



国际电信联盟

ITU-T

国际电信联盟
电信标准化部门

H.248.11

(11/2002)

H系列：视听和多媒体系统

视听业务的基础设施 — 通信规程

网关控制协议：媒体网关过载控制包

ITU-T H.248.11建议书

ITU-T H系列建议书
视听和多媒体系统

可视电话系统的特性	H.100-H.199
视听业务的基础设施	
概述	H.200-H.219
传输多路复用和同步	H.220-H.229
系统概况	H.230-H.239
通信规程	H.240-H.259
活动图像编码	H.260-H.279
相关系统概况	H.280-H.299
视听业务的系统和终端设备	H.300-H.399
多媒体的补充业务	H.450-H.499
移动性和协作程序	
移动性和协作、定义、协议和程序概述	H.500-H.509
H系列多媒体系统和业务的移动性	H.510-H.519
移动多媒体协作应用和业务	H.520-H.529
移动多媒体应用和业务的安全性	H.530-H.539
移动多媒体协作应用和业务的安全性	H.540-H.549
移动性互通程序	H.550-H.559
移动多媒体协作互通程序	H.560-H.569

欲了解更详细信息，请查阅ITU-T建议书目录。

网关控制协议：媒体网关过载控制包

摘 要

本建议书描述了用于媒体网关（MG）过载控制的包以和H.248.1网关控制协议结合使用。该包用于保护MG防止处理过载，这些过载防止了H.248.1的事务处理及时执行。

总而言之，在本建议书中，过载保护如下所述：

- 1) 无论何时只要MG接收到ADD命令，MG（或虚拟MG）检测到其处于过载状态就会向其所属的媒体网关控制器（MGC）通知这种状态。
- 2) MGC能够自适应地控制使用MG（或虚拟MG）建立呼叫的速率以最大化MG的有效吞吐量而同时限制其响应返回的时间。MGC通过控制与呼叫建立相关的事务处理速率来做到这一点，即建立新呼叫或发送新的呼叫接续给过载的MG。这样可以使MGC从过载的MG（或虚拟MG）接收到的过载通知速率会聚到一个合适的低水平。

MGC为每个其所控制的过载的MG（或虚拟MG）初始化独立的过载控制实例。这些独立的实例应独立地运行（也就是说它们之间并不明确地交换信息）。它们的过载控制参数应该可独立配置，例如，通过专有的管理接口来配置，或使用SNMP来调用配置功能。

控制所能够处理的最常见的过载事例是一个或多个MGC共同使单个MG过载，而该MG具有多个虚拟MG（虚拟MG ‘i’ 仅仅与MGC ‘i’ 交互作用）。控制不必知道是多少个MGC引起MG过载，也不必知道MG的容量是多少。资料性参考文件[1]提供了全面的解释如何实现控制，资料性参考文件[2]为设计过载控制提供了更进一步的材料。

过载控制是通过叙述其应该如何去做来规定的，而不是描述其应该如何实现来获得这种效果。这里有两个重要的结论。

第一个结论，部分包（见8.5节）定义了一系列的过载事例，这些包的任何完全顺从的实现必须自动地满足每个事例的全部要求（即不用考虑运营者的干预需求，从一个过载事例调整参数值转换到另一个事例）。

第二个结论，该包并不需要了解所有的可配置参数，因为这些参数的实现依靠控制的特定实现。不过存在这种要求，即实现应提供某种方法，运营者可以采用这种方法来改变影响控制性能的所有参数。与该包相关的管理要求请见第9节。一般期望这些要求将由专用的管理接口或使用SNMP来实现。

来 源

ITU-T第16研究组（2001-2004）起草并按照WTSA第1号决议规定的程序于2002年11月29日批准了ITU-T H.248.11建议书。

前 言

国际电信联盟（ITU）是从事电信领域工作的联合国专门机构。ITU-T（国际电信联盟电信标准化部门）是国际电信联盟的常设机构，负责研究技术、操作和资费问题，并且为在世界范围内实现电信标准化，发表有关上述研究项目的建议书。

每四年一届的世界电信标准化全会（WTSA）确定ITU-T各研究组的研究课题，再由各研究组制定有关这些课题的建议书。

WTSA第1号决议规定了批准建议书须遵循的程序。

属ITU-T研究范围的某些信息技术领域的必要标准，是与国际标准化组织（ISO）和国际电工委员会（IEC）合作制定的。

注

本建议书为简要而使用的“主管部门”一词，既指电信主管部门，又指经认可的运营机构。

知识产权

国际电联提请注意：本建议书的应用或实施可能涉及使用已申报的知识产权。国际电联对无论是其成员还是建议书制定程序之外的其他机构提出的有关已申报的知识产权的证据、有效性或适用性不表示意见。

至本建议书批准之日止，国际电联尚未收到实施本建议书可能需要的受专利保护的知识产权的通知。但需要提醒实施者注意的是，这可能不是最新信息，因此大力提倡他们查询电信标准化局（TSB）的专利数据库。

© 国际电联 2003

版权所有。未经国际电联事先书面许可，不得以任何手段复制本出版物的任何部分。

目 录

	页
1 范围	1
2 参考文献	1
2.1 规范性参考文献	1
2.2 资料性参考文献	1
3 定义	1
4 缩写	3
5 过载控制包	3
5.1 属性	3
5.2 事件	4
5.2.1 MG_Overload	4
6 信号	4
7 统计	4
8 程序	4
8.1 在过载的 MG（或虚拟 MG）上的行为	4
8.2 在 MGC 上的行为	4
8.2.1 在 MGC 上的控制初始化	4
8.2.2 在 MGC 上的呼叫限制方法	5
8.2.3 在 MGC 上的允许呼叫速率的自适应	5
8.2.4 在 MGC 上控制的终止	6
8.2.5 在 MGC 上使用优先级	6
8.3 在稳态下限定 MG 的响应时间	9
8.4 在初始过载瞬间限定提供给 MG 的速率	9
8.5 过载事例的范围	10
9 对 H.248.11 的管理要求	10
9.1 漏桶限流器的大致性能分析	10
9.2 在 MGC 上的漏桶限流器的配置	14
9.3 与 MG 上的过载检测相关的专用参数的配置	15
9.4 与 MGC 上的控制激活相关的专用参数配置	15
9.5 与 MGC 上所允许速率的自适应相关的专用参数的配置	15
9.6 在 MGC 上控制终止的配置	15
9.7 MGC 统计	16

网关控制协议：媒体网关过载控制包

1 范围

本建议书描述了用于H.248.1网关协议相关的媒体网关过载控制包。随着根终结点实现该包，网关被期望能够向媒体网关控制器报告MG_Overload事件。

下列MGC/MG特征需要用于本包以有效地工作：

- 1) 应用由两个终端用户之间的单个点到点呼叫和涉及到几个附加呼叫方呼叫（例如，会议呼叫，由MG做为会议桥）构成。
- 2) 使用MG的点到点呼叫建立典型地涉及到不多于两个ADD命令，该命令由MG的MGC所请求。

由这样的MGC所涉及到的应用的一个例子就是使用TDM、ATM和帧中继的实时会话语音传输。

2 参考文献

下列ITU-T建议书和其他参考文献的条款，通过在本建议书中的引用而构成本建议书的条款。在出版时，所指出的版本是有效的。所有的建议书和其他参考文献都面临修订，使用本建议书的各方应探讨使用下列建议书和其他参考文献最新版本的可能性。当前有效的ITU-T建议书清单定期出版。本建议书中引用某个独立文件，并非确定该文件具备建议书的地位。

2.1 规范性参考文献

- ITU-T Recommendation Q.543 (1993), *Digital Exchange Performance Design Objectives*.
- ITU-T Recommendation Q.714 (2001), *Signalling connection control part procedures*.

2.2 资料性参考文献

- [1] Oftel, PNO-ISC Information Document 015 ISUP Overload Controls, Ref. PD 6673:2001.
- [2] WHITEHEAD (M. J.) and WILLIAMS (P. M.): Adaptive Network Overload Controls, *BT Technology Journal*, Vol. 20, No. 3, July 2002.

3 定义

本建议书规定下列术语：

- 3.1 call 呼叫：**在该包中，“呼叫”意味着普通的术语，仅仅对MGC有意义，表示建立相关的路径以通过MG承载终端用户之间的用户数据。通常的，有必要在“呼叫”前后添加限定词以明确所考虑的方面，如呼叫尝试。
- 3.2 call attempt 呼叫尝试：**在该包中，“呼叫尝试”表示去建立呼叫的尝试。
- 3.3 load 负载：**在该包中，“负载”表示在给定时间间隔内的“呼叫尝试”的总数目（即所提供的负载）。

该定义的依据在于与ITU-T Q.543建议书的目标保持一致。

3.4 MG overload and MG capacity MG过载和MG容量: 在该包中, “MG Overload”意味着MG(或虚拟MG)几乎无法以充分及时的方式来响应MGC的事务, 从而避免主叫客户放弃处于建立之中的呼叫。

该定义的依据是基于最大化MG所能够处理的呼叫/秒的需要, 这样使大部分MG的响应时间足够短以防止客户在端到端的主叫和被叫之间的呼叫建立之前放弃该呼叫。

在该包中, “MG Capacity”意味着呼叫速率(呼叫/秒), MG通过检测该速率来判断MG过载。因此, MG Capacity依靠H.248.1命令(ADD、MODIFY等等)的特殊序列, 这些序列实现呼叫并占用MG的处理能力。

MG过载检测的精确方式很可能是特定实现的, 因为不同供应商的交换机很有可能具体不同的结构。例如, 过载检测可以是基于处理器占用门限的使用, 或者队列门限, 或内在的延迟门限。每种过载检测机制将具有一个可配置参数的相关集合, 当然过载控制包通常是不知道该集合的。无论怎样, 操作者将需要能够配置MG的过载检测机制以确保在过载情况下具有适宜的短的响应时间。因此, 对于MG/MGC的实现而言, 有必要为操作者提供一种方法来配置这些参数, 例如, 通过专用的管理接口, 或使用SNMP, 而不是H.248.1接口。第9节给出了对于该包的管理需求。

可能存在MG的某些资源(例如编解码器、ATM带宽、音调产生器), 这些资源的拥塞并不妨碍MG对MGC的事务的及时响应, 尽管妨碍了呼叫的建立。这些资源的拥塞必须不会导致MGC过载控制(如该包所定义的控制)启动。其原因为:

- a) 这些资源的拥塞并不会妨碍对MGC事务的及时响应; 而且
- b) 为了触发所拥塞资源的容量更新, 通过现有的H.248.1错误消息来标记出这些拥塞可能会更合适。

3.5 leaky bucket 漏桶: 该包要求MGC限制其对MG所允许的呼叫速率, 该MG已经报告处于拥塞状态。任何下列类型3漏桶都可接受用作具有该包的呼叫限流器, 只要其实现能够配置以覆盖所要求的允许速率范围, 该速率由8.5节中给出的过载控制实例所包含。近似的性能分析见第9节。

Type 1 Leaky Bucket 类型1漏桶: 这是一个计数器, 由可配置的LeakAmount在每个LeakInterval时递减(不会低于0), 由可配置的SplashAmount在每次呼叫间隔递增(不会超过可配置的MaximumFill计数器)。某个呼叫到达时, 计数器小于或等于{MaximumFill – SplashAmount}, 则该呼叫被允许(且计数器由SplashAmount递增); 否则呼叫被拒绝且计数器不递增。最大可持续的呼叫允许速率大约为 (LeakAmount/SplashAmount) /LeakInterval, 只要满足LeakAmount ≤ MaximumFill。LeakInterval是由过载控制进行自适应改变的。

Type 2 Leaky Bucket 类型2漏桶: 这是一个计数器, 由(Now – TimeOfLastDecrement) × LeakAmount/LeakInterval在每次呼叫到达瞬间递减(不会低于0)。如果计数器小于或等于{MaximumFill – SplashAmount}且呼叫被允许, 则计数器由SplashAmount递增; 否则呼叫被拒绝且计数器不递增。最大可持续的呼叫允许速率是(LeakAmount/SplashAmount)/LeakInterval, 只要LeakAmount ≤ MaximumFill。LeakInterval是由过载控制进行自适应改变的。

Type 3 Leaky Bucket 类型3漏桶：这是一个计数器，由LeakAmount在每个LeakInterval时递减（不会低于0）。LeakInterval是一个操作者可配置的固定时间间隔；LeakAmount是由过载控制进行自适应改变的。某个呼叫到达时，计数器小于或等于{MaximumFill – SplashAmount}，则该呼叫被允许（且计数器由SplashAmount递增）；否则呼叫被拒绝且计数器不递增。最大可持续的呼叫允许速率大约为 $(\text{LeakAmount}/\text{SplashAmount})/\text{LeakInterval}$ ，只要满足 $\text{LeakAmount} \leq \text{MaximumFill}$ 。

对于三种漏桶类型，SplashAmount和LeakAmount都不应超过MaximumFill。

漏桶限流器的使用在电话交换机中已经很普遍。这些限流器简单易于实现且更容易实现按比例丢弃，因为限流器被限制在最大可持续速率范围内，而与所提供的呼叫速率无关。

4 缩写

本建议书采用下列缩写：

ATM	异步转义模式
IP	互联网协议
MG	媒体网关
MGC	媒体网关控制器
OCP	过载控制包
SCCP	信令连接控制部分
SNMP	简单网络管理协议
TDM	时分复用

5 过载控制包

Package Name（包名称）： OCP

PackageID（包标识）： ocp, 0x0051

Description（描述）：

该包使得MG（或虚拟MG）控制其负载成为可能，以便MG能够以及时的方式来处理从其MGC所接受的所有事务，同时使过载的MG能够最大化其有效吞吐量（呼叫/秒）。

该包中的事件可以在MG中提供。

该包中的事件可能仅仅使用于根终结点（Root Termination）。

Version（版本）： 1

Extends（扩展）： 无

5.1 属性

未定义。

5.2 事件

5.2.1 MG_Overload

Event name（事件名称）： MG_Overload

EventID（事件标识）： mg_overload,（0x0001）

Description（描述）：

该事件仅仅在当MG（或虚拟MG）接收到来自MGC的ADD命令且MG确定其处于过载状态的时候发生。该事件由MGC定制或提供。

EventsDescriptor Parameters:

无。

ObservedEventsDescriptor Parameters:

无。

6 信号

无。

7 统计

无。

8 程序

8.1 在过载的MG（或虚拟MG）上的行为

MG（或虚拟MG）应能够检测到何时其处于过载状态（“MG Overload”定义在3.4节）。

一个过载的MG（或虚拟MG）接收到来自MGC的ADD命令后应该

- a) 继续进行正常的事务处理；并
- b) 尽可能地通知MGC：该MG已经处于过载状态（通过向MGC发送带有事件“MG_Overload”的Notify请求）。

注 — 对于本建议书范围内所列举的应用，对每个呼叫而言，通常有两个来自MGC的ADD命令。因此，对每个呼叫而言，通常至多有两个这样的MG过载通知，而且这种通知的产生早于MG和MGC之间互换的事务处理序列。而且，过载MG返回这种通知的速率将收敛在一个可配置的低水平速率上（即TargetMG_OverloadRate，见8.2.3节）— 少于每秒1次。因此，发送“MG_Overload”事务应不会造成严重的处理能力下降或过多的传输开销。

8.2 在MGC上的行为

8.2.1 在MGC上的控制初始化

在MGC上的控制应对MG被激活（尽可能地快），当：

- a) 每秒钟MGC从MG接收到的MG Overload的速率超过了TargetMG_OverloadRate（见8.2.3节）；且
- b) 从MGC至MG的控制当前并未被激活。

当控制被激活时，漏桶计数器应被初始化设置为操作者可配置的值（InitialFill）。

注1 — 这允许操作者去限制所允许呼叫的初始突发量，通过将漏桶的InitialFill值设置为接近或等于其最大值。

对于类型1和类型2的漏桶限流器，当控制被激活时，LeakInterval应被初始化设置为操作者可配置的值（InitialLeakInterval）。

对于类型3的漏桶限流器，当控制被激活时，LeakAmount应被初始化设置为操作者可配置的初始值（InitialLeakAmount）。

初始限制层面的参数值的范围（即对于类型1和类型2漏桶的InitialLeakInterval参数，以及对于类型3漏桶的InitialLeakAmount参数）应允许操作者去设置初始的允许速率以达到所需要的低速率或高速率。

注2 — 即使在最坏的过载事件中，这允许控制能够快速响应初始出现的大量呼叫。

8.2.2 在MGC上的呼叫限制方法

当过载控制已经为某个MG进行初始化时，为了选择某个应被阻塞的呼叫，MGC控制实例应使用3.5节中所定义的3种类型的漏桶限流器之一，或者相当于这3种类型之一的漏桶实现。

注 — “相当于”意味着“严格地允许同样的呼叫和严格地拒绝同样的呼叫”，当出现任意顺序的呼叫到达实例时。

如果在MGC的过载控制实例允许呼叫建立请求，则该实例必须允许所有随后的与该呼叫相关的呼叫控制信令消息和因此产生的H.248.1事务。在MG过载情况下，该实例对所允许的呼叫给予高优先级以进行随后的呼叫处理，相对新呼叫而言。

8.2.3 在MGC上的允许呼叫速率的自适应

在MGC上保护MG（或虚拟MG）的过载控制实例应能够自适应MG的呼叫允许速率（“允许速率”），以便其所接收到的“MG_Overload”事务的速率能够收敛到操作者可配置的TargetMG_OverloadRate速率（定义为MG_Overloads/s）。

注1 — 该要求确保了提供给过载MG的总共呼叫/秒（也许来自多个MGC）将会聚至接近过载的MG的容量，只要（每个MGC控制实例的）TargetMG_OverloadRate足够小（即小于1个每秒）。全面的解释可以在文献[1]和[2]中找到。这些参考文献表明在大范围内的MG容量和大范围内的为数众多的MGC上会发生这种会聚。

注2 — 以这种方式控制拒绝速率意味着在稳态下，MGC具有激活控制的过载MG的容量共享正比于其TargetMG_OverloadRate值，注意到这一点是很重要的。特别的，如果所有的MGC具有相同的TargetMG_OverloadRate，则过载的MG容量将在MG之间进行均分。见参考文献[1]和[2]。

注3 — 在MGC上的允许速率的改变应该是越来越大（或这种变化更加频繁），因为MGC所检测到的MG_Overload速率越来越偏离TargetMG_Overload速率，这一点正是可期待的。这是为了确保对所提供呼叫速率中的突然变化（增加或减少）、或MG容量突然变化、或引起MG过载的MGC数量突然变化进行快速响应，而同时维持控制的稳定性。

注4 — 对于类型1和类型2的漏桶限流器，所允许的呼叫/秒的数目可通过调整LeakInterval来改变。对于类型3的漏桶限流器，所允许的呼叫/秒的数目可通过调整LeakAmount来改变。

注5 — 在恒定的负载下，接近于稳态的会聚意味着提供给过载MG的所允许呼叫/秒的数目变化不超过MG的呼叫/秒容量的小范围（10至20%）这表示控制变量（即类型1或2漏桶的LeakInterval，以及类型3漏桶的LeakInterval）需要能够适当地对待小的变化。例如，对于类型1或类型2漏桶，为了保持所提供税率的变化范围

位于MG容量的20%以内，LeakInterval将需要能够做的变化量为最小所要求的LeakInterval的1/5。控制变量范围和粒度的分析见9.1节，这些都由8.5节中规定的示例范围所要求。

如果MGC具有几个激活的过载控制实例，每个实例对应于一个过载的所属MG（或虚拟MG），则这些实例将完全独立运作。

8.2.4 在MGC上控制的终止

对某个MG（或虚拟MG）的负载控制应在MGC上终止，仅当以下两种情况出现时：

- a) MG向MGC发送MG_Overload的速率；且
- b) MGC漏桶限流器拒绝对MG的呼叫的速率

在足够长的时间区间内（TerminationPendingPeriod，以秒来计量）都为0就可以指示过载已经被缓和了。

注 — 当MG仅仅处于轻微过载时，这对防止控制重复地和快速的结束和重新启动（在其初期，有可能是低的允许速率）是很必要的。

8.2.5 在MGC上使用优先级

对于关联，H.248.1提供了可选的16个优先级级别，编号从0（最低优先级）到15（最高优先级），并为关联提供了可选的紧急指示符。本建议书将紧急指示符视为一种附加优先级，其优先层次在16个优先级级别之上，是对优先级的一个扩展集。

在这里规定的优先级的使用是基于Q.714的2.6节和Q.714的5.2.4节所描述的SCTP流量限制机制的。

其基本思想即：在稳态时，在MGC上的控制应允许（从而为MG提供呼叫）呼叫速率接近于MG的容量，通过拒绝足够多的低优先级呼叫来达到驱使MG_Overload速率接近于TargetMG_OverloadRate的要求；而且如果这样做不可能，则通过拒绝所有的低优先级呼叫和一些（也可能是全部）邻近高优先级呼叫来实现，拒绝呼叫以此类推。最终，控制将达到一个稳定状态，而此时的优先级为0, 1, ..., P-1的所有呼叫都被拒绝（对于某个优先级P而言），某些优先级为P的呼叫也被拒绝，而优先级高于P的呼叫则不会被拒绝（如果有这样的呼叫）。优先级P被称为HighestControlledPriorityLevel。（大写字母P用于在本节中用于区别小写字母p，p用于表示任意优先级）。

其基本思想在图1中描述，图1描述了由单个MGC所允许并提供给MG的总呼叫/秒是一个时间的函数。图1假设具有优先级0、1和2的呼叫共同使MG过载。其所提供的呼叫速率（在MGC限制之前）以堆叠增加的方式显示。纵轴指示MG的呼叫/秒的容量。在这个例子中，MG容量位于优先级2所提供的速率和优先级1和优先级2所提供的总速率之间。在稳态下，控制应拒绝所有优先级0的呼叫，大约拒绝一半优先级1的呼叫，且不拒绝优先级为2的呼叫。

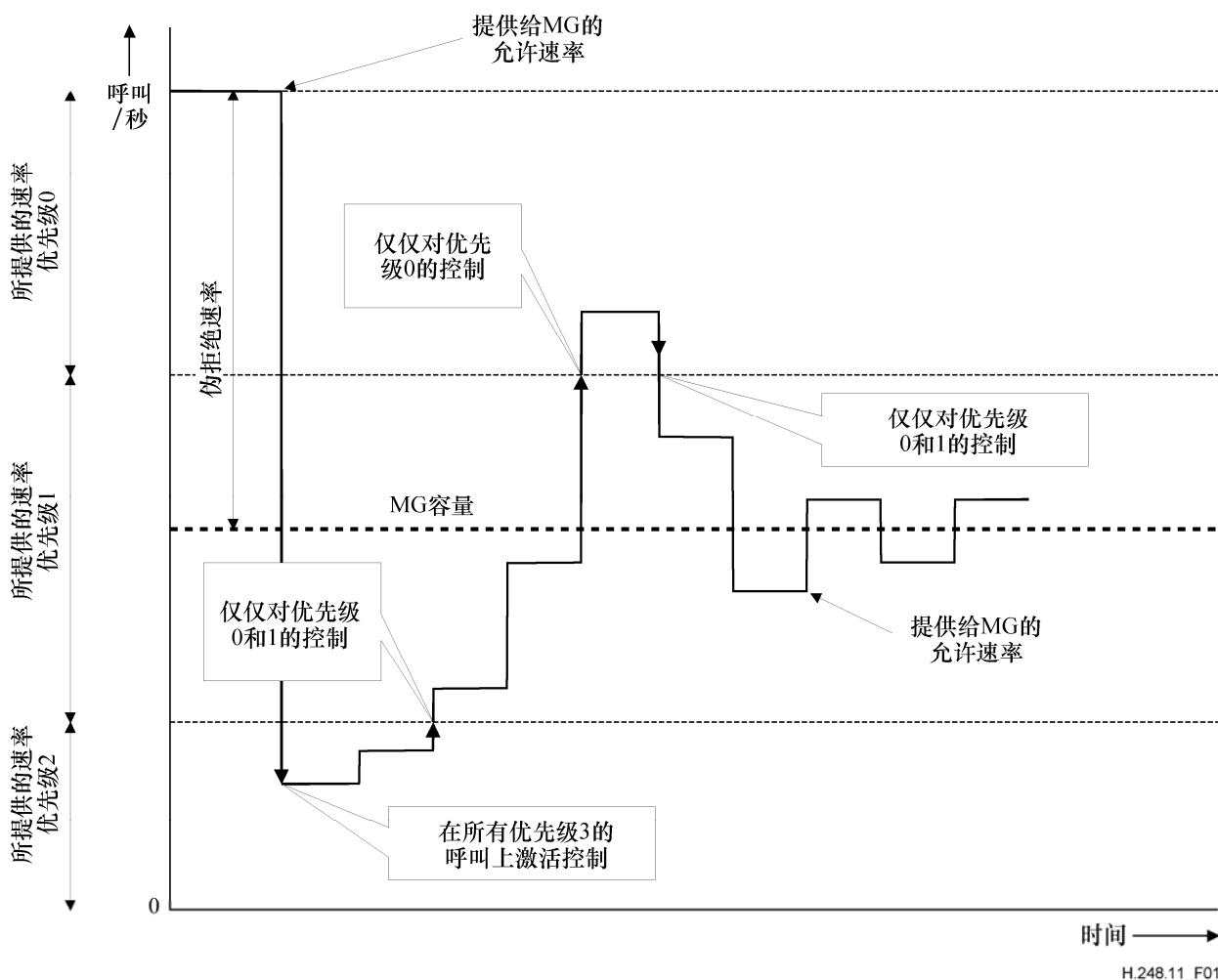


图 1/H.248.11—使用优先级控制的示例

在这个例子中，当过载开始时，MG_Overload速率（该速率大约等于所提供的呼叫速率减去MG的容量）足够快以激活在MGC上的过载控制。控制启动时将HighestControlledPriorityLevel初始化为（可配置的）InitialHighestControlledPriorityLevel，在这个例子中设置为2。相关的漏桶监视器也被初始化和激活。任何优先级低于HighestControlledPriorityLevel的呼叫建立请求都将被控制所拒绝。优先级对于HighestControlledPriorityLevel的呼叫建立请求将被提交给激活的漏桶限流器以确定呼叫是否被接纳或拒绝。

在这个例子中，控制最初超出限制范围（因此MG不返回MG_Overload通知），而限流器的允许速率能够自适应地增加到最大，在例子中所有优先级为2的呼叫都被允许接纳，但是MG_Overload速率仍然很低。因此HighestControlledPriorityLevel减少1且相关的漏桶限流器具有初始化的允许的速率和计数器（初始化为其各自的可配置最小和最大值以对应于最大优先级1的限制）。

限流器继续（在这个例子中）自适应地增加其所允许的速率（同时保持HighestControlledPriorityLevel为1）且最终达到其可配置的最大值，在到达最大值的时候，控制将HighestControlledPriorityLevel减少至0，并象从前一样重新初始化和重新激活漏桶限流器。

在最大值时，在例子中，控制仍然进行限制且接收到了MG_Overload通知。这些通知导致MGC过载控制实例增加限制，通过自适应地减少漏桶的所允许速率直至其最小值（表示对优先级为0的呼叫的最大限制）。在最大值时，MG_Overload速率仍然太高了，因此HighestControlledPriorityLevel增加到1且漏桶限流

器具有重新初始化的允许的速率和计数器（初始化为其各自的可配置最小和最大值以对应于最小优先级1的限制）。

此后，在这个例子中，控制保持HighestControlledPriorityLevel恒定为1（因此拒绝所有优先级为0的呼叫和某些优先级为1的呼叫）并自动调整漏桶的所允许速率以便使MG_Overload速率收敛接近至其目标速率。

以下为了总结优先级的使用：

- a) 控制的目的在于自适应调整其所允许的速率以便MG_Overload速率能够收敛接近至TargetMG_OverloadRate；
- b) 控制总是在高优先级呼叫前拒绝低优先级呼叫；
- c) 控制具有可配置的InitialHighestControlledPriorityLevel以快速地激活有效的控制，而不必从中间优先级开始进行自适应调整；而且
- d) 无论使用多少个优先级，仅仅需要一个漏桶限流器。

以上例子用于解释下列观点。下面的文本陈述了包的要求。

HighestControlledPriorityLevel应采取从MinimumHighestControlledPriorityLevel至MaximumHighestControlledPriorityLevel（这两个值都是操作者可配置的）范围内的整数值。

当控制在MGC上对某个特定MG被激活时，HighestControlledPriorityLevel应设置为对于（操作者可配置的）InitialHighestControlledPriorityLevel的值，且漏桶限流器应被激活和初始化。如果限流器是类型1或2，则其计数器应设置为InitialFill且其LeakInterval应设置为InitialLeakInterval。如果限流器是类型3，则其计数器应设置为InitialFill且其LeakAmount应设置为InitialLeakAmount。

当优先级“p”的呼叫建立请求到达某个激活的MGC控制时，则控制应按照如下处理：

- a) 如果 $p < \text{HighestControlledPriorityLevel}$ ，则呼叫被拒绝；
- b) 如果 $p = \text{HighestControlledPriorityLevel}$ ，则呼叫被提交给激活的漏桶限流器以确定该呼叫是否被接纳或拒绝；
- c) 如果 $p > \text{HighestControlledPriorityLevel}$ ，则呼叫被接纳。

对于HighestControlledPriorityLevel的每个值而言，控制应自适应调节漏桶的所允许速率以便使其接收到MG_Overloads的速率收敛至TargetMG_OverloadRate。

控制应自适应调整HighestControlledPriorityLevel以便使MGC接收到MG_Overload的速率收敛至TargetMG_OverloadRate。特别是：

- a) 增加HighestControlledPriorityLevel。如果MG_Overload速率超过了TargetMG_OverloadRate（即控制处于限制范围内）且漏桶的所允许速率等于其（可配置的）最小速率（因此对优先级为HighestControlledPriorityLevel的呼叫限制最大），则HighestControlledPriorityLevel的值应增加（不会超过MaximumHighestControlledPriorityLevel），漏桶的计数器应设置为MaximumFill且LeakInterval（或LeakAmount）应设置为其（可配置的）MinimumLeakInterval（或MaximumLeakAmount）以对应于新HighestControlledPriorityLevel上的最小限制。
- b) 减少HighestControlledPriorityLevel。如果MG_Overload速率低于TargetMG_OverloadRate（即控制超出限制范围）且漏桶的所允许速率等于其（可配置的）最大速率（因此对优先级为HighestControlled

PriorityLevel的呼叫限制最少），则HighestControlledPriorityLevel应减少（不会低于MinimumHighestControlledPriorityLevel），漏桶的计数器应设置为其（可配置的）MaximumFill且LeakInterval（或LeakAmount）应设置为其（可配置的）MaximumLeakInterval（或MinimumLeakAmount）以对应于新HighestControlledPriorityLevel上的最大限制。

8.3 在稳态下限定MG的响应时间

以下应该可能要去配置：

- a) MG过载检测；且
- b) MGC限制

在过载MG（或虚拟MG）上，对呼叫建立请求的H.248.1事务处理的响应时间而言，在稳态下，由控制所获得的95%的响应时间不应超过TargetMG_ResponseTime毫秒。

响应时间定义为从呼叫建立请求到达直到响应到达（见图2）。（例如，对于ISUP，响应时间是从IAM的到达直到相关的IAM被转发或接收到返回的ACM或REL。）

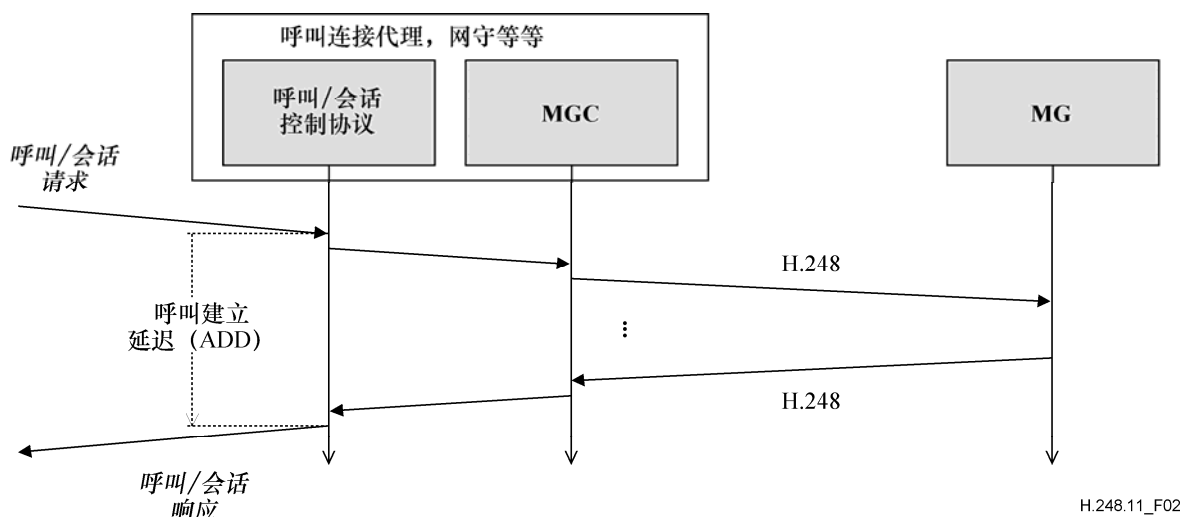


图 2/H.248.11—响应时间

注1 — TargetMG_ResponseTime应足够短以减少用户在呼叫建立过程中放弃呼叫的可能性。所建议的范围从0毫秒到100毫秒。

注2 — 这种要求服务于双重目的：首先确保，当控制已经停留在稳态时，用户不会在呼叫建立起来之前由于呼叫建立次数（这有必要限制用户发起的反复呼叫尝试次数）而放弃呼叫。这种要求也限制了从MGC到过载的MG再到MGC的往返时延，这样有助于改善过载控制的稳定性。

8.4 在初始过载瞬间限定提供给MG的速率

配置MG过载检测和MGC限制应该是有可能的，因此在控制的初始瞬间响应（即在到达稳态之前）过程中，由激活的MGC限流器所允许的并提供给过载的MG的总呼叫速率并未过多超过在过载的MG上所检测到的速率，该速率是在初始瞬间响应的连续的短时间测量期内测量的。

所允许的呼叫速率（以呼叫/秒计算）定义为在测量期内所允许的呼叫建立数目。

注一 这种要求确保了过载控制能够响应足够快以防止提供给过载的MG的负载不至于严重超过MG的容量。

8.5 过载事例的范围

该包已经设计用于在实时语音业务中的过载事例，其特征为：

- 使MG过载的MGC编号范围为1至10；
- 在提供给MGC（和去往过载MG）的总负载从0-5倍的MG容量进行“渐增”，而总负载保持高水平20分钟；
- 在20秒内，在提供给MGC（和去往过载MG）的总负载中快速的“斜坡增长”到达5倍的MG容量，随后10分钟内总负载呈现慢斜坡减少；
- 在MGC之间对总的所提供的负载进行任意分配；
- 过载MG的能力位于呼叫负载范围50呼叫/秒至500呼叫/秒。

注一 控制也可以适用于其他过载事例。

9 对H.248.11的管理要求

本节为本建议书中所定义的包收集了所有的管理要求。这些要求汇集在一起向MG/MGC的操作者提供了配置MGC和MG上所使用的所有过载控制参数的方法。这些要求将通过专用的管理接口或者使用SNMP来实现，这正是所期望的。

H.248.11的一个很重要的特性就是并不是所有的过载控制参数都能够预先知道。这是因为本建议书中的包并未详细无遗地规定过载控制应该如何实现。而且，建议仅仅规定了包的行为，例如，MGC应自适应调整其提供给MG的所允许速率以便其从MG接收到MG_Overload通知/秒的速率收敛至TargetMG_OverloadRate。因此，当对控制的某种实现完成时，才能够仅仅知道某些参数。

9.1 漏桶限流器的大致性能分析

本节给出了H.248.11所允许的类型1和类型3漏桶的大致性能分析。这应有助于控制的实现来确定：

- a) 应使用哪种漏桶类型；且
- b) 什么因素影响不同的可配置参数所要求的范围和粒度。

考虑这样的一般性的事例，其中N个MGC（编号1...N）共同使单个MG过载，超过其容量C呼叫/秒。

定义下列符号：

γ_i = 在MGC i 上由激活的控制所允许的呼叫/秒的数目，因此提供给MG

m_i = MGC i 上的MaximumFill

l_i = MGC i 上的LeakAmount $\leq m_i$

$s_i = \text{MGC } i \text{ 上的 SplashAmount} \leq m_i$

$t_i = \text{MGC } i \text{ 上的 LeakInterval (秒)}$

在大量的漏桶泄漏间隔上，漏桶填充增加的平均速率由 γs_i /秒给定，因此每个所允许呼叫导致漏桶填充增加 SplashAmount。而且，在每个漏桶泄漏间隔，漏桶填充泄漏为 l_i （或依据在泄漏瞬间的填充水平高度）。因此，漏桶填充减少的真实平均速率为 $\leq \frac{l_i}{t_i}$ 每秒。对于长期运行的所允许速率从而可得到以下上限：

$$\gamma \leq \frac{l_i}{s_i t_i}$$

可以考虑进行一些简化，假设提供给MGC i 的呼叫/秒对于所获得的最大可持续允许速率而言已足够大。在这种情况下，对于具有上限的真实允许速率的模拟对比显示了上限小于 1.1γ ，这对随机选择的漏桶而言，对75%的漏桶都成立。因此可以推断得到一个近似，即

$$\gamma_i = \frac{l_i}{s_i t_i}$$

现在假设所有 N 个MGC具有同样的TargetRejectRate参数值。则应遵循文献[1]和[2]，对于过载的MG的呼叫/秒的容量，每个MGC获得相等的份额。即

$$\gamma_i = \frac{l_i}{s_i t_i} = \frac{C}{N}$$

只要MaximumFill的值大于或等于LeakAmount: $m_i \geq l_i$ 。

这对每个MGC都成立，而且不管漏桶类型是1或3都成立。

现在有两种情况需要分析。情况 1在MGC i 上具有类型1漏桶且其控制变量是LeakInterval，情况 2在MGC i 上具有漏桶且其控制变量是LeakAmount（即类型3漏桶）。在每种情况下，我们都为相关的可配置参数推导出表达式。

情况1：控制变量 = LeakInterval

对于这种情况，在稳态下，LeakInterval由以下给定：

$$t_i = \frac{l_i}{s_i} \cdot \frac{N}{C}$$

在LeakInterval上的小改变 dt_i 应导致所允许速率的足够小的改变，即：

$$|d\gamma| \leq f \cdot \frac{C}{N}$$

其中 f 是一个小数（即20%）。这个不等式将保持成立，如果有

$$\left| \frac{d\gamma_i}{dt_i} \right| \cdot dt_i = \left| \frac{-l_i}{s_i t_i^2} \right| \cdot |dt_i| \leq f \cdot \frac{C}{N}$$

也就是，如果

$$|d\gamma| \leq f \cdot t_i$$

从这里，可以推导出在LeakInterval上的最小必然改变：

$$\min |dt_i| = f \cdot \min t_i = f \cdot \frac{l_i}{s_i} \cdot \frac{\min N}{\max C} = f \cdot \frac{l_i}{s_i} \cdot \frac{1}{500}$$

只有使这个最小改变等于MGC i 上的最短精确有效的时钟间隔才有意义，这个间隔用 τ_i 表示。典型地， τ_i 将位于10-100毫秒范围内。所以，对于所需要的LeakInterval粒度、LeakInterval所需要的最小值和最大值，可以获得以下简单的公式：

$$\begin{aligned} \min |dt_i| &= \tau_i \\ \min t_i &= \frac{l_i}{s_i} \cdot \frac{\min N}{\max C} = \frac{l_i}{s_i} \cdot \frac{1}{500} = \frac{\tau_i}{f} \\ \max t_i &= \frac{l_i}{s_i} \cdot \frac{\max N}{\min C} = \frac{l_i}{s_i} \cdot \frac{10}{50} = 100 \cdot \frac{\tau_i}{f} \end{aligned}$$

（这些等式里的数值来自8.5节。）

最后，由于MaximumFill必须不小于LeakAmount，即 $m_i \geq l_i$ ，可以推导出在每个LeakInterval内所允许的呼叫的最大数目，即 $\frac{m_i}{s_i}$ ，必须满足以下不等式：

$$\frac{m_i}{s_i} \geq \frac{l_i}{s_i} = 500 \cdot \frac{\tau_i}{f}$$

为了描述这些结果，假定 $f = 0.2$ 并在表格里列举4个例子。

表 1/H.248.11—情况 1：控制变量 = LeakInterval

最短时钟周期 τ_i (秒)	LeakInterval 粒度 $ dt_i $ (秒)	最小LeakInterval $\min t_i$ (秒)	最大LeakInterval $\max t_i$ (秒)	在每个LeakInterval内 所允许的最大呼叫 $\max \frac{l_i}{s_i}$
0.002	0.002	0.01	1	5
0.01	0.01	0.05	5	25
0.05	0.05	0.25	25	125
0.1	0.1	0.5	50	250

对于更大值的时钟周期，该表揭示了一个潜在的严重问题。在每个LeakInterval上，对于更大值的时钟周期（50和100毫秒），漏桶将趋向于允许更大的呼叫突发，而与泄漏间隔的持续期无关。（即使对10毫秒的时钟周期，所提供的高速呼叫速率将导致在每个泄漏间隔开始时产生25个呼叫的突发）。这将导致在过载的MG上产生密集的突发呼叫到达，这样可能会降低过载控制的效果。

情况2: 控制变量= LeakAmount

对于这种情况，在稳态下，LeakAmount由以下给定：

$$l_i = t_i s_i \cdot \frac{C}{N}$$

在LeakAmount上的小改变 dl_i 应导致所允许速率的足够小的改变，即：

$$|d\gamma| \leq \frac{C}{N}$$

其中f是一个小数（即20%）。这个不等式将保持成立，如果有

$$\left| \frac{d\gamma_i}{dl_i} \right| \cdot dl_i = \left| \frac{1}{s_i t_i} \right| \cdot |dl_i| \leq f \cdot \frac{C}{N}$$

也就是，如果

$$|d\gamma| \leq f \cdot l_i$$

从这里，可以推导出在LeakAmount上的最小必要改变（LeakAmount的粒度）：

$$\min |dl_i| = f \cdot \min l_i = f s_i t_i \cdot \frac{\min C}{\max N} = f s_i t_i \cdot \frac{50}{10} = 5 f s_i t_i$$

使LeakInterval t_i 等于MGC i 上的最短时钟周期 τ_i ，对于所需要的LeakAmount粒度、LeakAmount所需的最小值和最大值，可以获得以下简单的公式：

$$\begin{aligned} \min |dl_i| &= 5 f s_i \tau_i \\ \min l_i &= s_i \tau_i \cdot \frac{\min C}{\max N} = s_i \tau_i \cdot \frac{50}{10} = 5 s_i \tau_i \\ \max l_i &= s_i \tau_i \cdot \frac{\max C}{\min N} = s_i \tau_i \cdot \frac{500}{1} = 500 s_i \tau_i \end{aligned}$$

最后，由于MaximumFill必须不小于LeakAmount，即 $m_i \geq l_i$ ，可以推导出在每个LeakInterval内所允许的呼叫的最大数目，即 $\frac{m_i}{s_i}$ ，必须满足以下不等式：

$$\frac{m_i}{s_i} \geq \max \frac{l_i}{s_i} = 500 \tau_i$$

为了描述这些结果，假定 $f = 0.2$ ，（稍微任意地）选择SplashAmount $s_i = 100$ ，并在表格里列举4个例子。

表 2/H.248.11—情况 2：控制变量 = LeakAmount

最短时钟周期 τ_i (秒)	LeakAmount 粒度 $ dl_i $ (秒)	最小LeakAmount $\min l_i$ (秒)	最大LeakAmount $\min l_i$ (秒)	在每个LeakInterval 内所允许的最大呼叫 $\max \frac{l_i}{s_i}$
0.002	0.2	1	100	1
0.01	1	5	500	5
0.05	5	25	2500	25
0.1	10	50	5000	50

表2揭示了在类型1漏桶上的改善（与表1相比）。即在每个LeakInterval内所允许的呼叫的最大数目现在仅仅是类型1漏桶的1/5。例如，对于时钟周期50毫秒和类型3的漏桶，在每个突发内最多允许25个呼叫（并且是仅仅对最高要求的所允许速率），然而对于同样的时钟周期，类型1漏桶在每个LeakInterval内最多允许125个呼叫，这都是对所允许速率的整个范围而言的。

从上述分析所得到的结论是当设计H.248.11控制的实现时，控制实现者应做这种性能研究（和更多）。

另一个结论就是短时钟周期（理想的周期是小于50毫秒）的有效性是满足H.248.11的要求的一个重要部分。

第三个结论是类型3漏桶与类型1漏桶相比具有极大的优势，因为类型3漏桶并不完全地依赖短时钟周期来获得所需要的控制变量粒度。

9.2 在MGC上的漏桶限流器的配置

如果使用类型1或类型2的漏桶，则参数MaximumFill、LeakAmount和SplashAmount应该是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。在本建议书所定义的过载事例的完整范围内，实现应确保每个这些参数的范围和粒度都是可配置的且足够满足H.248.11的要求。

如果使用类型3的漏桶，则参数MaximumFill、LeakInterval和SplashAmount应该是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。在本建议书所定义的过载事例的完整范围内，实现应确保每个这些参数的范围和粒度都是可配置的且足够满足H.248.11的要求。

如果相当于这三种类型漏桶之一的漏桶限流器由过载控制的实现所使用，则相当于MaximumFill、LeakAmount、LeakInterval和SplashAmount的参数应该是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。在本建议书所定义的过载事例的完整范围内，实现应确保每个这些参数的范围和粒度都是可配置的且足够满足H.248.11的要求。

注 — “相当于”意味着“严格地允许同样的呼叫和严格地拒绝同样的呼叫”，当出现任意顺序的呼叫到达实例时。

如果提供了具有优先级划分功能的限流器，则InitialHighestControlledPriorityLevel、MinimumHighestControlledPriorityLevel和MaximumHighestControlledPriorityLevel应该是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。

9.3 与MG上的过载检测相关的专用参数的配置

实现应提供与检测MG Overload相关的参数集，这些参数（和呼叫限制相关的可配置参数一起）允许操作者去配置MG过载检测以便满足在8.3节内所规定的MG要求。这些（实现特定的）参数应是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。

9.4 与MGC上的控制激活相关的专用参数配置

如果控制使用类型1或类型2的漏桶限流器，则当控制实例激活时，应用在MGC上的InitialLeakInterval（秒）和InitialFill应是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。在本建议书所定义的过载事例的完整范围内，实现应确保这些参数的范围和粒度都是可配置的且足够满足H.248.11的要求。

如果控制使用类型3的漏桶限流器，则当控制实例激活时，应用在MGC上的InitialLeakAmount和InitialFill应是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。在本建议书所定义的过载事例的完整范围内，实现应确保这些参数的范围和粒度都是可配置的且足够满足H.248.11的要求。

实现应该为操作者提供某种方法以配置与控制激活相关的任何其他参数，这些参数的调整功能对于满足本建议书的要求可能是必要的。实现应规定每个这些参数的范围和可配置的粒度。

与控制激活相关的（实现特定的）参数应是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。

9.5 与MGC上所允许速率的自适应相关的专用参数的配置

TargetMG_OverloadRate（MG_Overloads/s）应是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。TargetMG_OverloadRate可能采用0-1中的任何数值，步长为0.1。TargetMG_OverloadRate的缺省值应为每秒0.5倍MG_Overload。

实现应该为操作者提供某种方法以配置与MGC上所允许速率（呼叫/秒）的自适应相关的任何其他参数，这些参数的调整功能对于满足本建议书的要求可能是必要的。实现应规定每个这些参数的范围和可配置的粒度。

由允许速率自适应方案所使用的（实现特定的）参数应是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。

9.6 在MGC上控制终止的配置

TerminationPendingPeriod（以秒来计量）应是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。TerminationPendingPeriod可能采用1-300秒中的任何数值，步长为1秒。TerminationPendingPeriod的缺省值应为120秒。

实现应该为操作者提供某种方法以配置与MGC上控制终止相关的任何其他参数，这些参数的调整功能对于满足本建议书的要求可能是必要的。实现应规定每个这些参数的范围和可配置的粒度。

由控制终止所使用的（实现特定的）参数应是操作者可配置的，通过专用的管理接口或通过使用SNMP来进行配置。

9.7 MGC统计

当MGC对某个MG（或虚拟MG）初始化过载控制（见8.2.1节），下列信息应由MGC记录以用于随后的检索，通过专用接口或使用SNMP进行：date、time、MGC identity、MG identity。

当MGC对某个MG（或虚拟MG）终止过载控制（见8.2.4节），则下列信息应由MGC记录以用于随后的检索，通过专用接口或使用SNMP进行：date、time、MGC identity、MG identity、所提供和所拒绝的总呼叫数目、在整个过载持续期的限流器。

ITU-T 系列建议书

A系列	ITU-T工作的组织
B系列	表示方法：定义、符号、分类
C系列	综合电信统计
D系列	一般资费原则
E系列	综合网络运行、电话业务、业务运行和人为因素
F系列	非话电信业务
G系列	传输系统和媒质、数字系统和网络
H系列	视听和多媒体系统
I系列	综合业务数字网
J系列	有线网和电视、声音节目及其他多媒体信号的传输
K系列	干扰的防护
L系列	线缆的构成、安装和保护及外部设备的其他组件
M系列	TMN和网络维护：国际传输系统、电话电路、电报、传真和租用电路
N系列	维护：国际声音节目和电视传输电路
O系列	测量设备技术规程
P系列	电话传输质量、电话装置、本地线路网络
Q系列	交换和信令
R系列	电报传输
S系列	电报业务终端设备
T系列	远程信息处理业务的终端设备
U系列	电报交换
V系列	电话网上的数据通信
X系列	数据网和开放系统通信
Y系列	全球信息基础设施和互联网的协议问题
Z系列	电信系统中使用的语言和一般性软件情况

30333