment to sub clause 2.11 of reference [2], and the related text in section 4.2.2.2. Note that the mechanism specified in the section entitled ‘Optional use of priorities’ that allows ACC to throttle calls to emergency numbers (999, 112, 18000, 9112, 918000) for high volumes of such calls is not required – see Requirement 6 below.

2. A feedback control per route, which uses a leaky bucket throttle to adjust the calling rate a source node offers to a target node so that the measured reject rate due to congestion is brought close to a configurable goal reject rate (that is configurable per traffic route).

3. Control rapidly settles down to a steady state and thereafter there should be no oscillations in acceptance/ rejection rates large enough to noticeably affect call throughput or response times.

4. Fair allocation of the target node's capacity between several controlling sources: the target’s capacity should be divided between the sources in proportion to their goal reject rates (this is a source’s ‘fair share’). Averaged over 1 minute intervals, the rate each source’s ACC throttle admits calls should lie within +/- 10% of its fair share.

5. To guard against sudden step increases in load during a period of light overload, the adaptation of ACC bucket leak rates should not allow an increase in leak rate during periods when the bucket is rejecting no calls and none of the calls admitted by the bucket are rejected by the target (with ACL equal to 1 or 2 and with cause #42).

6. Call with a CPC marking of "11" will be subject to control but shall only be rejected on the basis of a higher leaky bucket threshold than ordinary calls.

7. Call Throughput maximised at the target (overloaded) node – minimum reduction in throughput as load is increased.

8. 95^{th} percentile response times to IAM at overloaded node no greater than 500 ms in steady state.

9. Fast activation of ACC: the load offered to the target shall be reduced to the leak rate configured for control activation within one second of a sharp increase in load at the target (causing the target to detect it is overloaded).

10. The bucket fill thresholds used by the leaky bucket restrictor to reject calls shall be configurable.

10.1.2 Message Flows

This SoR defines no new message flows. The existing ISUP call flows relevant to ISUP ACC and specified in ND1007 [1] shall be used.

10.1.3 Time-Outs

Any timers associated with the adaptive ACC implementation to meet the requirements of section 10.1.1 of this document shall be defined. That is, for each timer, its valid range of values, default value and method of change if appropriate, shall be defined.

10.2 Admin Control Procedures

The administration shall have the following control features available via the element manager GUI (if any) and via OSS interfaces:

- It shall be possible to enable and disable adaptive ACC throttling on a per-route basis. Note: This refers to the R&R behaviour at the source node.
- It shall be possible to enable and disable ACC overload detection on a per-node basis. Note: This refers to the D&T behaviour at the target node.
- Overload detection parameters at a target node shall be configurable on a per-node basis.
- Adaptive ACC parameters at a source node shall be configurable on a per-route basis. The parameters should comprise as a minimum the following: per route goal reject rate, per route initial leak rate, per route minimum and maximum leak rates and the bucket thresholds for ordinary calls and priority/emergency calls.
- It shall be possible to configure the mechanism so that it does NOT reject calls to emergency numbers and/or priority calls.
- It shall be possible to configure the control to be non-adaptive with any desired bucket leak rate.
- The treatment of calls which fail because of ACC throttling shall be data configurable on a per incoming route basis. It shall be possible to configure the data to implement the call treatment required by the operator.

10.3 Measurements for Network Management Visibility

The measurements specified here are for use in near ‘real-time’ management.

On activation: Each automatic control instance will raise an alarm identifying the control instance and the time, presented immediately to the OSS.

Whilst active: Each control instance shall generate data on the number of requests accepted; the number of ordinary requests rejected and the number of priority requests rejected over every 5 minute interval (see note).

To facilitate Real-Time Traffic Management (RTTM), this data will be reported to the OSS at five minute intervals with latency short enough so that the information is presented in real-time to network management centres within two minutes.

On deactivation: Each control instance will cancel the alarm. Aggregate stats on number of requests accepted; number of ordinary requests rejected and number of priority requests rejected.

Note: The five minute intervals used for data generation should be correlated to the clock face, i.e. 12:00 to 12:05, 12:05 to 12:10 etc. so that all the data gathered represents equivalent and recognisable time intervals.

10.4 Feature Interactions

10.4.1 Management Controls

Adaptive ACC shall be implemented such that it operates independently from other network management controls (such as telephony or IN call gapping).

10.4.2 Overload Controls

Adaptive ACC shall be implemented such that when multiple overload controls are active at a particular host, the controls will still converge to a steady state with the resource allocated to the traffic sources controlled by each overload control being apportioned in a configurable manner.

History

Document history		
1.1.1	Oct 2011	Initial Publication