ASR5x00 MME过载保护功能

目录

简介

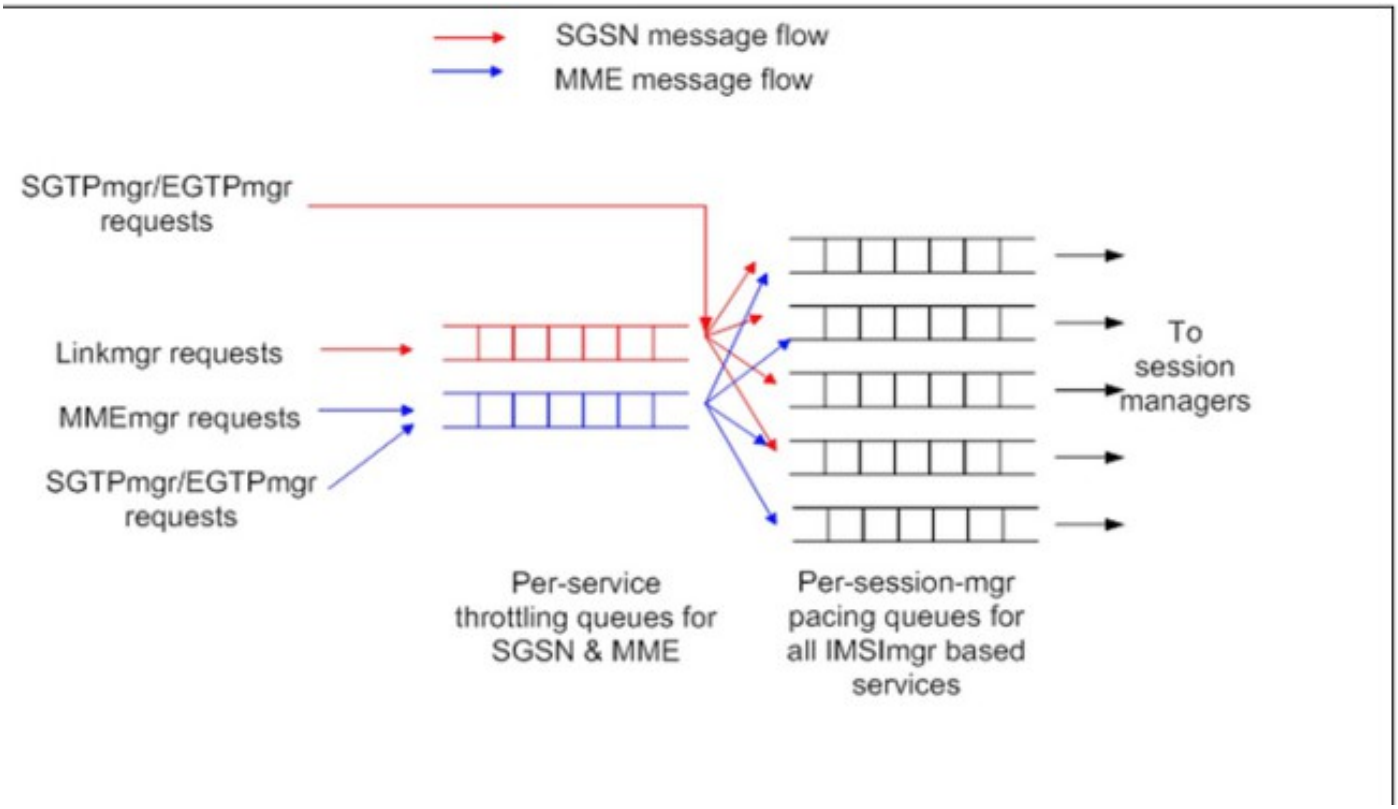
本文档重点介绍思科聚合服务路由器(ASR)5000系列上提供的各种移动管理实体(MME)过载保护方法和功能。在ASR 5000系列中，思科为客户提供了各种控制手段，本文解释了这些功能和相关的CLI命令。

MME保护

网络过载保护：配售率限制

附加速率限制可保护邻居网络元素，如家庭用户服务器(HSS)、策略和计费规则功能(PCRF)和在线计费服务器(OCS)，以及内部MME资源，如imsimgr和sessmgr。附加速率限制处理到达的新呼叫，例如连接和MME间/服务GPRS支持节点(SGSN)跟踪区域更新(TAU)。

此图显示呼叫和限制队列的消息流。



为了保护MME（imsimgr和sessmgr以后），应定义*限制速率*、*队列等待时间和*队列大小时间。由于*MME容量*取决于呼叫模型，因此限制速率取决于MME呼叫模型。

对于MME，限速率计算相对简单，以网络中的标准每秒呼叫事件数(CEPS)加容差为例。 此外，如果需要HSS保护，您可能还需要考虑HSS数据库容量。

示例

在繁忙时段，MME每秒最多可处理170到200个呼叫(Actons+ Inter TAU)。 如果一个站点发生故障，每秒最多可能会有350到370个呼叫到达一个MME。在此呼叫速率下，MME利用率将上升近80%，每秒*400个*呼叫是限制限制速率的最佳级别，以避免MME机箱内的过多信令负载。

默认*情况下，队*列等待时间为5秒。这对客户来说是最佳选择。默认*情况下*，队列大小为2500。这对客户来说是最佳选择。

配置命令如下所示。

```
asr5k(config)#network-overload-protection mme-new-connections-per-second
new_connections action attach { drop | reject-with-emm-cause
{ congestion | network-failure | no-suitable-cell-in-tracking-area}
tau { drop | reject-with-emm-cause { congestion | network-failure
| no-suitable-cells-in-tracking-area | no-sec-ctxt-in-nw} fwd-reloc
{ drop | reject} }{wait-time <wait-time>} {queue-size <queue-size>}
```

new_connections
定义每秒要接受的新MME连接数。必须是介于50和5000之间的整数。默认值为 500。

动作
定义定步队列变满时要执行的操作。每当在MME处收到新连接时，它们都会在步调队列中排队，并且imsimgr会以配置的速率处理来自队列的消息。当队列溢出（由于高传入速率）时，根据配置的“操作”，数据包将被丢弃或拒绝。

队列大小
定义用于缓冲数据包的步调队列的最大大小。必须是介于250和25000之间的整数。默认值为2500。

配置示例

```
network-overload-protection mme-new-connections-per-second 400 action attach
reject-with-emm-cause no-suitable-cell-in-tracking-area tau
reject-with-emm-cause no-suitable-cell-in-tracking-area fwd-reloc drop
```

现在，每秒呼叫率设置为400，并且操作是智能拒绝，原因为#15使用户设备(UE)重新连接到不同的无线接入技术(RAT)。 等待时间设置为默认值（5秒），队列大小为2500。

> **注意**：EMM原因为#15 "no-cull-in-tracking-area"的操作“reject”优先，因为#15拒绝的呼叫大多不会重新到达MME，并且将转到不同的RAT层(3G、2G)。 服务无线网络子系统(SRNS)重新定位的“丢弃”操作将供将来使用，并会防止在拒绝后快速重新连接到MME。

网络过载保护：寻呼限制

分页限制将内部MME资源(mmgr)作为eNodeB/无线电资源（如果需要）进行保护。 此速率限制阈值应适用于与MME关联的给定ASR 5000机箱的所有eNodeB。对eNodeB的S1寻呼请求应限制在此阈值的速率。对超过此阈值的eNodeB的S1寻呼请求将被丢弃。

对于MME，节流速率计算相对简单，采用网络中的标准出口寻呼速率加容差。（这完全取决于设计团队的决策。）

示例

在繁忙时段，每个MME每秒最多处理35000个寻呼消息。如果一个站点发生故障，一个MME每秒最多可能会发送70000页。在此寻呼速率下，MME利用率(memmgr)将上升近80%，每秒*70000到80000页将是限制带宽限制速率的最佳级别，以避免通过memmgr发送过多的S1信令。*

但是，每个平均eNodeB的速率有限。每eNodeB（在6500 eNodeB的情况下）的平均速率是*每秒*10页。但是，跟踪区域(TA)在用户数量上不相等，并且各种TA/成员eNodeB的分页加载方式不同。如果TA大小与每个TA的平均用户数之差为2倍，则每个eNodeB的速率为*20*。如果TA大小与每个TA的平均用户数之差为20倍，则每个eNodeB的速率为*200*。当TA（用户数）平均加载时，该功能将变得最高效。

另一个应同时采取的操作是激活智能寻呼。请参阅《ASR 5000 MME管理指南》中的“TAI mgmt db和LTE寻呼”部分。

配置命令如下：

```
asr5000(config)# network-overload-protection mme-tx-msg-rate-control enb s1-paging
```

- network-overload-protection标识网络过载保护
- mme-tx-msg-rate-control enb标识每平均eNodeB的MME消息速率控制
- s1寻呼标识S1寻呼的消息速率控制
- <rate>以每秒消息数为单位指定速率阈值每eNodeB — 范围(1到65535)

配置示例

```
network-overload-protection mme-tx-msg-rate-control enb s1-paging 200
```

> **注意：**
> — 速率限制是进一步调整的主题，方向是减小。调整的基础是TA上的用户数（寻呼数）（需要TA级统计信息）。
>
> — 当TA（按用户数/每TA分页数）平均加载时，该功能将变得最高效。

网络过载保护：DDN限制（服务GateWay功能，保护MME）

下行链路数据通知(DDN)限制是控制从服务网关(SGW)侧向MME发送DDN请求的速率的功能。它可保护MME资源（如mmemgr和sessmgr）免受DDN（即入口分页请求）浪涌的影响。

此功能有两个部分，一个用于符合Rel-10的MME，另一个用于不符合Rel-10的MME:

- 对于符合版本10的MME，在SGW服务中设置DDN限制分配和保留优先级(ARP)水印以启用该功能。
- 对于版本10不合规的MME，需要在SGW服务中与ARP水印（如限制因子、限制时间、稳定时间、轮询间隔等）一起设置一些其他参数。

当在SGW上启用此功能时，它会向MME发送DDN请求中的ARP水印。作为回复，MME发送限制延迟单元、限制延迟值和限制因子。延迟值和延迟单元的组合计算限制时间。收到这些值后，SGW会丢弃特定ARP的DDN请求，直到限制计时器过期。

对于使用本地配置的非Rel-10兼容MME，SGW将DDN Req与特定ARP水印一起限制。

Cisco ASR5x00 MME版本16和17不支持自动DDN限制，因此在DDN限制方面，它与非版本10兼容。

> **注意：**DDN限制在入口端(S11)而不是出口端(S1)的MME分页限制之上提供更精细的粒度。如果配*置了分页*限制，则Cisco不要求您实施DDN限制，但它提供了较早的过载检测和消除。

技术规范(TS)23.401,MME参考：

限制DDN请求

在异常情况下（例如，当MME负载超过运营商配置的阈值时），MME可能会限制其SGW在其上生成的信令负载（如果配置为这样做）。

MME可以拒绝UE在空闲模式下对低优先级流量的DDN请求或进一步卸载MME。MME可以请求SGW根据限制因子和DDN确认消息中指定的限制延迟，有选择地减少它为空闲模式下为UE接收的下行链路低优先级业务发送的DDN请求数。

SGW基于承载的ARP优先级和运营商策略（即，运营商在SGW中将ARP优先级配置为优先级或非优先级流量的配置）确定承载是否用于低优先级流量。 MME根据从SGW和运营商策略接收的ARP优先级确定DDN请求是否针对低优先级流量。

如果UE的空闲状态信令减少(ISR)不活动，在节流延迟期间，SGW丢弃在其所有低优先级承载上为

未连接的用户平面接收的下行链路分组(即，SGW上下文数据指示该MME不按节流系数提供下行链路用户平面隧道终端标识符(TEID))，并发送DDN消息仅针对非限制承载发送给MME。

如果ISR在限制延迟期间对UE处于活动状态，SGW不会向MME发送DDN，而仅向SGSN发送DDN。如果MME和SGSN都请求负载减少，SGW将丢弃在其所有低优先级承载上收到的下行链路数据包，这些下行链路数据包被称为未连接的用户平面（即，SGW上下文数据指示没有下行链路用户平面TEID），与限制因素成比例。

SGW在限制延迟到期后恢复正常操作。限制系数和限制延迟的最后接收值将取代从该MME接收的任何先前值。收到限制延迟会重新启动与该MME关联的SGW计时器。

对于SGW与MME，节流速率计算相对简单。采用允许的最大入口分页速率，即每MME框每秒1100条消息。

配置命令如下：

```
#configure

#context saegw-gtp

#sgw-service sgw-svc

#ddn throttle arp-watermark <arp_value> rate-limit <limit> time-factor <seconds>
throttle-factor <percent> increment-factor <percent> poll-interval <second>
throttle-time-sec <seconds> throttle-time-min <minutes> throttle-time-hour <hour>
stab-time-sec <seconds> stab-time-min <minutes> stab-time-hour <hour>
```

throttle arp-watermark arp_value
如果配置了ARP水印，并且MME/SGSN在DDN ACK消息中发送了限制因子和延迟，则ARP值大于配置值的所有DDN将被指定延迟的限制因子限制。
*arp_value*是1到15之间的整数。

速率限制
配置速率限制（仅当MME为非版本10 MME时，才使用此令牌和后续令牌进行速率限制）。
*limit*是介于1和999999999之间的整数。

时间因子秒
配置SGW做出限制决策的时长。
*seconds*是介于1和300之间的整数。

throttle-factor percent
配置DDN限制因子。输入在检测到DDN浪涌时要丢弃的DDN的百分比。
*百分*比是介于1和100之间的整数。

增量因子百分比
配置DDN限制增量因子。输入DDN限制应增加的百分比。
*百分*比是介于1和100之间的整数。

轮询间隔秒数
在DDN限制中配置轮询间隔。
*seconds*是介于2和999999999之间的整数。

throttle-time-sec seconds
配置DDN限制时间（以秒为单位）。输入DDN在SGW上被限制的时间段（以秒为单位）。

*seconds*是0到59之间的整数。

throttle-time-min minutes
配置DDN限制时间（以分钟为单位）。输入DDN在SGW上被限制的时间段（以分钟为单位）。
*分*钟是0到59之间的整数。

throttle-time-hour hour
配置DDN限制时间（以小时为单位）。输入DDN在SGW上被限制的时间段（以小时为单位）。
*hour*是0到310之间的整数。

stab-time-sec秒
配置DDN限制稳定时间（以秒为单位）。输入一个时间段（以秒为单位），如果系统稳定，则禁用限制。
*seconds*是0到59之间的整数。

stab-time-min分钟
配置DDN限制稳定时间（以分钟为单位）。输入一个时间段（以分钟为单位），如果系统稳定，则禁用限制。
*分*钟是0到59之间的整数。

stab-time-hour hour
配置DDN限制稳定时间（以小时为单位）。输入一个时间段（以小时为单位），如果系统稳定，则禁用限制。
*hour*是0到310之间的整数。

配置示例

```
ddn throttle arp-watermark 1 rate-limit RATE time-factor 120 throttle-factor 50
increment-factor 10 poll-interval 30 throttle-time-sec 0 throttle-time-min 1
throttle-time-hour 0 stab-time-sec 0 stab-time-min 2 stab-time-hour 0
```

- 1100页/秒是允许的最大入口速率（包括DDN）
- 1100页/秒（如果DDN浪涌对应1100 DDN/秒）
- 每个MME站点4xSGW >速率= 275 DDN/秒/SGW最大允许的区域
- 每个MME站点3xSGW >速率= 366 DDN/秒/SGW最大允许
- 每个MME站点2xSGW >速率= 550 DDN/秒/SGW最大允许的区域
- 每个MME站点1xSGW >速率= 1100 DDN/秒/SGW最大允许的区域

网络过载保护：EGTP路径故障限制

此功能可保护MME资源(sessmgr、memmgr)和4G资源，防止在IP主干和IP BackHaul中传输失败时EGTP路径故障浪涌，以及侧网元failures/restarts.The功能启用每会话限制检测到的EGTP路径故障事件，并在用户管理上定义更精细的粒度s1寻呼限制。根据空闲用户和已连接用户之间的划分，应设定限制。它非常特定于网络，需要根据eUTRAN和UE状态进行调整。

示例

用户被拆分为大约80:20空闲到已连接。在最坏情况下，IDLE用户的EGTP PF会导致寻呼浪潮，这可能导致memmgr过载，这是链中最窄的瓶颈。首先，此类EGTP寻呼因子(PF)浪涌（针对空闲用户）会导致寻呼浪涌，并且此浪涌会达到memmgr瓶颈，因此您需要首先保护memmgr免受此影响。因此，EGTP PF的IDLE可能被视为意外的入口寻呼浪涌，允许最大*1100*页/秒。

- 对于IDLE用户，建议的*限制限*制是1000 msg/秒。
- CONNECTED子网的数量比IDLE少5到7倍。
- CONNECTED用户不会发生寻呼浪涌，因此建*议将2000 msg/sec*安全地应用于CONNECTED用户。

注意：EGTP PF限制在入口端(S11, Sv)而不是出口端(S1)的MME分页限制上提供更精细的粒度。 如果配*置了寻呼*限制，思科不要求您实施EGTP PF限制，但它提供了较早的过载检测和消除。

此配置适用于接口类型为“interface-mme”的EGTP服务。

配置命令如下：

```
asr5000(config)# network-overload-protection mme-tx-msg-rate-control egtp-pathfail ecm-idle
< rate in sessions per second > ecm-connected < rate in sessions per second >
```

- network-overload-protection标识网络过载保护
- mme-tx-msg-rate-control标识MME消息速率控制
- egtp-pathfail识别EGTP路径故障的消息速率控制
- ecm-idle标识ECM-Idle模式下MME UE会话的速率
- ecm-connected标识ECM-Connected模式下MME UE会话的速率
- <rate in sessions per second>指定每秒会话数的速率阈值，范围为1到5000

配置示例

```
network-overload-protection mme-tx-msg-rate-control egtp-pathfail ecm-idle
1000 ecm-connected 2000
```

增强的拥塞控制

使用增强的拥塞控制功能，MME可以向它所连接的eNodeB发送信号，以将流量重定向到MME池中的其他MME。这通过S1接口过载程序（TS 36.300和TS 36.413）完成。

当配置了过载控制并达到拥塞阈值时，MME可以配置为向MME所连接的eNodeB的百分比发送S1AP接口过载开始消息。为了反映MME希望减少的负载量，此百分比是可配置的。在发送到eNodeB的过载响应信息元素(IE)中，MME可以请求eNodeB拒绝或允许特定类型的会话，包括：

- 拒绝非紧急会话
- 拒绝新会话
- 允许紧急会话
- 允许高优先级会话和移动终止服务
- 拒绝容迟访问

拥塞控制功能允许您设置策略和阈值，并指定系统在面临重负载情况时如何反应。拥塞控制监控系统是否存在在系统负载过重时可能降低性能的条件。通常，这些情况是临时的（例如，CPU或内存利用率高），并且会快速解决。但是，在特定时间间隔内持续或大量出现这些情况可能会影响系统为用户会话提供服务的能力。拥塞控制有助于识别此类情况并调用策略来解决该情况。

拥塞条件阈值

- 系统CPU使用率
- 系统服务CPU使用率（解复用卡CPU使用率）
- 系统内存使用
- 许可证使用情况
- 每个服务的最大会话数

阈值和容差级别

当您为严重、主要和次要拥塞级别配置阈值和容差时，阈值级别和容差不应重叠。考虑以下阈值级别不重叠的示例配置：

- 严重拥塞在95%时触发，在90%时清除
- 主要拥塞在90%时触发，在85%时清除
- 轻微拥塞在85%时触发，在80%时清除

服务控制CPU阈值

此阈值由系统的解复用CPU计算。阈值根据五分钟的平均CPU使用率计算。

考虑了解复用器CPU的两个CPU核心的最高CPU使用值。例如，如果CPU核心0的5分钟CPU使用率为40%，而CPU核心1的5分钟CPU使用率为80%，则CPU核心1将考虑用于阈值计算。

系统CPU阈值

此阈值使用所有CPU（备用CPU和SMC CPU除外）的五分钟CPU使用率平均值计算。

考虑所有CPU中两个CPU核心的最高CPU使用值。

系统内存阈值

此阈值是使用所有CPU（备用CPU和SMC CPU除外）的五分钟内存使用平均值计算的。

配置拥塞操作配置文件

拥塞操作配置文件定义一组操作，这些操作可在超过相应阈值后执行。

将拥塞操作配置文件与拥塞控制策略关联

每个拥塞控制策略（关键、主要、次要）必须与拥塞控制配置文件关联。

配置过载控制

当在MME上检测到过载情况时，系统可以配置为将该情况报告到指定百分比的eNodeB，并对传入会话执行配置的操作。

这些过载操作也可用（除reject-new-sessions外）：

- permit-emergency-sessions-and-mobile-terminated-services

- permit-high-priority-sessions-and-mobile-terminated-services
- reject-delay-tolerant-access
- reject-non-emergency-sessions

配置说明示例

这将启用拥塞控制功能：

```
congestion-control

This monitors the overall CPU Utilization including the sessmgr and demux mgrs

congestion-control threshold system-cpu-utilization critical 90

congestion-control threshold system-cpu-utilization major 85

congestion-control threshold system-cpu-utilization minor 80


Memory utilization thresholds:

congestion-control threshold system-memory-utilization critical 85

congestion-control threshold system-memory-utilization major 75

congestion-control threshold system-memory-utilization minor 70


CPU utilization on DEMUX card:

congestion-control threshold service-control-cpu-utilization critical 85

congestion-control threshold service-control-cpu-utilization major 75

congestion-control threshold service-control-cpu-utilization minor 70


Defining tolerance margins:

congestion-control threshold tolerance critical 5

congestion-control threshold tolerance major 5

congestion-control threshold tolerance minor 5
```

定义拥塞操作配置文件（严重、主要和次要）

```
lte-policy

congestion-action-profile criticalCogestionProfile

reject s1-setups time-to-wait 60

drop handovers

drop combined-attaches

drop service-request
```

```
drop addn-brr-requests

drop addn-pdn-connects

exclude-voice-events

exclude-emergency-events

report-overload permit-emergency-sessions-and-mobile-terminated-service enodeb-percentage 50

congestion-action-profile majorCogestionProfile

report-overload permit-emergency-sessions-and-mobile-terminated-service enodeb-percentage 50

congestion-action-profile minorCogestionProfile

report-overload permit-emergency-sessions-and-mobile-terminated-service enodeb-percentage 30

end
```

应用拥塞策略

```
configure

congestion-control policy critical mme-service action-profile criticalCogestionProfile

congestion-control policy major mme-service action-profile majorCogestionProfile

congestion-control policy minor mme-service action-profile minorCogestionProfile

end

.
```

相关信息

- Cisco ASR 5000移动管理实体管理指南
- 技术支持和文档 - Cisco Systems