NICC ND 1657 V2.1.1 (2023-07)

NICC Document

SIP – Overload Control

NICC Standards Limited

c/o TWP ACCOUNTING LLP,
The Old Rectory,
Church Street,
Weybridge,
Surrey KT13 8DE

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 SIP Overload Control Task Group.

1 Scope

The present document provides overload control requirements for UK CPs across all external SIP interfaces (NNI and UNI). Its use in relation to other SIP interfaces is not precluded.

SIP overload control refers to the management of load on network nodes by limiting the rate of SIP requests in order to maintain or maximise successful throughput and to limit signalling delays.

Media congestion is not managed through SIP overload control, although where the media resource is a component of the overall SIP server its throughput may also be protected through these procedures.

2 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.

2.1 Normative references

[1] RFC 3261 SIP Session Initiated Protocol
[2] Ofcom National Telephone Numbering Plan
[3] ND1035 SIP Network to Network Interface

2.2 Informative references

[i1] ND1033 NGA Telephony SIP User Profile
[i2] ND1034 UK SIPconnect Endorsement
[i3] ND1037 SIP - ISUP Interworking
[i4] ND1653 Overload Control for SIP in UK Networks
[i5] RFC 7339 Session Initiation Protocol (SIP) Overload Control
[i6] RFC 7415 Session Initiation Protocol (SIP) Rate Control

3 Definitions, symbols and abbreviations

3.1 Definitions

Congested SIP trunk	A SIP trunk which is attempting to carry in excess of the maximum configured limit of sessions and, thus, will reject new session requests.
Grade of Service	Grade of Service is the probability that calls are lost owing to a lack of network capacity, rather than equipment failure or an engaged terminating station. For example, a Grade of Service of 0.01 represents one call lost in 100 offered.
Node	A point across a UNI or NNI where a decision is made to accept, reject or re-attempt a call.

3.2 Abbreviations

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

CAC	Call Admission Control
CP	Communication Provider
CPE	Customer Premises Equipment
CPU	Central Processing Unit
DDoS	Distributed Denial of Service
DoS	Denial of Service
DSP	Digital Signal Processor
DTMF	Dual Tone Multi-Frequency
GPU	Graphics Processing Unit
IP	Internet Protocol
ND	NICC Document
NNI	Network to Network Interface
PBX	Private Branch Exchange
SBC	Session Border Controller
SIP	Session Initiation Protocol
UK	United Kingdom
UNI	User-Network Interface

4 SIP overload control

This section provides some background to the content of this ND.

4.1 Background

This document presents overload control principles and configuration settings which operational experience has shown will help to mitigate against overloads in SIP networks. The baseline requirements and the mitigations described in this ND are designed to significantly reduce the likelihood of an overload incident having drastic consequences.

The scope of this document is the whole UK SIP network. The overload control mitigations described in this document need to be applied as appropriate on the user to Communication Provider (CP) and CP to CP interfaces (i.e. UNI and NNI).

In order to start to apply these mitigations, the bottlenecks in the network/systems must be understood. The bottleneck may vary depending on the call mix, and may change as the network grows.

The majority of the mitigations mentioned here are low cost and can be implemented using existing features on network devices such as SBCs. It is important that CPs work through these recommendations and implement them, configuring SBCs/devices accordingly to set these mitigations in place.

For NNIs, CPs should wherever possible reach pragmatic bilateral agreements regarding the extent to which re-attempts are permitted, mindful of the worst case scenario when all paths are attempted.

It should be understood that SIP overload control mechanisms (by their very nature) while protecting platform integrity and maximising throughput, will result in some calls being dropped during overload.

Application of the Requirements in this ND depends on the nature of the specific UNI or NNI, i.e:

- Across National NNI, this ND applies both ways.
- For International NNI, this standard applies only to the UK CP side. However, UK CPs are encouraged to request the International CP to also apply the controls where appropriate.
- For UNI, this standard applies only to the UK CP side. However, UK CPs are encouraged to request their customers to also apply the controls where appropriate.

4.2 Comparison to ND1653

The ND1653 [i4] specification was published in 2019 and was based on two existing RFCs. At the time of writing ND1657, as far as is known by NICC, ND1653 had not yet been developed by any vendors, or deployed in any UK CPs' networks. Vendors are reluctant to develop something just for the UK market. It is also considered that ND1653 will be complex for CPs to test across NNI.

Instead, operational experience and learning has led to careful configuration of existing load control measures to mitigate SIP overload, not only across the NNI but also over the UNI. These pragmatic measures are specified in ND1657. ND1657 is standalone and implementable.

CPs may choose to implement ND1653 [i4] as well as ND1657. If a CP deploys ND1653, any other CP who doesn't deploy ND1653 will be regarded as a non-compliant source from an ND1653 perspective; ND1657 on its own does not give ND1653 compliance.

5 Baseline requirements for overload control

Experience has shown there are a number of baseline requirements for successful handling of voice calls under SIP overload conditions. CPs should ensure they have deployed appropriately configured equipment to comply with these requirements. Unless otherwise stated these requirements apply to:

- Calls entering the CP's network
- Calls between SIP nodes within the CP's network
- Calls leaving the CP's network.

These baseline requirements should be considered not only for known bottlenecks in the network but for all routes through the network because, under SIP overload conditions, *any* route can quickly become a bottleneck once calls are re-routed around the initial bottleneck.

The baseline requirements are contained in the following subsections.

5.1 Limit ingress traffic to avoid breakdown of the CP's network

It is important that CPs protect their own networks from catastrophic breakdown caused by a flood of traffic coming into their network. It is recommended that CPs deploy an SBC on the edge of their network to perform this role.

In most networks, there will be two capacities to limit: the calls/second rate (CPU constraint) and the number of simultaneous calls (memory and/or bandwidth constraint).

Within their network, CPs should ensure that their network nodes have sufficient capacity to handle load up to this limit. In large networks, CPs may wish to split their network into several smaller networks with rate limiting SBCs between them.

NOTE: CPs should also consider limiting the amount of egress traffic to protect their peering CPs, particularly to smaller CPs who may not have the same ability as the larger CPs to withstand loads, but this should only be done under bilateral agreement.

Requirement 1: CPs shall deploy an SBC or similar device on the edge of their network, configured to prevent the amount of ingress traffic (calls/second rate and simultaneous calls) exceeding the design limits of their network.

5.2 Maintain throughput near to the nominal limit when subjected to overload

Traffic limiting should be designed so that good throughput of successful voice calls is maintained during overload conditions in order to avoid causing traffic spikes elsewhere in the network which could further degrade throughput. This good throughput should be maintained as stable as possible even when the ingress load increases well beyond the rated traffic capacity. Physical constraints (e.g. CPU) ultimately mean that, in order to gracefully reject the increasing ingress traffic, it will be necessary to reduce the throughput but this should be done with a long tail rather than a sudden decrease.

See section 6.2 Stable overload control

Requirement 2: CPs shall ensure that throughput can be maintained at, or near, the rated traffic capacity even when subjected to overload well beyond that limit.

5.3 Allow existing calls to continue; rejecting only new calls

Under SIP overload conditions, existing calls should not be dropped to alleviate the problem as the participants will likely redial, causing overload to get worse. Furthermore, there are regulatory consequences of dropping emergency calls.

SIP messages within a call should not be dropped because it breaks the SIP protocol flow, resulting in unstable calls and protocol retransmissions that worsen rather than alleviate overload.
See sections 6.2 Stable overload control and 6.5 SIP request prioritisation during overload.

Requirement 3: Under overload conditions network nodes shall allow existing, stable, calls to continue.

5.4 Prioritise emergency calls over other calls

Network nodes should give priority to emergency calls (see the Ofcom National Telephone Numbering Plan [2] and ND1035 [3]) over non-emergency traffic.

See section 6.3 Emergency calls

Requirement 4: Network nodes shall prioritise the set-up of emergency calls over the set-up of non-emergency calls.

5.5 Do not rely on other CPs to protect your network

Some existing SIP overload specifications (e.g. RFC 7339 [i5] and RFC 7415 [i6]) are based on the 'receiving' SIP node communicating with the 'sending' SIP node to tell it to reduce the amount of traffic it is sending. These specifications depend on all SIP networks supporting a common mechanism for this communication. At present there are no CPs in the UK (and possibly none elsewhere in the world) that deploy such mechanisms because there is no advantage to doing so unless every other SIP network does the same; logistics and commercial pressures mean this is unlikely to happen unless forced by strict regulation.

A CP must therefore always have its own mechanisms in place to protect its own networks rather than assuming that all other peering CPs will implement measures to protect it. These mechanisms should exist even if other mutually agreed mechanisms or bilateral agreements are in place.

Requirement 5: CPs shall protect their own network without reliance on the capabilities of other CPs' networks.

6 Maintaining good network service during periods of high demand

When a voice network has reached its maximum capacity, it is inevitable that some calls will be dropped. However, with careful design it is still possible to maintain a good level of service. This section provides guidance for best practice to reduce the impact on the consumers of the voice service when a network is nearing, or has reached, its maximum capacity.

6.1 Headroom capacity

Headroom capacity is the additional network capacity available above that required for normal traffic conditions. It is this headroom capacity which is utilised before reaching a true overload scenario.

6.1.1 Headroom for temporary traffic spikes

Traditional TDM switch capacity was always dimensioned to include some headroom which helps to carry modest overload traffic levels, such as temporary spikes of traffic caused by the stochastic bunching of calls during a normal day.

Similarly, SIP networks should also be dimensioned to include some headroom. It is up to individual CPs to choose the exact margin but a typical recommendation would be that each network node should run at no more than 75% of capacity under normal busy hour traffic levels, leaving 25% headroom for peak events.

Requirement 6: CPs shall ensure that the capacity of equipment deployed in their networks has sufficient headroom to cope with normal traffic spikes.

6.1.2 Headroom for rejecting traffic

For nodes such as SBCs, which protect the network by rejecting excess traffic, the node should incorporate enough headroom so that it is possible to reject many times more traffic than it accepts. As a recommendation, the node should be designed to handle x5 overload (i.e. accepting x1 load and rejecting the excess x4 load).

It is important that the processing cost of rejecting a call is as low as possible. By rejecting early, the rejection is much cheaper than handling a complete call, meaning minimal extra resource is required to handle it. However, it is important to get far enough through the processing stack so that emergency calls are identified and handled appropriately (see section 6.3 Emergency calls), and that only whole new calls are subject to rejection, rather than rejecting SIP messages relating to existing calls in progress (see section 6.5 SIP request prioritisation during overload).

Requirement 7: CPs shall ensure that nodes which reject traffic are dimensioned to handle the rejection of excess traffic well beyond normal load levels.

6.1.3 Headroom for cloud deployments

There will be commercial pressure to reduce headroom, particularly when CPs deploy network nodes in public and private clouds where it is easier to spin up additional capacity on demand. However, CPs should be cognisant that “on demand” does not mean “instant”. It can take many minutes to spin up a new node, even if it is done automatically, by which time the spike could have

passed. Also, even public clouds don't have limitless capacity and during periods of heavy demand requests to allocate more resources may be temporarily blocked.

Requirement 8: CPs should ensure there is headroom rather than relying on spinning up capacity on demand as the primary solution for preventing overload.

6.2 Stable overload control

During overload conditions, the priority should be to deliver a high, and stable, quality of service to established calls. If this is not achieved then customers will hang up and re-dial, which would increase overload further.

To that end, network nodes should reject new work such as new SIP dialogues, rather than rejecting individual SIP messages, to avoid impacting established sessions. Existing work (dialogues in progress) should be prioritised over new work (new INVITEs, SUBSCRIBEs, REGISTERs). See section 6.5 SIP request prioritisation during overload.

The SIP protocol is designed to be resilient to temporary failures by automatically retransmitting failed protocol messages. If messages within a SIP dialogue are dropped then these automatic retransmissions will result in yet more SIP messages being sent, traffic increasing and overload getting worse. Also, if SIP messages such as BYE are rejected then existing calls, and the resources they are using, may never be released.

Overload controls should be configured to accept up to a target rate of new dialogues, and reject the excess, whilst continuing to process as close as possible to the target rate. Hence, the throughput of the network as a whole is maintained as high as possible, until the capacity of the first bottleneck is reached. Only the excess load is then subject to re-attempt amplification.

Overload controls which reach a threshold and then reject all call attempts for an extended period should not be implemented. Such a control reduces total system throughput. Although this gives immediate protection to the next node in the network, the wider implications have been shown to be devastating during network-to-network overloads.

Requirement 9: During overload conditions network nodes shall reject out-of-dialogue SIP messages in preference to rejecting, or discarding, subsequent messages within a SIP dialogue. For voice calls this means rejecting new call attempts whilst allowing existing calls to continue.

6.3 Emergency calls

Requirement 4, section 5.4 requires that network nodes shall prioritise emergency calls over non-emergency calls.

Overload controls should be priority call aware. Higher priority should be given to emergency calls so that emergency calls will only be rejected when all new non-emergency calls are being rejected and there is still not enough capacity in the network to handle new emergency calls.

However, new emergency calls should not pre-empt existing stable non-emergency calls.

There are many different methods for giving priority to emergency calls, including;

- Expediting emergency calls to the front of traffic queues.

- Having a higher threshold before emergency calls are rejected compared to non-emergency calls.
- Reserving a portion of trunks for emergency calls.
- Routeing emergency calls via specific network nodes and routes that are dedicated to emergency calls.

Any of these methods are acceptable. However, all except the first of the above require traffic analysis to ascertain exactly what threshold to set, or how much space to reserve for emergency calls and are potentially wasteful in terms of reserving unused resources.

NOTE: Continuous Retry was used to provide additional measures to ensure the delivery of emergency calls in TDM networks, by persistently re-attempting the call at the originating local exchange. Continuous Retry should not be used in SIP networks.

Requirement 10: Network nodes shall not pre-emptively drop existing stable (non-emergency) calls in order to accept a new emergency call.

NOTE: For clarity, Radio Access Networks are out of scope for Requirement 10.

6.4 Call mix

Different calls may require different amounts, or different types, of resources. For example, if a call requires transcoding or DTMF interworking then it may require significantly more CPU capacity, or it may require special ring-fenced hardware (e.g. DSPs, GPUs), compared to other calls.

These resources should be taken into account when deciding whether to accept or reject a new call. If resources are tight then consideration should be given to whether to allow the call to continue but with lower resource usage (for example, by negotiating a codec that avoids transcoding).

SIP applications may request to change resources in mid-call (with re-INVITE or UPDATE), for example to uplift a voice call to G.711 for fax. If resources are tight then consideration should be given as to whether it is better to reject the change but allow the call to continue (by responding with 488), or to drop the call completely (with BYE). The appropriate answer will often depend on the application, or on the source or destination of the application. As an example, an emergency voice call that attempts to uplift to video should be allowed to continue as voice only without video, whereas it might be better to drop a voice call that attempts to uplift to fax.

Requirement 11: CPs shall ensure that they take into account all the different types of resource shortages that can occur in their networks when dimensioning their network nodes; and whether, in some cases, it is preferable to force a call to use less resources rather than rejecting it outright.

6.5 SIP request prioritisation during overload

As discussed in the previous sections, during overload conditions it will be necessary to reject some SIP requests but care must be taken to prioritise which SIP requests are rejected in order to reduce the impact of rejection.

Table 1 below shows the assignment of priority values which results from the application of the principles in the previous section where:

- 1 is the highest priority and is applied to in-dialogue methods [i.e. those that are the last candidates for rejection to be applied to].
- 2 is the next highest priority and is applied to out-of-dialogue methods that are associated with emergency calls.
- 3 is the lowest priority and is applied to out-of-dialogue methods that are not associated with emergency calls [i.e. those that would be the first candidates for rejection to be applied to].
- The list of SIP methods in table 1 is more extensive than those specified to provide basic voice services in ND1033 [i1], ND1034 [i2] and ND1035 [3]. It is intended to give guidance on handling the methods under overload conditions and should not be taken as extending the behaviour of the UNI and NNI specifications.

Request Method	Within Dialogue?	Emergency call	Priority Level
ACK	yes	no	1
		yes	1
BYE	yes	no	1
		yes	1
CANCEL	yes	no	1
		yes	1
INFO	yes	no	1
		yes	1
INVITE	no	no	3
		yes	2
	yes	no	1
		yes	1
MESSAGE	no	no	3
		yes	2
	yes	no	1
		yes	1

Table 1: SIP request priority levels (continued on next page)

Request Method	Within Dialogue?	Emergency call	Priority Level
NOTIFY	no	no	3
		yes (note 1)	2
	yes	no	1
		Yes	1
OPTIONS	no	no	3
		yes (note 1)	2
	yes	no	1
		yes	1
PRACK	yes	no	1
		yes	1
PUBLISH	no	no	3
		yes (note 1)	2
REFER	no	no	3
		yes	2
	yes	no	1
		yes	1
REGISTER	no	no	3
		yes (note 1)	2
SUBSCRIBE	no	no	3
		yes (note 1)	2
	yes	no	1
		yes	1
UPDATE	yes	no	1
		yes	1
Note 1: Out-of-dialogue NOTIFY, OPTIONS, PUBLISH, REGISTER or SUBSCRIBE messages cannot, strictly, be deemed to be associated with an 'emergency call'. However, they could still potentially be treated as 'emergency' messages if the request can be identified as being of an 'emergency' type e.g. if it has the Resource-Priority header associated with it.			

Table 2: SIP request priority levels (continued from previous page)

Requirement 12: When limiting traffic during overload conditions, network nodes should accept or reject SIP requests following the priority allocations specified in Table 1.

6.6 Load testing

It is important that CPs test their network overload controls, both under mild and severe overload conditions. Based on historical operational events, it is recommended that CPs test for at least five times the rated traffic capacity for an SBC / network device. Testing of overload controls can be conducted in the CP's test lab or reference model. Alternatively, test evidence can provided by the solution vendor, such as an overload test demonstration or overload test report.

Requirement 13: CPs shall ensure that all overload controls deployed are tested to ensure they work as intended.

7 Overload control and re-attempt strategies

This section unpacks the baseline requirements described in section 5 into more specific requirements covering rate limiting, contention, re-attempts and response codes. It also considers the interlock with destination overload control.

7.1 NNI overload control

It is expected that CPs may operate NNIs with an element of contention for ports and calls/second processing capacity. CPs will monitor this, and adjust capacity over time if the contention ratio changes, to maintain the engineered Grade of Service for interconnected CPs.

There should be controls limiting the concurrent calls and calls/second which the CP can admit for each NNI. Also, there shall be a higher control limiting the total concurrent calls and calls/second which can be supported across all of the CP's NNIs (mindful of any contention ratio). This is known as Nested Controls. For example:

- If there are 4 NNIs each with 1000 calls/second, but the CP has a 2000 calls/second overall limit:
    - The individual NNI overload controls should restrict calls when 1000 calls /second is reached for that specific NNI.
    - The overall control shall restrict calls when 2000 calls/second is reached (across the set of NNIs).

Requirement 14: CPs should limit the number of concurrent calls and calls/second capacity allowed for each NNI. Also, there shall be a higher control limiting the total concurrent calls and calls/second which can be supported.

7.2 Destination overload control

Destination overload controls shall be applied for high volume or peaky traffic streams. Destination overload control is often referred to as Call Gapping, an umbrella term used for all forms of call restriction, including:

- Rate limiting: a 'leaky bucket' method which admits calls at a predefined rate and rejects excess calls when the bucket is full.

- Gapping: where calls are blocked for a specific period of time after a call has been admitted.

There are various algorithms and implementations of both mechanisms.

These rate controls are typically deployed close to the destination. However, the larger the event, the closer to the source the controls should be applied, so as not to tie up network capacity inadvertently. So;

- For a small event, it may be sufficient to apply controls in the destination network itself.

- For larger events, controls should be deployed in transit networks which feed into the destination network in question.

- In the case of very large events, controls should be applied as close to the source as possible, i.e. back in the originating networks.

It is recommended that peaky traffic streams are rate limited to a sensible level, especially in cases when there is a high ineffective call rate.

The destination overload controls are addressing a different problem to the NNI overload controls, however the two are linked. Traffic to high volume / peaky destinations originates both from within the UK and from global networks. So the ability to limit peaky traffic streams frees up NNI capacity much more readily than if all those calls reached the destination, and the majority of calls then received engaged tone. The scenario of high calling rates to doctors' surgeries at 8am, where the terminating capacity is often limited is an example where it may be beneficial for a terminating CP to request originating and/or transit CPs to apply destination overload control. Another example is a televote event, where the combined traffic to the set of voting numbers is controlled, often with bilateral agreement.

Response code 486 or 600 is mandated for destination overload controls since it gives a clearer end-user experience (busy tone), thus encouraging the end user not to try again immediately but at some time later.

Requirement 15: Call Gapping, rate controls, or centralised measures shall be put in place to limit the traffic to high volume or peaky destinations. Sufficient traffic shall be admitted to fully populate the terminating capacity (lines / agents, etc), yet bound the traffic to limit the ineffective rate. Calls rejected shall not hunt, and shall return response code 486 or 600.

Nodes receiving a call rejected with response code 486 or 600 shall not hunt, and shall reject the call, sending response code 486 or 600 further backwards.

Requirement 16: CPs shall discuss and agree bilaterally the extent to which peaky traffic streams should be gapped across NNIs, when the total size of the NNI between them is considered of significant impact.

7.3 Number of re-attempts and hops

It is important to keep the number of re-attempts to a sensible level when the first choice path is not available. As a general rule, if there are lots of available routes, only a sub-set should be tried by any individual call.

As an example, in the scenario where a call will traverse 10 nodes each of which has 5 hunting options, there will be 5^{10} distinct paths across the network, any number of which could be tried in a congestion scenario. Many of these paths would probably not be tried in reality as SIP timers etc. would kick in, but the damage would have been done with load amplification of this nature.

Requirement 17: There shall not be more than 6 attempts in total per call across all the CP's egress nodes.

NOTE: The routeing strategies implemented to enforce Requirement 17 may differ between CPs and, in practise, CPs may even implement different strategies for different segments of their network.
A simple implementation would be to restrict calls to a single node for NNI egress and enforce a maximum of 6 NNI egress choices from any node. An alternative that delivers egress node diversity would be to increase the number of egress nodes that can be selected but enforce a smaller limit on the number of NNI egress choices available at each node. In this case, the product of egress nodes and NNI egress choices must be 6 or less (e.g. 3 egress nodes each with a maximum of 2 possible NNI connections). A more sophisticated solution would allow differing limits on the number of NNI egress choices at each node however this would likely require the implementation of a central routeing policy server (except in very small networks).

7.4 Response codes

SIP response Codes are defined in RFC 3261 [1] and give an indication that preceding nodes can use to select a next action. SIP was specified for use across the public Internet and, in common with many other Internet signalling systems, is designed to permissively allow re-attempts by the same path or by alternate paths if available. During congestion conditions in CP networks, excessive re-attempts can lead to load amplification and create a positive feedback loop which has the effect of worsening the overload condition. Therefore calls rejected due to network wide overload should be backward released without re-routeing since network wide overload is significantly more serious than route congestion. This eliminates the risks of multiplying the overload many times and spreading it network-wide within a CP or worse across multiple CPs. Collaborative real-time Network Management action may enable specific temporary expansive routeing once the nature of the overload is understood.

The following SIP response codes should be used to signal different categories of congestion in order to avoid excessive re-attempts.

- 486 Busy Here or 600 Busy Everywhere
    - The called party is busy or there is insufficient terminating capacity to complete the call.
    - Do not re-attempt the call.
    - Return 486 or 600 on each hop back to the originating network so that it plays busy tone to the caller.

- 500 Server Internal Error
    - A significant portion of the network is congested.
    - Do not re-attempt the call, because to do so will likely amplify the congestion.
    - Return 500 on each hop back to the originating network so that it plays Congestion tone to the caller.

- 503 Service Unavailable
    - A route is congested.
    - Do not re-attempt the call on the same route.
    - Re-attempting the call on alternate routes may be permissible subject to compliance with Requirement 17.

- A SIP 503 response code should be returned on each hop back to the originating network, except where a node determines that the re-attempt limits specified in Requirement 17 has been exceeded (in which case a SIP 500 response code must be returned).

Depending on the trigger conditions, the extent of a congestion event can range from a single node (e.g. due to a software or hardware issue) to network wide (e.g. due to a large volume of calls triggered by a disaster event). In the former case, a re-attempt may result in completion of the call while in the latter case it is highly unlikely hence the SIP response code chosen needs to reflect the specific scenario.

The following broad classifications identify the key congestion events that can be identified by SIP nodes:

- Ingress congestion
  'Ingress congestion' is where a node detects that it is close to reaching pre-defined limits and needs to protect itself, and the network behind it, by reducing the processing impact of received load. Possible symptoms for detection include high CPU, resource constraints, configured incoming rate exceeded, etc.

  Following detection, a 503 Service Unavailable response should be sent to allow preceding nodes an option to re-attempt on alternative routes.

- Egress congestion – route
  'Egress congestion' at a route is where a node is unable to place a call over a selected route. This might be because the engineered capacity has already been consumed, the destination node has sent a SIP response indicating failure (but permitting a re-attempt over an alternative route etc). Possible symptoms for detection include no response to INVITE request, all channels busy, receipt of 503 Service Unavailable response.

  Following detection, the node can select a different routeing choice except where this will lead to breach of Requirement 17 and detection of "Egress congestion – network wide". Where no further routeing choices are available, a 503 Service Unavailable response will be sent to allow preceding nodes an option to re-attempt on an alternative route.

- Egress congestion – network wide
  'Egress congestion – network wide' is where multiple routes across multiple nodes are congested or multiple nodes are overloaded resulting in a significant restriction of possible paths through the CP's network.

  This will be detected by routeing policy following failure to set up a call over an individual routeing choice and where the call has previously attempted routeing over other choices in compliance with Requirement 17.

  Following detection, a 500 Server Internal Error response will be sent to prevent preceding nodes from re-attempting the call.

- Egress congestion - terminating node

‘Egress congestion - terminating node’ is where a network is unable to deliver calls due to overload at a terminating CP node, and there is no other way to reach the destination.

Following detection, a 500 Server Internal Error response will be sent to prevent preceding nodes from re-attempting the call.

- Destination overload control
  ‘Destination overload control’ is discussed in section 7.2

The scope of this document is restricted to UNI and NNI but, as a network option, the classifications above could be usefully applied to SIP connections within a CP network.

Explicit rejection with a response code is better than discard (but this only works up to a certain overload level). Discard should only be used under extreme overload. The pragmatic measures outlined here will help avoid high amplification, with the aim of keeping overload within the bounds where explicit rejection works.

Requirement 18: The following SIP response codes shall be used, as stipulated in Table 2.

Scenario	Response code for rejected calls
‘Ingress congestion’	503 Service Unavailable (note 2)
‘Egress congestion –route’	503 Service Unavailable (note 2)
‘Egress congestion – network wide’	500 Server Internal Error
‘Egress congestion – terminating node’	500 Server Internal Error
‘Destination overload control’	486 Busy Here/600 Busy Everywhere
NOTE 1: Response code 503 with Retry-After X means do not send me any new traffic for X seconds. This applies both to an originator such as a CPE device, and also to a previous network node (which is likely to have many calls in time X). The specific call which was rejected should be forwarded to an alternate server, if one exists. Retry-After is not recommended as a robust overload control mechanism for SIP INVITEs as it is ignored by many CPE devices. **NOTE 2:** When SIP networks are connected to TDM networks, CPs are referred to ND1037 [i3] for the appropriate ISUP cause codes to which each SIP response code should map.	

Table 2: SIP response codes for rejected calls

7.5 Bilateral reviews

It is important that CPs work together to mitigate network overloads. This is especially so for CPs which have large interconnect routes between them.

Requirement 19: When the total size of the NNI between CPs is considered of significant size or importance to either of them, or the UK network or UK customers, they shall conduct bilateral periodic reviews of the capacity, the rate controls, the re-attempt policy and the response codes, in line with Table 2 above, in order to check for mutual overload protection.

8 Originating network overload mitigation

8.1 Call Admission Control (CAC)

It is important to use CAC to limit the calling rate and number of simultaneous sessions a large access trunk group can admit into a CPs network to prevent overload.

Requirement 20: CAC (number of sessions, calls/second rate) shall be agreed with the customer and implemented, in the case of customer access trunk groups such as SIP Trunking / Unified Comms.

8.2 Device auto-re-attempt

End user devices and CPE should not automatically re-attempt the call many times. This is not good practice, since it adds to the level of traffic amplification during an overload.
It is much better to play an appropriate tone or announcement to the user, and allow them to re-attempt at a time of their choosing.

Requirement 21: End user devices / phones shall not automatically re-attempt failed calls.

NOTE: In the case of customer owned and managed CPE, Requirement 21 is beyond the scope of CPs, however CPs must ensure that Requirement 21 is enforced for any CPE under their control.

8.3 IP layer rate limiting and call rate control

It is good practice to limit the rate of call attempts per source IP address (in access SBC), as part of traffic shaping. Businesses/Call Centre customers will need higher thresholds than Residential customers.

Requirement 22: Call rate control shall be configured to bound the rate of call attempts per source IP Address.

Furthermore, extreme IP signalling rates, way in excess of normal traffic levels, shall be limited as part of cyber defence. This provides a safety net beyond the SIP layer controls specified in this document. See section 8.5 DoS / DDoS attack.

8.4 General overload of the UK network

A general overload is defined as a general increase in traffic spread across a large range of destination numbers, such as would be caused by a national disaster.

Rate controls shall be deployed in each originating network, to control the total traffic each originating network can send to the wider network. This protects the wider network in the case of a general overload from one or more originating networks.

In the case of high traffic to a destination number or number range, Call Gapping, rate controls or centralised measures are recommended and, if the event is to be sufficiently large, should be deployed in originating networks (see section 7.2 Destination overload control)

Requirement 23: To prevent general overload of the UK network by originating networks, rate controls shall be applied to help bound the total traffic delivered to the network.

8.5 DoS / DDoS attack

Denial of Service (DoS) and Distributed Denial of Service (DDoS) are special cases of overload because they are deliberate and designed to cause overload by known (or expected) weaknesses in networks rather than by accidental or freak events. These must be stopped by deploying suitable SBCs or firewalls, with appropriate capacity, at the outermost edge of a CP's originating network.

There are many different types of DoS/DDOS attack. These may utilise strategically manufactured or malformed SIP or IP packets, together with large numbers of hacked devices, to amplify the attack. CPs should ensure their SBCs and/or firewalls are designed with this in mind.

As attackers develop new attack vectors, it is essential that CPs update their SBCs or firewalls with security patches from their vendors.

Requirement 24: CPs shall deploy SBCs and/or firewalls on exposed IP edges of their originating network, such as connections from the public internet.

Requirement 25: CPs shall ensure their SBCs and firewalls have the necessary capacity to withstand an anticipated DDoS attack and assess this capacity at regular intervals.

Requirement 26: CPs shall ensure their SBCs and firewalls are kept up to date with security patches.

Requirement 27: IP layer rate limiting shall be configured to bound extreme IP signalling traffic when the signalling traffic rate is several times higher than expected. This shall be done in such a way that the normal SIP layer controls specified in this document will be invoked first in the majority of overload scenarios.

8.6 Scam and nuisance calls

Scam calls, and other types of nuisance calls, can increase the load on a network but are unlikely, by themselves, to cause a network overload.

The mitigations described in this document do not deal specifically with scam traffic unless the level of traffic is sufficiently high and leads to overload conditions. Other types of controls are required to detect and deal with scam and nuisance call traffic. Such controls for scam and other types of nuisance call are outside the scope of this document and are, instead, addressed by other NICC documents.

Annex A (informative):

Call flow example scenarios

Annex A presents a number of examples of the scenarios described in ND1657 Table 2 Requirement 18 (Section 7.4). This is not a complete list; these examples are not exhaustive, but rather are possible scenarios for illustrative purposes.

- Section A.1 covers the 'Ingress congestion' scenario. See example 1.
- Section A.1 also covers 'Egress congestion – route' and 'Egress congestion – network wide'. See examples 2, 3 and 4.
- Section A.2 covers the 'Egress congestion – terminating node' scenario. See example 5.
- Section A.3 covers 'Destination overload control' scenario. See examples 6 and 7.

Each numbered node in the example diagrams which follow is an SBC or equivalent network node responsible for routeing the next hop of the SIP call. Nodes A and B are IP CPE devices or IP PBXs. The key for the diagrams is given in Table 1.

————	SIP trunk (NNI or UNI)
- - - - - - - - - - - -	SIP trunk (intra-network)
————	SIP trunk (NNI or UNI) in scope for the example call
- - - - - - - - - - - -	SIP trunk (intra-network) in scope for the example call
————	SIP trunk (NNI or UNI) which is congested for the example call
— — — —	SIP trunk (intra-network) which is congested for the example call
← 486/500/503/600	The return path of a SIP response code
1	A node
1	A node which is congested and rejects the example call
1	A node which rejects the example call due to destination overload control (rate control)

Table 1: Key for diagrams

A.1 'Ingress congestion', 'egress congestion – route' and 'egress congestion – network wide' scenarios

Example 1a (left), and 1b (right)

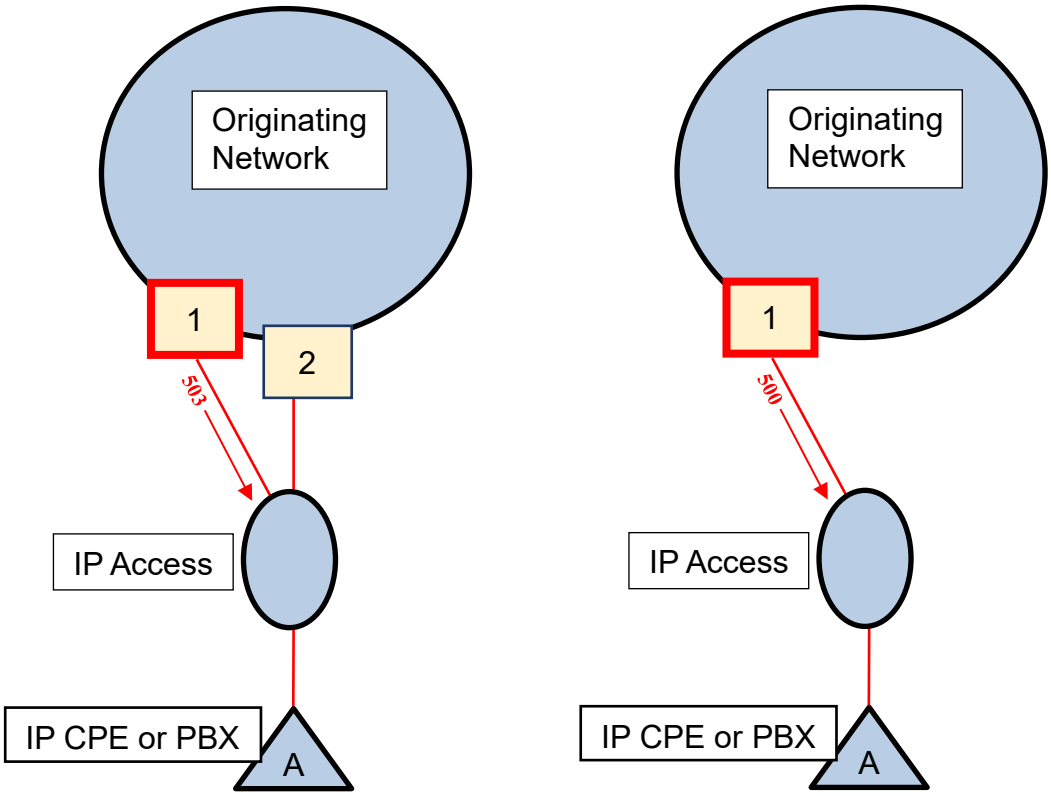

Figure 1: network diagrams for examples 1a and 1b

Example 1a (left): in this scenario, a call from CPE A has arrived at Node 1 in the originating network.

- In this scenario CPE A connects to the originating network via Node 1 or Node 2.
- Node 1 is unable to onward route the call due to 'ingress congestion', so rejects the call returning a 503. CPE A could try to route via Node 2. In this case the CP would need to satisfy themselves that the CPE will not re-attempt the call on the same route.

NOTE: Sending of SIP 503 response code is the default action for ingress congestion. Where a customer has a connection to a single node as in example 1b (right) with Node 1 only, CPs may provide an option that a SIP 500 response code is sent, to minimise load amplification.

Example 2

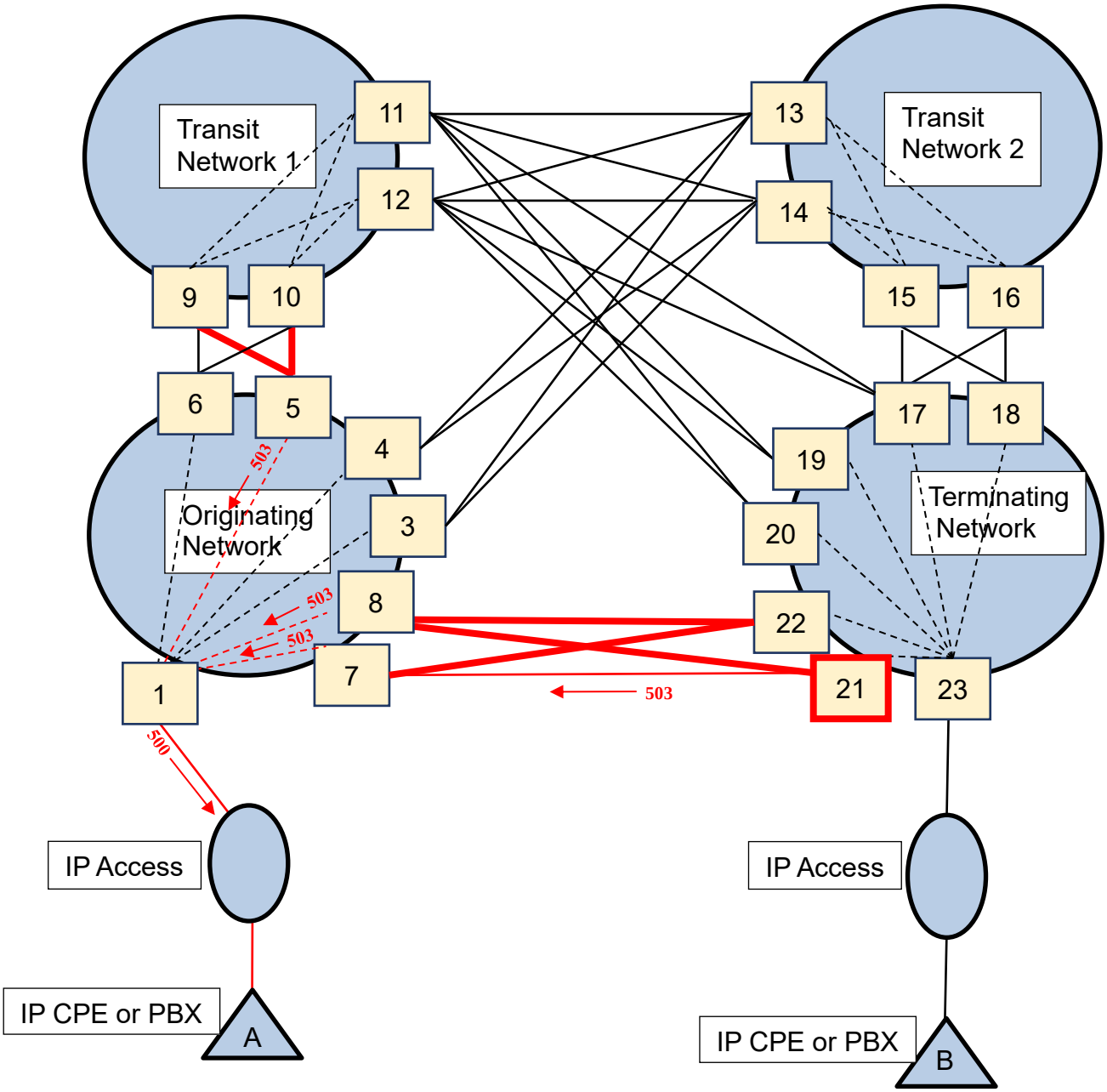

Figure 2: network diagram for example 2

Example 2: in this scenario, a call from CPE A has arrived at Node 1 in the originating network and is destined for CPE B which is connected to the terminating network.

- In order to comply with Requirement 17, the originating network can make a maximum of 6 attempts to route the call (see the note after Requirement 17 for options on how this can be done). In this case, the sub-set of routes to try are the direct SIP trunks from Node 7 to Nodes 21 and 22, and from Node 8 to Nodes 21 and 22, and also the SIP trunks from Node 5 to Nodes 10 and 9 in transit network 1.

- Node 1 routes the call across the internal SIP trunk to Node 7.
    - Node 7 routes the call across the external SIP trunk to Node 21. Node 21 is experiencing 'ingress congestion', so rejects the call returning a 503 to Node 7. Thus, Node 7 has encountered 'egress congestion – route', for the route to Node 21.
    - Node 7 selects the next choice in its routeing list and attempts to place the call via Node 22, but the SIP trunk is congested, so the call does not reach Node 22, hence, Node 7 has encountered 'egress congestion – route'.
    - Hence, Node 7 has exhausted its routing options, so rejects the call returning a 503 to Node 1.

- Node 1 receives the 503 from Node 7, and selects the next choice in its routeing list, routeing the call across the internal SIP trunk to Node 8.
    - Node 8 attempts to place the call via Node 21, but the SIP trunk is congested. Thus, Node 8 has encountered 'egress congestion – route', for the SIP trunk to Node 21.
    - Node 8 selects the next choice in its routeing list and attempts to place the call via Node 22, but the SIP trunk is congested. Thus, Node 8 has encountered 'egress congestion – route', for the SIP trunk to Node 22.
    - Hence, Node 8 has exhausted its routing options, so rejects the call returning a 503 to Node 1.

- Node 1 receives the 503 from Node 8, and selects the next choice in its routeing list, routeing the call across the internal SIP trunk to Node 5.
    - Node 5 attempts to place the call via Node 10 in transit network 1, but this external SIP trunk is congested, so the call does not reach Node 10, hence, Node 5 has encountered 'egress congestion – route' for the SIP trunk to Node 10.
    - Node 5 selects the next choice in its routeing list and attempts to place the call via Node 9.
    - If the call is successful in reaching Node 9, transit network 1 will attempt to route the call to the terminating network (see Example 3).
    - But in this case the external SIP trunk to Node 9 is congested, so the call does not reach Node 9, hence, Node 5 has encountered 'egress congestion – route'.
    - Hence, Node 5 has exhausted its routing options, so rejects the call returning a 503 to Node 1.

- Node 1 receives the 503 from Node 5 and that exhausts the maximum allowable re-attempts defined in Requirement 17. Hence, Node 1 has encountered 'egress congestion – network wide' and it rejects the call, sending back a 500 to CPE A.

Example 3

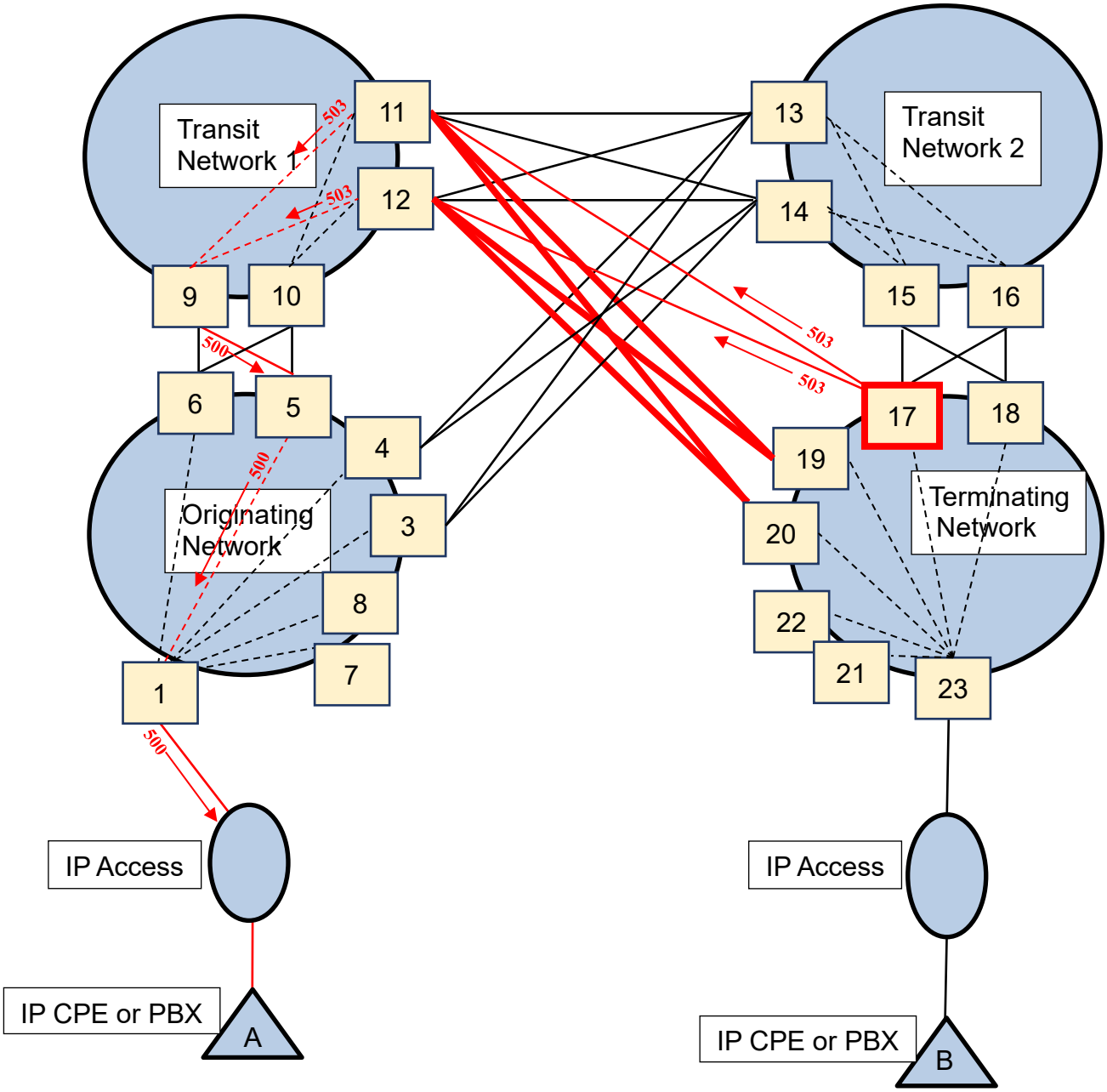

Figure 3: network diagram for example 3

Example 3: in this scenario, a call from CPE A has been routed via Node 1 and Node 5 and has arrived at Node 9 in transit network 1, destined for CPE B which is connected to the terminating network.

- In order to comply with Requirement 17, transit network 1 can make a maximum of 6 attempts to route the call (see the note after Requirement 17 for options on how this can be done). In this case, the sub-set of routes to try are from Node 11 to Nodes 17, 19 and 20, and from Node 12 to Nodes 17, 19 and 20.

- Node 9 routes the call across the internal SIP trunk to Node 11.
    - Node 11 routes the call across the external SIP trunk to Node 17. Node 17 is experiencing 'ingress congestion', so rejects the call returning a 503 to Node 11. Thus, Node 11 has encountered 'egress congestion – route', for the route to Node 17.
    - Node 11 selects the next choice in its routeing list and attempts to place the call via Node 19, but the SIP trunk is congested, so the call does not reach Node 19, hence, Node 11 has encountered 'egress congestion – route'.
    - Node 11 selects the next choice in its routeing list and attempts to place the call via Node 20, but the SIP trunk is congested, so the call does not reach Node 20, hence, Node 11 has encountered 'egress congestion – route'.
    - Hence, Node 11 has exhausted its routing options, so rejects the call returning a 503 to Node 9.

- Node 9 receives the 503 from Node 11, and selects the next choice in its routeing list, routeing the call across the internal SIP trunk to Node 12.
    - Node 12 routes the call across the external SIP trunk to Node 17. Node 17 is experiencing 'ingress congestion', so rejects the call returning a 503 to Node 12. Thus, Node 12 has encountered 'egress congestion – route', for the route to Node 17.
    - Node 12 selects the next choice in its routeing list and attempts to place the call via Node 19, but the SIP trunk is congested, so the call does not reach Node 19, hence, Node 12 has encountered 'egress congestion – route'.
    - Node 12 selects the next choice in its routeing list and attempts to place the call via Node 20, but the SIP trunk is congested, so the call does not reach Node 20, hence, Node 12 has encountered 'egress congestion – route'.
    - Hence, Node 12 has exhausted its routing options, so rejects the call returning a 503 to Node 9.

- Node 9 receives the 503 from Node 12 and that exhausts the maximum allowable re-attempts defined in Requirement 17. Hence, Node 9 has encountered 'egress congestion – network wide' and it rejects the call, sending back a 500 to Node 5 in the originating network, which upon receiving a 500 does not hunt but rejects the call, sending a 500 back to Node 1. Node 1 does not hunt but rejects the call, sending a 500 back to CPE A.
- NOTE: in example 2, if the call at Node 5 had been successful in reaching Node 9, and then Node 9 had proceeded as per example 3, then there would have been a total of 10 attempts to reach the terminating network for this call (4 from the originating network in example 2 plus 6 from transit network 1 in example 3).

Example 4

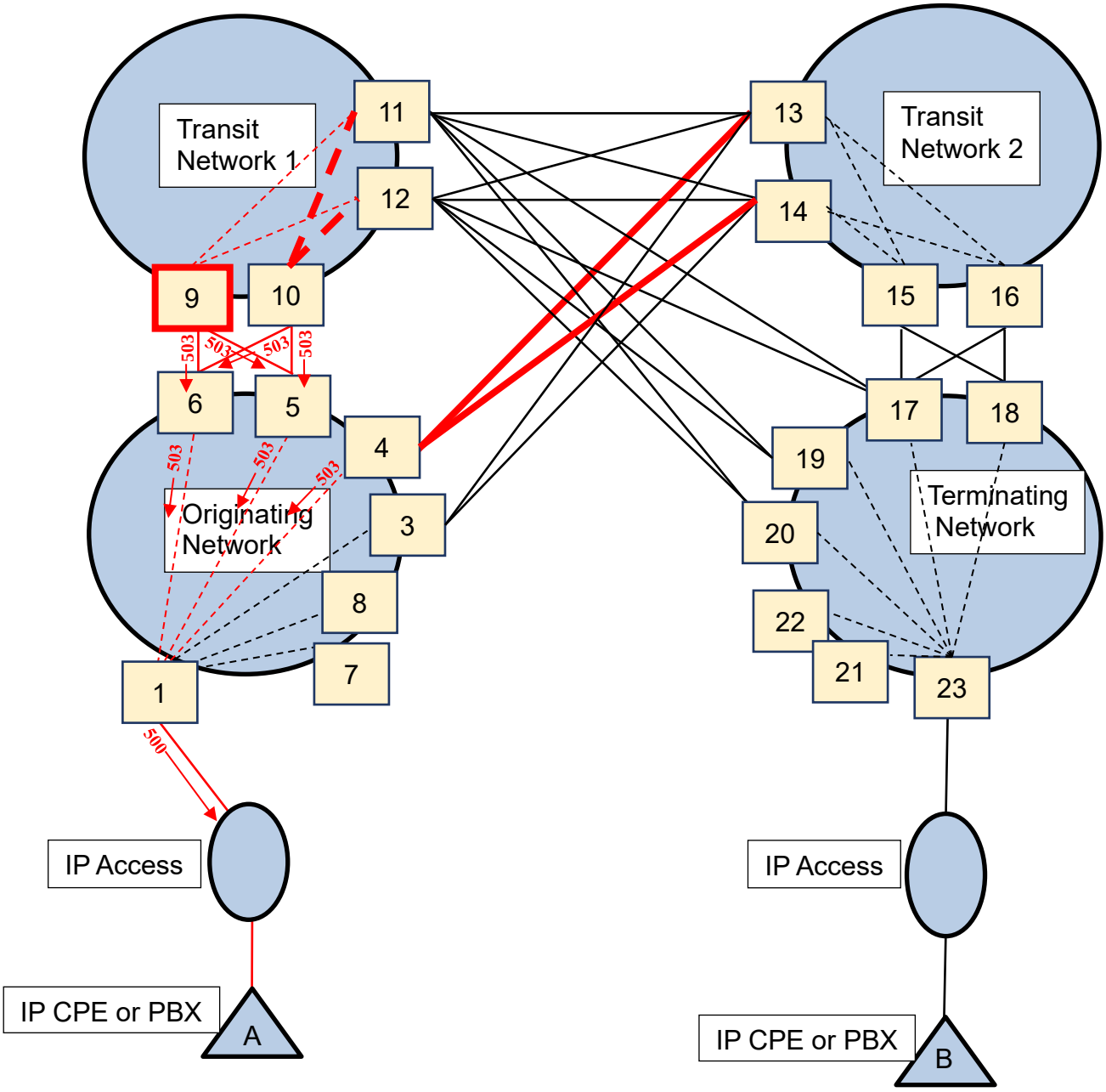

Figure 4: network diagram for example 4

Example 4: in this scenario, a call from CPE A has arrived at Node 1 in the originating network and is destined for CPE B which is connected to the terminating network.

- In this case there is no direct connection from the originating network to the terminating network. The originating network can attempt to route to the terminating network via transit network 1 and transit network 2.

- In order to comply with Requirement 17, the originating network can make a maximum of 6 attempts to route the call (see the note after Requirement 17 for options on how this can be done). In this case the sub-set of routes to try are the SIP trunks from Node 5 to Nodes 9 and 10 in transit network 1, and from Node 6 to Nodes 9 and 10 in transit network 1, and also the SIP trunks from Node 4 to Nodes 13 and 14 in transit network 2.

- Node 1 routes the call across the internal SIP trunk to Node 5.
    - Node 5 routes the call across the external SIP trunk to Node 9. Node 9 is experiencing 'ingress congestion', so rejects the call returning a 503 to Node 5. Thus, Node 5 has encountered 'egress congestion – route', for the route to Node 9.
    - Node 5 selects the next choice in its routeing list and attempts to place the call via Node 10. The call reaches Node 10, which attempts to route the call across transit network 1 to Nodes 11 and 12, but the internal SIP trunks to Nodes 11 and 12 are both congested, so the call does not reach Node 11 or 12. Hence, Node 10 has exhausted its routing options, so rejects the call returning a 503 to Node 5.
    - Hence, Node 5 has exhausted its routing options, so rejects the call returning a 503 to Node 1.

- Node 1 receives the 503 from Node 5, and selects the next choice in its routeing list, routeing the call across the internal SIP trunk to Node 6.
    - Node 6 routes the call across the external SIP trunk to Node 9. Node 9 is experiencing 'ingress congestion', so rejects the call returning a 503 to Node 6. Thus, Node 6 has encountered 'egress congestion – route', for the route to Node 9.
    - Node 6 selects the next choice in its routeing list and attempts to place the call via Node 10. The call reaches Node 10, which attempts to route the call across transit network 1 to Nodes 11 and 12, but the internal SIP trunks to Nodes 11 and 12 are both congested, so the call does not reach Node 11 or 12. Hence, Node 10 has exhausted its routing options, so rejects the call returning a 503 to Node 6.
    - Hence, Node 6 has exhausted its routing options, so rejects the call returning a 503 to Node 1.

- Node 1 receives the 503 from Node 6, and selects the next choice in its routeing list, routeing the call across the internal SIP trunk to Node 4.
    - Node 4 attempts to place the call via Node 13 in transit network 2, but the SIP trunk is congested, so the call does not reach Node 13, hence, Node 4 has encountered 'egress congestion – route'.
    - Node 4 selects the next choice in its routeing list and attempts to place the call via Node 14.
    - If the call is successful in reaching Node 14, transit network 2 will attempt to route the call to the terminating network.
    - But in this case the external SIP trunk to Node 14 is congested, so the call does not reach Node 14, hence, Node 4 has encountered 'egress congestion – route'.
    - Hence, Node 4 has exhausted its routing options, so rejects the call returning a 503 to Node 1.

- Node 1 receives the 503 from Node 4 and that exhausts the maximum allowable re-attempts defined in Requirement 17. Hence, Node 1 has encountered ‘egress congestion – network wide’ and it rejects the call, sending back a 500 to CPE A.
- NOTE: in this case, it is transit network 1 which is overloaded. Hence, since transit network 1 returns a 503, the originating network can try alternative routes (such as via transit network 2).
- NOTE: if transit network 1 had returned a 500 to the originating network then the originating network would not hunt, and not try to route the call via transit network 2. For example, 500 is returned to the originating network in the following cases:
    - Example 3, where transit network 1 has encountered ‘egress congestion – network wide’ towards the terminating network
    - Example 5, where the terminating network has encountered ‘egress congestion – terminating node’, and transit network 1 is relaying the 500 further backwards.
- NOTE: if transit network 1 had returned a 486 or 600 to the originating network (as per example 6 and example 7), then the originating network would not hunt, and not try to route the call via transit network 2.

This page is intentionally left blank.

A.2 ‘Egress congestion – terminating node’ scenario

Example 5

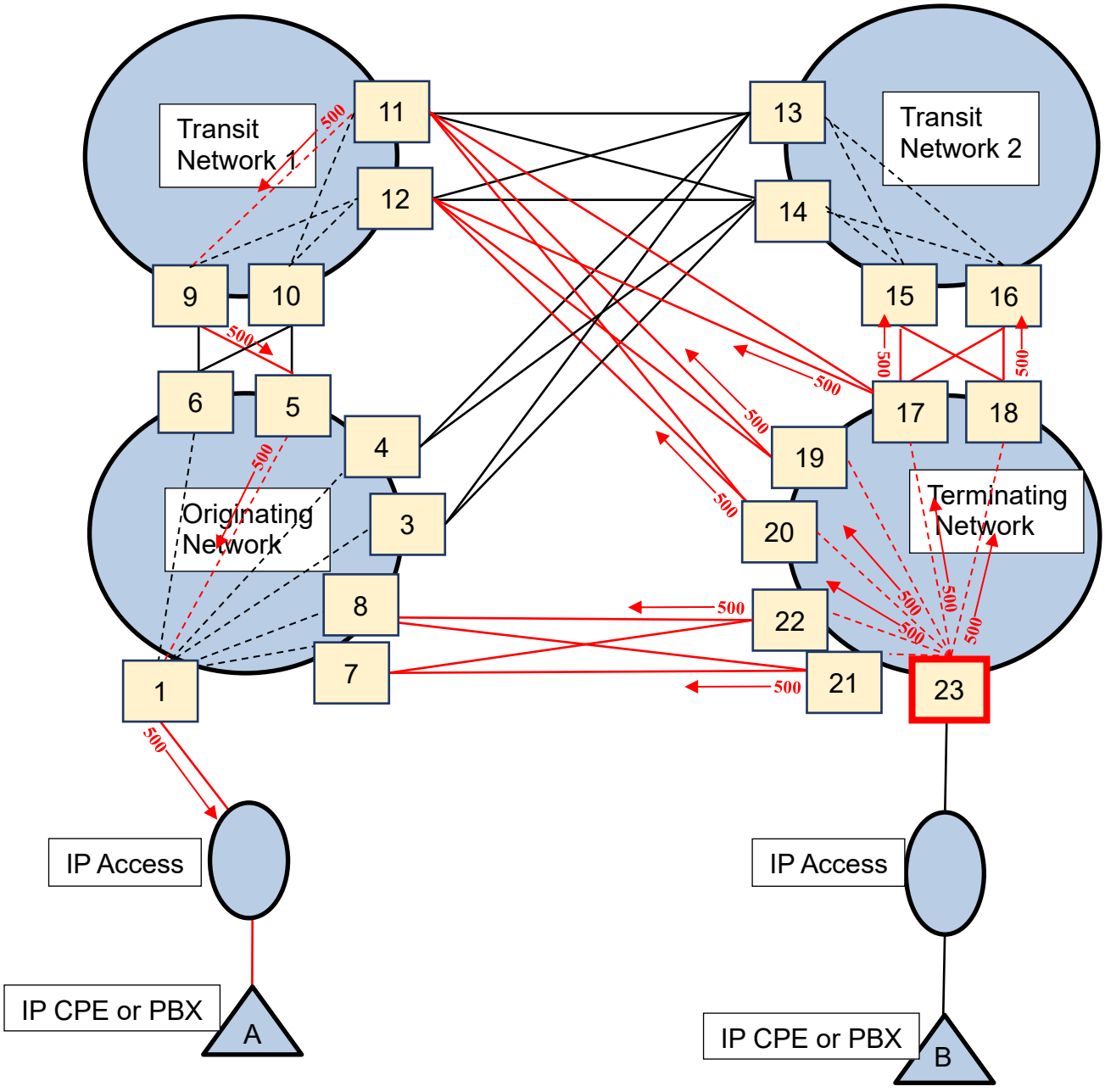

Figure 5: network diagram for example 5

Example 5: in this scenario, lots of calls are arriving the the terminating network, headed for destinations which are only reachable via Node 23.

- Due to the high volume of calls, the terminating network is unable to deliver all calls to their destinations due to overload at Node 23, and there is no other way to reach these destinations.
- Hence, Node 23 has encountered ‘egress congestion – terminating node’.
- Node 23 will reject excess traffic by sending a SIP 500 response code.
- Any node receiving a 500 shall not attempt any alternate routes and shall send a 500 back to any preceding node. This would apply to Nodes 17, 18, 19, 20, 21 and 22 in the terminating network; these Nodes will also send a 500 back to Nodes 15 and 16 in transit network 2, Nodes 11 and 12 in transit network 1, and Nodes 7 and 8 in the originating network.
- In the example shown, a call from CPE A has routed via Node 1, Node 5 and Node 9 to reach Node 11. When Node 11 receives a 500 back from the terminating network, it does not hunt but rejects the call, sending a 500 back to Node 9. Upon receipt of the 500, Node 9 does not hunt but rejects the call, sending a 500 back to Node 5. Upon receipt of the 500, Node 5 does not hunt but rejects the call, sending a 500 back to Node 1. Upon receipt of the 500, Node 1 does not hunt but rejects the call, sending a 500 back to CPE A.

A.3 ‘Destination overload control’ scenario

Example 6

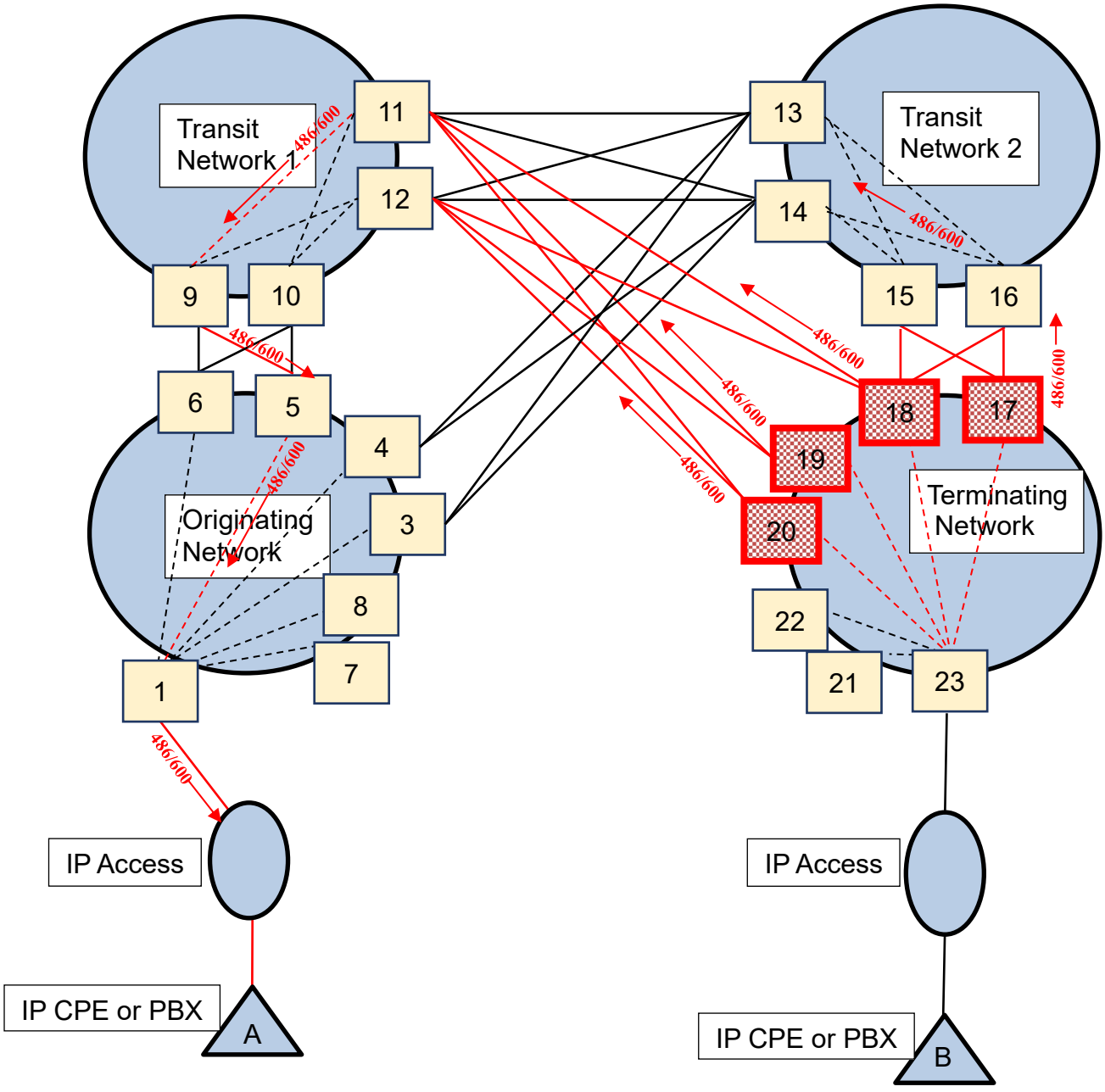

Figure 6: network diagram for example 6

Example 6: in this scenario the terminating network hosts a high volume traffic destination number range on CPE B accessed via Node 23. In this case there are no routes directly from the originating network to the terminating network Nodes 21 and 22; all traffic destined for the high volume number range is routed via transit networks 1 and 2, and ingress into the terminating network is via Nodes 17, 18, 19 and 20.

- Nodes 17, 18, 19 and 20 in the terminating network will control the rate of traffic sent to this destination number range, sufficient to fill the terminating capacity, in accordance with Requirement 15. For example, a rate of X calls/second is desired, and is achieved by applying destination number rate control to this specific number range, at a rate of X/4 calls/second at each of Nodes 17, 18, 19 and 20.
- Excess traffic rejected by this rate control on Nodes 17, 18, 19 and 20 does not hunt, and returns 486 or 600 to the previous nodes in the transit networks (e.g. Nodes 11, 12, 15 and 16), who, upon receipt of a rejected call with 486 or 600, will not hunt, but will reject the call sending back 486 or 600.
- In the example shown, a call from CPE A has routed via Node 1, Node 5 and Node 9 to reach Node 11. When Node 11 receives a 486/600 back from the terminating network, it does not hunt but rejects the call, sending a 486/600 back to Node 9. Upon receipt of the 486/600, Node 9 does not hunt but rejects the call, sending a 486/600 back to Node 5. Upon receipt of the 486/600, Node 5 does not hunt but rejects the call, sending a 486/600 back to Node 1. Upon receipt of the 486/600, Node 1 does not hunt but rejects the call, sending a 486/600 back to CPE A, which plays busy tone back to the caller.

Example 7

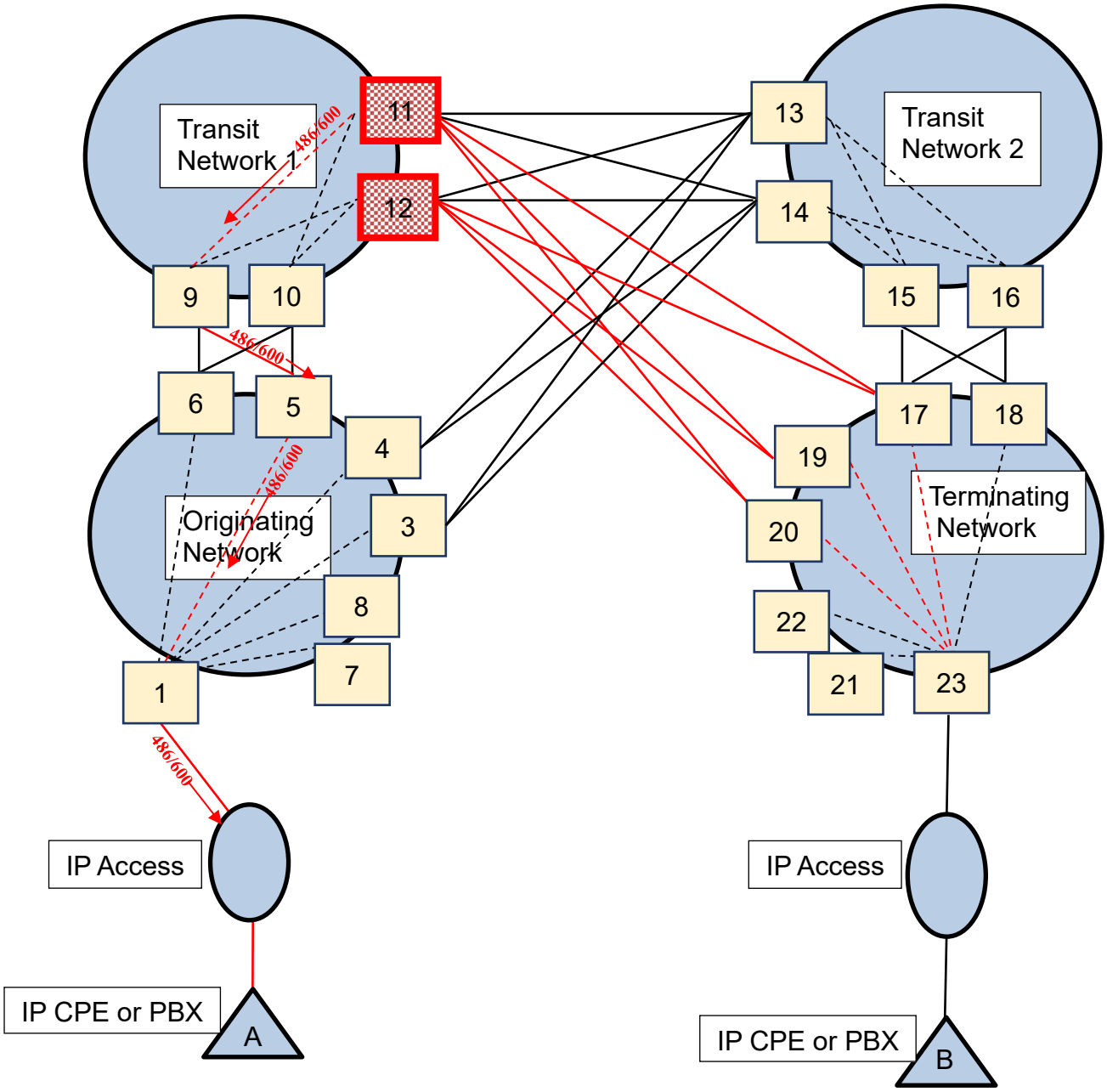

Figure 7: network diagram for example 7

Example 7: in this scenario, following bilateral discussion, transit network 1 has agreed to deliver no more than Y calls/second to a peaky destination number at CPE B hosted in the terminating network, in accordance with Requirement 16. Traffic is shared across Nodes 11 and 12.

- Hence, destination number rate control is applied to this specific number, at a rate of Y/2 calls/second, on each of Node 11 and Node 12.
- Excess traffic does not hunt at Nodes 11 or 12 but is rejected, returning a 486 or 600 further back (such as to Nodes 9 and 10, which will reject the call without hunting, returning a 486 or 600 further back).
- In the example shown, a call from CPE A has routed via Node 1, Node 5 and Node 9 to reach Node 11. Node 11 rejects the call due to the destination rate control, sending a 486/600 back to Node 9. Upon receipt of the 486/600, Node 9 does not hunt but rejects the call, sending a 486/600 back to Node 5. Upon receipt of the 486/600, Node 5 does not hunt but rejects the call, sending a 486/600 back to Node 1. Upon receipt of the 486/600, Node 1 does not hunt but rejects the call, sending a 486/600 back to CPE A, which plays busy tone back to the caller.

History

Document history		
Version	Date	Milestone
1.1.1	30th March 2023	Initial publication
2.1.1	28th July 2023	Second publication including the new annex