Load Balancing and Rebalancing and VoLTE Offloading

Feature Description

The sections below describe the load balancing and rebalancing functionality available on the MME. The MME also supports VoLTE Offloading.

Load Balancing

Load balancing on the MME permits UEs that are entering into an MME pool area to be directed to an appropriate MME in a more efficient manner, spreading the load across a number of MMEs.

Load Rebalancing

The MME load rebalancing functionality permits UEs that are registered on an MME (within an MME pool area) to be moved to another MME in the pool. The rebalancing is triggered using an exec command on the mme-service from which UEs should be offloaded.

When initiated, the MME begins to offload a cross-section of its subscribers with minimal impact on the network and users. The MME avoids offloading only low activity users, and it offloads the UEs gradually (configurable from 1-1000 minutes). The load rebalancing can off-load part of or all the subscribers.

The eNodeBs may have their load balancing parameters adjusted beforehand (e.g., the weight factor is set to zero if all subscribers are to be removed from the MME, which will route new entrants to the pool area into other MMEs).

VoLTE Offloading

Offloading of a certain percentage of users can be configured using the **mme offload** command. The MME sends S1 Release (with cause "load balancing TAU required" for offload) to the configured percentage of

UEs attached to the MME. The MME does not distinguish between VoLTE and Non-VoLTE subscribers. Some subscribers with voice bearers are also offloaded as a result calls are dropped. This feature enhancement is targeted to preserve VoLTE voice bearers during MME offloading. A new CLI keyword is added to the **mme offload** command to preserve VoLTE subscribers (QCI = 1) from offloading until voice calls are terminated.

Note This feature enhancement is license controlled. Contact your Cisco Account or Support representative for information on how to obtain a license.

Relationships to Other Features

MME load balancing can be used in conjunction with congestion control. For more information on congestion control, refer to the *Congestion Control* section in the Mobility Management Entity Overview chapter of the *MME Administration Guide*.

How it Works

Load Balancing

Load balancing is achieved by setting a weight factor for each MME so that the probability of the eNodeB selecting an MME is proportional to its weight factor. The weight factor is set by the operator according to the capacity of an MME node relative to other MME nodes. The **relative-capacity** mme-service level command is used to specify this relative weighting factor.

Once set, the Relative MME Capacity IE is included in the S1AP S1 SETUP RESPONSE message from MME to relay this weight factor. If the relative MME capacity is changed after the S1 interface is already initialized, then the MME CONFIGURATION UPDATE message is used to update this information to the eNodeB.

Load Rebalancing

The MME uses the **mme offload mme-service** exec level command to enable the operator to offload UEs for a particular mme-service for load rebalancing among MMEs in a MME pool area. The command enables the operator to specify a percentage of UEs to offload, and the desired time duration in which to complete the offload.

The operator can also include the keyword option **disable-implicit-detach**. By default, if the UE context is not transferred to another MME within 5 minutes, the UE will be implicitly detached. This option disables this implicit detach timer.

To offload ECM-CONNECTED mode UEs, the MME initiates the S1 Release procedure with release cause "load balancing TAU required".

To offload UEs which perform TA Updates or Attaches initiated in ECM-IDLE mode, the MME completes that procedure and the procedure ends with the MME releasing S1 with release cause "load balancing TAU required".

To offload UEs in ECM-IDLE state without waiting for the UE to perform a TAU or perform Service request and become ECM CONNECTED, the MME first pages the UE to bring it to ECM-CONNECTED state.

Call Handling and Other Messaging Considerations

New calls are processed normally (as per the new call policy configuration). The offloading process does not reject INIT UE messages for new subscribers. To prevent new calls from entering this MME, set the **relative-capacity** on this mme-service to 0.

When Init UE messages are received for an existing offloaded subscriber, the ue-offloading state is set as MARKED and the offload procedure continues until the UE is offloaded.

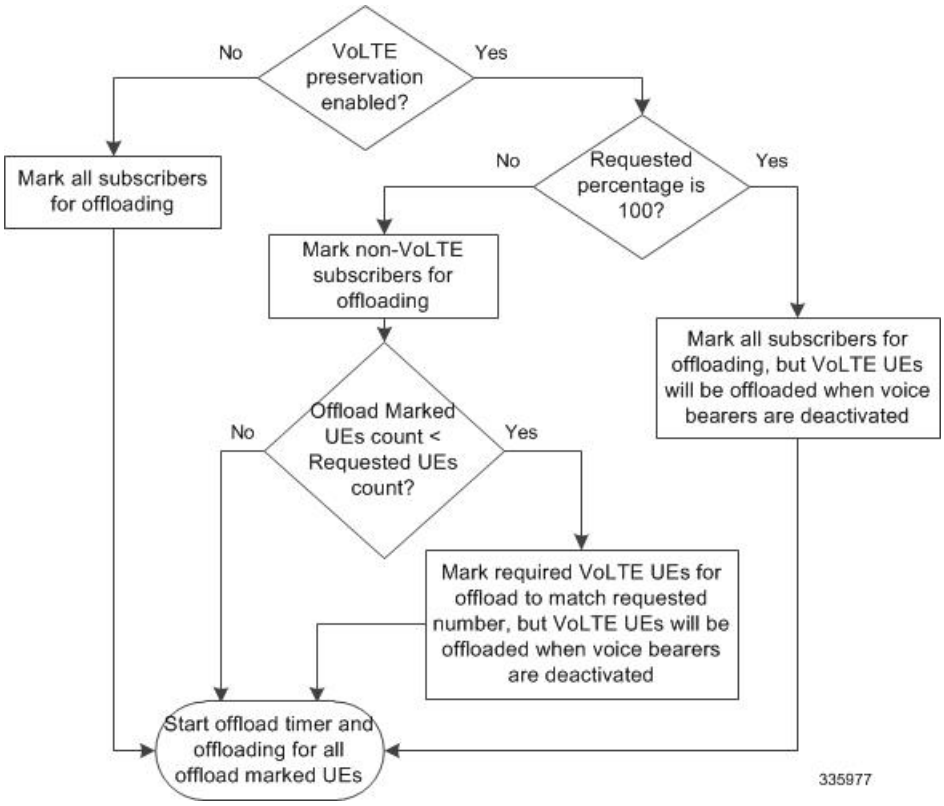
Once a UE is offloaded, messages such as EGTP events, Create bearer, Update bearer, Idle mode exit, and Paging trigger are be rejected. HSS initiated events also will be rejected for offloaded UEs.

Detach events are processed as usual.

Important Emergency attached UEs in Connected or Idle mode are not considered for offloading.

VoLTE Offloading

The **mme offload** command is enhanced with the keyword **preserve-volte-subscribers** , this keyword enables preservation of subscribers with voice bearers (QCI=1) from offloading until the voice bearers are deactivated. In any MME service both VoLTE and Non-VoLTE subscribers are present. The offload command now has options to configure the percentage of total subscribers to be offloaded and to preserve VoLTE subscribers from offloading until voice calls are terminated. With this feature enhancement if VoLTE preservation is not enabled, all subscribers are marked for offloading. But when the keyword **preserve-volte-subscribers** is enabled, Non-VoLTE subscribers are first marked for offloading based on configured offload-percentage. If the configured offload-percentage is greater than the available Non-VoLTE subscribers, VoLTE subscribers are also marked for offloading but the VoLTE UEs will be offloaded only when voice bearers are deactivated.



Configuring Load Balancing and Rebalancing

Configuring Load Balancing

Set the relative capacity of an MME service to enable load balancing across a group of MME services within an MME pool.

Use the following example to set the relative capacity of this MME service. The higher the value, the more likely the corresponding MME is to be selected.

```
config
  context context_name
    mme-service mme_svc -noconfirm
      relative-capacity rel_cap_value
      exit
```

Notes:

- **relative-capacity** *rel_cap_value* -- This command specifies a weight factor such that the probability of the eNodeB selecting this MME is proportional to this value in relation to other MMEs in a pool. *rel_cap_value* define the relative capacity by entering an integer from 0 to 255. The default relative capacity for an MME service is 255.
- The weight factor of the MME is sent from the MME to the eNodeB via S1-AP messages using the Relative MME Capacity S1AP IE in the S1AP S1 Setup Response. If the relative MME capacity is

changed after the S1 interface is already initialized, then the MME Configuration Update message is used to update this information to the eNodeB.

Verifying Load Balancing

Enter the **show mme-service all** causes the MME to generate a display similar to the following to indicate the configured relative capacity:

```
show mme-service all
Relative Capacity:      50
```

Performing Load Rebalancing (UE Offloading)

Start Offloading

The following example command rebalances (offloads) 30 percent of all UEs from the specified MME service (to other MME services in the MME pool) over the course of 10 minutes.

mme offload mme-service mme_svc time-duration *10* **offload-percentage** *30* **-noconfirm**

This command can also be entered with the **disable-implicit-detach** option. By default, if the UE context is not transferred to another MME within 5 minutes, the UE will be implicitly detached. This option disables this implicit detach timer.

mme offload mme-service *mme_svc* **time-duration** *10* **offload-percentage** *30* **disable-implicit-detach -noconfirm**

Stop Offloading

To stop the offloading process, issue the command with the **stop** keyword option.

mme offload mme-service *mme_svc* **stop -noconfirm**

Verifying Load Rebalancing (UE Offloading)

The following command shows the offload configuration as well as the status of the rebalancing.

show mme-service name *svc_name* **offload statistics**

```
show mme-service name mme1 offload statistics
Current Offload Status: In Progress
Implicit Detach Status: Enabled
Time Duration Requested: 600 secs
Percentage of Subscribers Requested: 30
Total Number of Subscribers: 0
Total Number of Subscribers to be Offloaded: 0
Total Number of Subscribers Offloaded: 0
Total Number of Subscribers Received Context Transfer: 0
Remaining Time:  0 secs
```

Where the Current Offload Status field will report one of the following:

- **Not Started** No UEs marked for offloading and no UEs currently being offloaded.
- **In Progress** MME is currently offloading marked UEs.
- **Completed** Offload procedure is completed or has been terminated by operator using **stop** keyword.

These counters are reset each time an offload procedure is initiated, or when the following command is entered:

clear mme-service statistics offload

Configuring VoLTE Offloading

The following configuration command is used to configure preservation of VoLTE subscribers from offloading during active calls (QCI=1); the offload command is enhanced with the key word**preserve-volte-subscribers** :

mme offload mme-service *mme_svc_name* **{ time-duration** *minutes* **offload-percentage** *percent* **[disable-implicit-detach | preserve-volte-subscribers]] | stop } [- noconfirm]**

By default, the subscribers with voice bearer with QCI = 1 will not be preserved during MME offloading. Configuring the keyword **preserve-volte-subscribers** enables preservation of subscribers with voice bearer.

The following example command re-balances(offloads) 30 percent of Non-VoLTE subscribers from the specified mme-service (to other mme-services in the MME pool) over the course of 30 minutes with VoLTE preservation.

mme offload mme-service *mmesvc* **time-duration** *30* **offload-percentage** *30* **preserve-volte-subscribers**

Verifying VoLTE Offloading

The following show command display is used to verify if VoLTE preservation is enabled and the number of VoLTE subscribers preserved during offloading:

show mme-service name *svc_name* **offload statistics**

```
Current Offload Status : Completed
Implicit Detach Status : Disabled
Preserve VoLTE subscribers Status : Enabled
Time Duration Requested : 60 secs
Percentage of Subscribers Requested : 1
Total Number of Subscribers : 0
Total Number of Subscribers Marked for Offloading: 1
Total Number of Subscribers Offloaded : 0
Total Cumulative Number of Subscribers Offloaded: 2
Total Number of VoLTE Subscribers Preserved : 0
Total Cumulative Number of VoLTE Subscribers Preserved:7
Total Number of Subscribers Received Context Transfer: 0
Remaining Time : 0 secs
```

Monitoring and Troubleshooting

The following sections describe commands available to monitor and troubleshoot this feature on the MME.

Show Command(s) and/or Outputs

This section provides information regarding show commands and their outputs in support of load rebalancing (UE offload).

The following show command displays current statistics for the Load Rebalancing feature.

show mme-service name *mme_svc* **offload statistics**

Table 1: show mme-service name <mme_svc_name> offload statistics

Field	Description
Current Offload Status	Current offload status of the specified mme-service. Possible values are Not Started, In Progress and Completed.
Implicit Detach Status	The Implicit Detach Status specified in the **mme offload** command. When enabled, if the UE context is not transferred to another MME within 5 minutes then it will be implicitly detached.
Preserve VoLTE subscribers Status	Is displayed as "Enabled" when the keyword **preserve-volte-subscribers** is configured in the mme offload command. The status is displayed as "Disabled", when VoLTE preservation is not configured. By default VoLTE preservation is disabled.
Time Duration Requested	The time-duration value specified in the **mme offload** command (in seconds). This is the maximum allowed time for the offload procedure to complete.
Percentage of Subscribers Requested	The offload-percentage specified in the **mme offload** command (specified as a percentage of all UEs on this mme-service).
Total Number of Subscribers	The total number of UEs on the specified mme-service.
Total Number of Subscribers Marked for Offloading	Displays the total number of subscribers marked for offloading during the current MME offload.
Total Number of Subscribers to be Offloaded	Total number of UEs on the specified mme-service selected for offloading.
Total Number of Subscribers Offloaded	The total number of UEs which have been successfully offloaded from this mme-service (UE offloading State/Event = Done).
Total Cumulative Number of Subscribers Offloaded	Displays the cumulative count of subscribers offloaded.
Total Number of VoLTE Subscribers Preserved	Displays the number of preserved VoLTE subscribers during and after MME offload.
Total Cumulative Number of VoLTE Subscribers Preserved	Displays the total numbers of subscribers preserved before starting the offload timer when the **mme offload** command is executed.

Field	Description
Total Number of Subscribers Received Context Transfer	Total number of UEs which has been successfully context transferred to another MME.
Remaining Time	The number of seconds remaining to complete the offload procedure.

The following command also provides information relating to load balancing:

show mme-service session full all

Only the output field which relates to load rebalancing is shown.

Table 2: show mme-service session full all

Field	Description
UE Offloading	Displays the UE offload state. Possible values are None, Marked, In-Progress and Done.