White Paper

March 2026

Protecting Networks in Large-Scale IoT Deployments

A GSMA IoT Community Publication

GSMA™

About the GSMA

The GSMA is a global organisation unifying the mobile ecosystem to unlock the full power of connectivity so that people, industry and society thrive.

Led by our members, we represent the interests of over 1,100 operators and businesses in the broader ecosystem. The GSMA also unites the industry at world-leading events, such as MWC (in Barcelona, Kigali, Las Vegas and Shanghai) and the M360 Series.

Unlock the benefits of GSMA membership

As a member of the GSMA, you join a vibrant community of industry leaders and visionaries – helping to shape the future of mobile technology and its transformative impact on societies worldwide.

Our unique position at the heart of the mobile industry means you get exclusive access to our technical experts, data and analysis – as well as unrivalled opportunities for networking, innovation support and skills acceleration.

For more information, please visit:
http://www.gsma.com/membership/

Security Classification: Non-confidential

Access to and distribution of this document is restricted to the persons permitted by the security classification. This document is confidential to the Association and is subject to copyright protection. This document is to be used only for the purposes for which it has been supplied and information contained in it must not be disclosed or in any other way made available, in whole or in part, to persons other than those permitted under the security classification without the prior written approval of the Association.

Antitrust Notice

The information contained herein is in full compliance with the GSM Association's antitrust compliance policy.

Copyright Notice

Disclaimer

The GSM Association ("Association") makes no representation, warranty or undertaking (express or implied) with respect to and does not accept any responsibility for, and hereby disclaims liability for the accuracy or completeness or timeliness of the information contained in this document. The information contained in this document may be subject to change without prior notice.

Contents

1. Introduction

1.1 Purpose of this document

With the rapid rise in Mobile IoT traffic, operators are getting more and more often confronted with adverse network effects caused by rogue devices or devices not adhering to the GSMA-recommended IoT Device Connection Efficiency Guidelines[1].

In many cases, those adverse effects are caused by roaming users, necessitating a coordinated approach between inbound and outbound roaming partners to:

- Implement measures to mitigate adverse effects on both the home and visited networks (signalling storm, cell congestion),
- Prevent wider service degradation or disruption, in particular for uninvolved customers
- Identify long-term actions to further increase the IoT network resilience to harmful behaviour in the future

In the first section of this whitepaper, the different Harm-to-Network scenarios are described and the root cause for their occurrence explained.

In the second section, recommended actions are provided for preventive, immediate as well as longer-term measures to be initiated in order to counteract harmful device behaviours.

This white paper embraces the complete chain of the IoT ecosystem and eventually provides recommendations focused on the following ecosystem players:

- Mobile Operators
- Module and Chipset Manufacturers
- IoT Device Application developers
- eUICC/eSIM/SIM Application Providers

2. Problem Statement

When developing IoT applications, device application developers must comply with various standards and guidelines to ensure their applications are working efficiently and optimally in both their own IoT communication providers' network and those of their providers' roaming partners.

This includes the TS.34 "IoT Device Connection Efficiency Guidelines"[1] which outlines rules and procedures to handle abnormal situations and prevent network overload or dysfunction.

For communication providers, guidelines like IR.73 "Steering of Roaming Implementation Guidelines"[2] or IR.88 "EPS Roaming Guidelines"[5] offer recommendations on steering mechanisms and associated error codes to effectively manage customer roaming traffic.

Eventually, 3GPP TS 29.272 "Evolved Packet System (EPS); Mobility Management Entity (MME) and Serving GPRS Support Node (SGSN) related interfaces based on Diameter protocol"[3] provides a mapping table between 3GPP S6a and NAS cause code values for all possible connectivity request reject cases.

Unfortunately, those standards and guidelines are not always strictly followed by IoT device application developers or IoT communication providers. Also, guidelines sometimes leave room for interpretation as to which specific action shall be taken when handling a particular scenario.

Eventually, this partial adherence to the IoT Device Connection Efficiency Guidelines combined with individual guideline interpretations creates the conditions for the inconsistent IoT Device behaviour and user experience seen across different networks and ultimately the harmful IoT Device behaviours that will now detail in the next section of this document.

Definitions

TERMS	DESCRIPTION
IoT	Internet of Things, a generic term for the network of physical objects that contain embedded technology to communicate and sense or interact with their internal states or the external environment. IoT offers functions and services which go beyond the pure M2M scope. MIoT is a subset of the far bigger IoT concept, for example a bunch of sensors connected together via Wi-Fi or Bluetooth are a part of IoT but not MIoT.
M2M	Machine-to-Machine, a general term referring to any network technology allowing devices to communicate with each other. For example, two industrial robots connected to each other via Ethernet in a factory is a part of M2M but not MIoT.
MIoT	Mobile Internet of Things, a GSMA term which refers to the 3GPP standardised LPWA technologies using the licenced band (aka LTE-M, NB-IoT and EC-GSM-IoT). From 3GPP Release 13 and the following Releases, the Category of UEs that support power consumption optimisations, extended coverage and lower complexity are part of MIoT (CAT M1, CAT NB1 from Release 13 and CAT M2, CAT NB2 from Release 14). As this particular term is widely used throughout GSMA, it is utilised also in this document.
LTE-M	LTE-M is the simplified industry term for the LTE-MTC low power wide area (LPWA) technology standard published by 3GPP in the Release 13 specification. It specifically refers to LTE Cat M, suitable for the IoT. LTE-M is a low power wide area technology which supports IoT through lower device complexity and provides extended coverage, while allowing the reuse of the LTE installed base.
CAT-M NTN	Generic category for CAT-M NTN devices supporting Release-17 or later, as specified in 3GPP TS 36.306[14].
CAT-NB NTN	Generic category for NB-IOT NTN devices supporting Release-17 or later, as specified in 3GPP TS 36.306[14].

Abbreviations

TERMS	DESCRIPTION
3GPP	3rd Generation Partnership Project
ACB	Access Class Barring
CE	Coverage Enhancement
EAB	Extended Access Barring
EMM	EPS Mobility Management
EPS	Evolved Packet System
EF_IMSI	Elementary File - International Mobile Subscriber Identity
eSIM	Embedded SIM
GPRS	General Packet Radio Service
HPLMN	Home Public Land Mobile Network
HSS	Home Subscriber Server
IMSI	International Mobile Subscriber Identity
IoT	Internet of Things
KPI	Key performance Indicator
LPWA	Low Power Wide Area
LTE-M	Long-Term Evolution Machine Type Communications
MNO	Mobile Network Operator
MME	Mobility Management Entity
NAS	Non-Access Stratum
NB-IoT	Narrowband IoT
ODB	Operator Determined Barring
HPLMN	Home Public Land Mobile Network
IPAe / LPAe	Embedded IoT Profile Assistant / Embedded Local Profile Assistant
PDP	Packet data Protocol
PLMN	Public Land Mobile Network
PSM	Power Saving Mode
RAT	Radio Access Technology

TERMS	DESCRIPTION
RPM	Radio Policy Manager
S6a	Diameter interface between MME and HSS for subscriber data & mobility services management
SIM	Subscriber identity Module
SGSN	Serving GPRS Support Node
OTA	Over-The-Air
UICC	Universal Integrated Circuit Card
VPLMN	Visited Public Land Mobile Network

References

REF	DOC NUMBER	DOC NUMBER
[1]	GSMA TS.34	"IoT Device Connection Efficiency Guidelines" https://www.gsma.com/iot/gsma-iot-device-connection-efficiency-guidelines/
[2]	GSMA IR.73	"Steering of Roaming Implementation Guidelines" https://www.gsma.com/newsroom/gsma_resources/steering-of-roaming-implementation-guidelines-v10-0/
[3]	3GPP TS 29.272	"Evolved Packet System (EPS); Mobility Management Entity (MME) and Serving GPRS Support Node (SGSN) related interfaces based on Diameter protocol" http://www.3gpp.org/DynaReport/29272.htm
[4]	3GPP TS 36.321	"Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification" http://www.3gpp.org/DynaReport/36321.htm
[5]	GSMA IR.88	"EPS Roaming Guidelines" https://www.gsma.com/newsroom/wp-content/uploads/IR.88-v25.0-17.pdf
[6]	3GPP TS 24.301	"Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS)" http://www.3gpp.org/DynaReport/24301.htm
[7]	3GPP TS 24.501	"Non-Access-Stratum (NAS) protocol for 5G System (5GS); Stage 3 http://www.3gpp.org/DynaReport/24501.htm

3. Harm-to-Network scenarios

3.1 Radio network congestion

3.1.1 Background: Large scale IoT deployments

Example scenario: Smart Metering

With the device densities accompanying large scale smart meter deployments, it's not uncommon for some operators to be seeing several thousand mobile IoT devices in a single radio sector. Whilst great care is often exercised in the development of battery powered smart meters used for water and gas to deliver a 10+ year service life, many operators are finding that the same cannot be said for the advanced metering infrastructure deployments used for smart electrical metering For these deployments, often using LTE-M, smart meters are typically configured to provide an "always on" connection with the smart meter closely monitoring the wireless connection for loss of connectivity.

It is during a loss of connectivity that the stark difference in design methodology between battery and mains powered smart meters is most noticeable. Expecting an "always on" connection, many electrical smart meters will aggressively seek to reestablish connectivity. In contrast, battery powered smart meters with energy constraints, generally use a much more conservative connectivity retry algorithm. In aggressively trying to reestablish connectivity, the mains powered meter can create signalling storms which may continue indefinitely if the MNO does not step in and take remedial measures.

3.1.2 Radio cell impact

In understanding the impact of aggressively trying to reestablish connectivity, we need to consider how a device attaches to the network.

Before attempting to connect to a network, a mobile IoT device will perform a cell search to select the best possible radio cell to connect to. As part of that radio cell search, mobile IoT devices read system information broadcast by each radio cell. This system information provides guidance for the mobile IoT device on how to connect to that radio cell. For example, system information assists the mobile IoT device to determine what power level it should initially transmit with, when and where it should send its initial radio transmission and identifiers to be used for that initial transmission.

When trying to connect to a mobile network, a mobile IoT device will go through what's called the random access procedure. It's through the random access procedure that the mobile IoT device is able to establish connectivity with the radio base station ahead of authenticating itself to the network and fully attaching.

In the case of a mains failure, upon restoration of mains power, it is possible that thousands of electrical smart meters will immediately try to reconnect to the network at the very same instant. It's with thousands of smart meters all trying to perform the random access procedure at the same time that problems can occur. Think of it as a bit like a freeway with thousands of vehicles all trying to enter the freeway at the same time. Whilst under normal traffic conditions the freeway may be able to handle that number of vehicles, the on ramp becomes the choke point.

It is not uncommon for smart meters to monitor the amount of time elapsed to reestablish connectivity. Under such high load conditions, delays are to be expected. However, after a period of time that the smart meter software developer considers sufficient, some smart meters may perform a hard reset on the radio module in the smart meter and the cycle starts all over again.

From a base station perspective, it must manage a finite amount of radio resources. The sending of the multiple messages used to complete the random access procedure, consumes radio resources. Resources that may otherwise be used for the sending and receiving of user data. Ordinarily, this is not a problem but under the high load condition of thousands of devices trying to reconnect, the base station faces a challenge. Should it prioritise resources so the devices that are already connected can send/receive data or should it prioritise the reconnection of new devices. There may be insufficient resources to do both. The handling of this is determined by algorithms in the base station with many base stations favouring the reconnection of devices.

If the base station prioritises the reconnection of new devices, until the large number of devices can be reconnected, those devices that have connected may be unable to send/receive data due to radio resources being consumed by the random access procedure. Remembering that some smart meters power cycle their radio modules if there has been no connectivity after a period of time and so starved of uplink/downlink data, do just that. They reset the modem and start all over again.

A contributing factor to this problem is an increase in radio noise at the base station. On first attempting the random access procedure, a mobile IoT device will determine an optimal output power to use based on the receive signal strength it measures, the transmit power output the base station has indicated it is using and included in the system information as well as the desired signal level the base station would like to receive, also identified in system information.

Due to the large number of devices trying to simultaneously attempt the random access procedure, it is possible the base station may not respond to a random access attempt. In this situation, the mobile IoT device may assume that the base station did not hear that attempt and as a consequence, if it's not already operating at full power, it will ramp up the power level it transmits with based on ramping parameters provided by the base station in system information.

Another contributing factor is the number of random access preambles available for the mobile IoT device to use as an initial identifier. For a mobile network, the maximum number of preambles available to choose from is theoretically 64 although in practice, it is much more common for networks to use 52 or 56. The number of preambles available is provided in system information and a mobile terminal will randomly select one out of the pool available.

For mobile IoT devices using LTE-M, the number of preambles available for use is significantly less. That is, a MNO will configure their available preamble pool such that a small proportion of preambles are available for LTE-M devices in coverage enhancement level 0 (CE0) and a separate small number of preambles are available for LTE-M devices using CE1. It is not uncommon for a radio base station delivering a LTE-M service to be configured with a total preamble pool size of 52, out of which 4 preambles (numbers 0-3) are allocated for LTE-M devices using CE0, 3 preambles (numbers 4-6) allocated for LTE-M using CE1 and the remainder of the pool allocated for 4G LTE. With 1000s of LTE-M based smart meters randomly choosing from such a small number of preambles, contention resolution becomes challenging.

The root cause of these issues are the ill-conceived connectivity reestablishment algorithms in the smart meter and this should be the first place to start.

Application developers therefore need to do a much better job of implementing connectivity management software. Whilst simplistic measures make appear to work well in a lab or in the field for small numbers of devices, when deployed at a scale, the only difference between an ill-conceived smart meter application and a denial-of-service attack, is malicious intent.

3.2 Core network performance degradation

3.2.1 Background: Harmful IoT Application behaviour

Example scenario: Smart Metering

With the always increasing relative proportion of IoT devices in their network, IoT communication providers are unfortunately more and more often confronted with situations where the misbehaviour of IoT applications causes a significant impact on their network.

Misbehaviour in error scenarios

Often as a result of a poor design, those IoT device applications fail to properly deal with a particular session or mobility error situation, leading to a dead lock situation where the device application is in vain reattempting the same action again and again. In a worst-case scenario, such a situation may cause a signalling storm, eventually impacting both the Roaming Partner's local network, the IoT communication provider core network and the end-customer device.

Example of such situations, whereby IoT devices remain in a powered-on state and endlessly try to establish a connection, includes for instance:

— Attach rejects sent after a temporary or long-term SIM barring by the IoT Customer.
— Attach rejects sent after a suspension or termination of the connectivity service by the IoT Service Provider.

To make the situation even worse, and as explained in the next sections in this chapter, the same error situation can lead to different reject causes and error codes, depending on the network the IoT device is roaming in and which HPLMN the SIM-card is provisioned in.

Misbehaviour in standard scenarios

Even in scenarios where no network failures or rejects occur, IoT device applications sometimes also do not comply with basic rules from the IoT Device Connection Efficiency Guidelines[1] relating to standard communication behaviours. This may for instance include:

— Applications systematically detaching after every successful data transmission and re-attaching only a few seconds or minutes later instead of remaining in idle mode or using the PSM feature.
— Applications initiating new Service Request Messages every few minutes instead of optimising their data transfer, e.g. buffering data and sending them within one single RRC connection at longer intervals.

Those behaviours generate an excessive amount of RRC and NAS messages towards both the Radio and Access network, eventually impacting the overall network performance and capacity of the VPLMN network.

3.2.2 Network Reject implementation by MNOs

As described in the problem statement, although recommendations exist on how IoT communication providers should react to specific error situations, these recommendations can sometimes be interpreted differently by roaming partners.

For example, one HPLMN might use the DIAMETER_ERROR_RAT_NOT_ALLOWED (5421) code to prevent NB-IoT roaming in a VPLMN network, while another might use DIAMETER_ERROR_UNKNOWN_EPS_SUBSCRIPTION (5420) with an Error Diagnostic of GPRS_DATA_SUBSCRIBED,. This could lead to different NAS Cause Codes by the VPLMN and varying reactions from the IoT Module or Application (should the NAS Cause Code be known to the latter).

Table 1 illustrates Diameter error codes that an HPLMN may choose when facing roaming error scenarios that require denying an attach request:

USE CASE SCENARIO	POSSIBLE DIAMETER REJECT CODE BY HPLMN
The SIM-card is not yet provisioned in the HPLMN HSS	DIAMETER_ERROR_USER_UNKNOWN (5001)
The SIM-card has been barred (ODB Barring) by the HPLMN (e.g. "Barring of Roaming")	DIAMETER_ERROR_ROAMING_NOT_ALLOWED (5004) , without Error Diagnostic DIAMETER_ERROR_ROAMING_NOT_ALLOWED (5004), with Error Diagnostic of ODB_HPLMN_APN or ODB_VPLMN_APN DIAMETER_ERROR_ROAMING_NOT_ALLOWED (5004), with Error Diagnostic of ODB_ALL_APN
Roaming for the SIM-card is not allowed by the HPLMN network for that specific RAT-Type only	DIAMETER_ERROR_UNKNOWN_EPS_SUBSCRIPTION (5420) without Error Diagnostic, or DIAMETER_ERROR_UNKNOWN_EPS_SUBSCRIPTION (5420) with Error Diagnostic of GPRS_DATA_SUBSCRIBED DIAMETER_ERROR_RAT_NOT_ALLOWED (5421)
Roaming for the SIM-card is not allowed by the HPLMN network for all RAT-Types (no Roaming agreement in place)	DIAMETER_ERROR_ROAMING_NOT_ALLOWED (5004) without Error Diagnostic
SIM-card is steered away by the SoR-implementation of the HPLMN	DIAMETER_ERROR_RAT_NOT_ALLOWED (5421), or DIAMETER_ERROR_ROAMING_NOT_ALLOWED (5004), without Error Diagnostic DIAMETER_UNABLE TO COMPLY (5012) DIAMETER_ERROR_UNKNOWN_EPS_SUBSCRIPTION (5420) without Error Diagnostic DIAMETER_ERROR_UNKNOWN_EPS_SUBSCRIPTION (5420) with Error Diagnostic of GPRS_DATA_SUBSCRIBED
Other scenarios (non-exhaustive list)	DIAMETER_UNABLE_TO_COMPLY (5012), DIAMETER_INVALID_AVP_VALUE (5004) DIAMETER_AVP_UNSUPPORTED (5001) DIAMETER_MISSING_AVP (5005) DIAMETER_RESOURCES_EXCEEDED (5006) DIAMETER_AVP_OCCURS_TOO_MANY_TIMES (5009) DIAMETER_AUTHENTICATION_DATA_UNAVAILABLE

Table 1, Use Case scenario to Diameter reject codes mapping in roami

3.2.3 Mapping of Network Reject between Roaming partners

Upon receiving Diameter Reject codes from the HPLMN, VPLMN networks in turn may apply different NAS Cause codes for the same Diameter Reject code. Which mappings of Diameter Reject codes to NAS Cause codes are allowed is defined in the Annex A, Table A.1 of the 3GPP specification TS29.272[3].

For example, upon receiving DIAMETER_ERROR_RAT_NOT_ALLOWED (5421), one mobile operator might map it to NAS cause code #15 "No suitable cells in tracking area", while another operator might use NAS cause code #13 "Roaming not allowed in this tracking area". As explained previously, these different cause codes could eventually lead to different reactions by the device module and IoT device application, potentially resulting in different reactions from the IoT device.

Further recommendations are provided by the GSMA in their Roaming Guidelines (see IR.73 "Steering of Roaming Guidelines"[2] and IR.88 "EPS Roaming Guidelines") that are further limiting the eligible mapping options in roaming, yet it is unclear whether those recommendations are strictly enforced by all mobile carriers today.

Depending on the actual visited and home PLMN network being used, IoT Applications may therefore be confronted for the very same scenario with different NAS cause codes, leading to an unexpected and potentially harmful device behaviour towards the network.

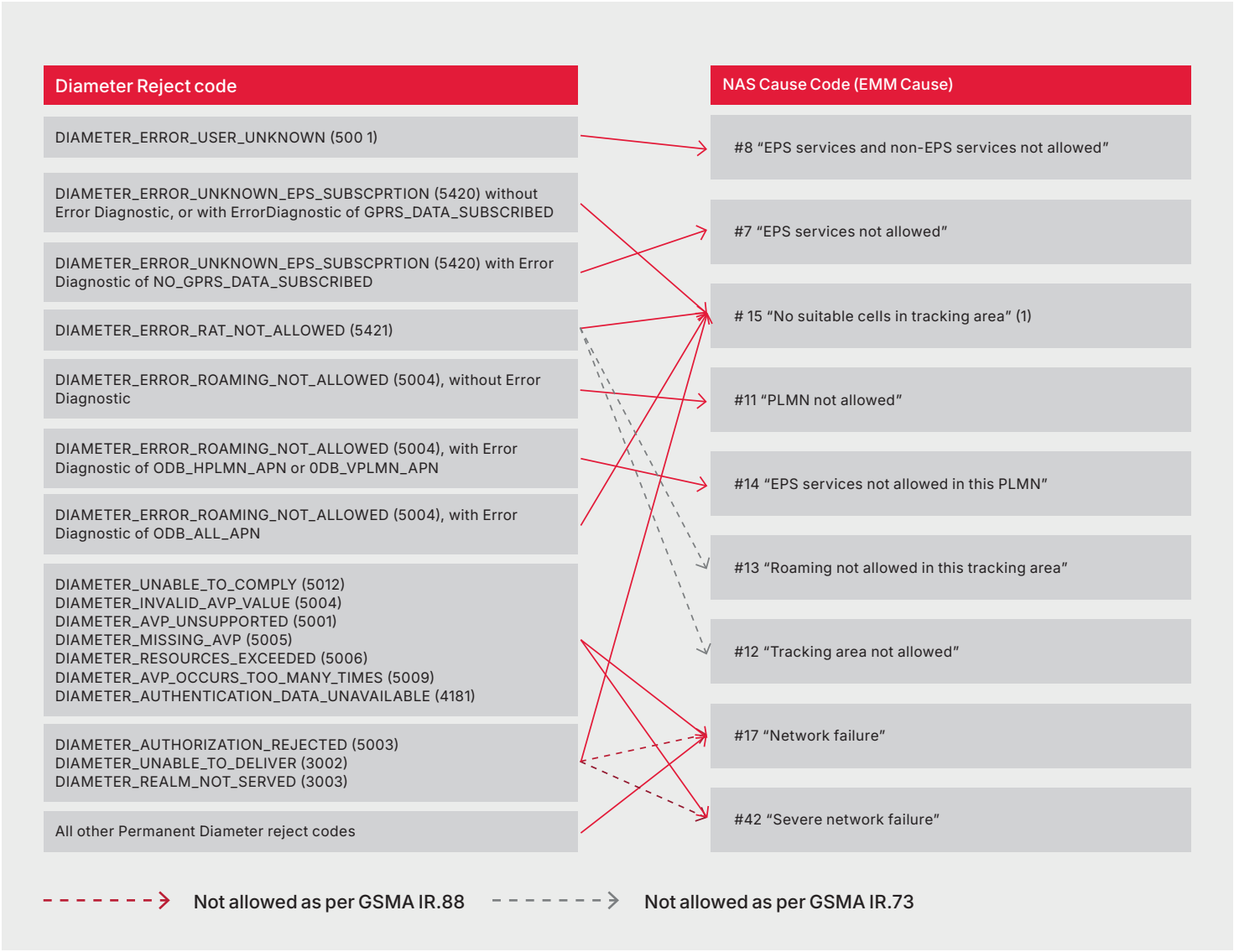

Figure 1, Diameter Reject codes and associated NAS Cause Code mapping

3.2.4 Network reject handling by the UE

Whenever an attempt by the UE to attach to a network or set up a PDP context is denied, the NAS cause code is the error code delivered to the device's radio module by the visited network to provide further context about the reason for the rejection.

How a device module shall react upon receiving such a code is specified in 3GPP document TS24.301[6] and a high-level description of this behavior for the codes identified in the previous section is provided in the table below.

Note: All cases pertaining to each cause code cannot be reflected in this high-level view that only aims to highlight the most common behavior of the module in a standard situation.

NAS Cause Code (EMM Cause)		High-level module behavior description
#8 "EPS services and non-EPS services not allowed"	→	Module shall only attempt a new attach after a reset or wait for T3245 to expire if available in EF_ NASCONFIG.
#7 "EPS services not allowed"	→	Module must consider the SIM as invalid for EPS service until it is reset or after expiration of the T3245 timer if available in EF _NASCONFIG.
#15 "No suitable cells in tracking area" (1)	→	Module may immediately search for another eel l or tracking/location area and may only reattach to the same tracking area after a module reset or the erasure of the "forbidden tracking areas for roaming" list (every 12 to 24 hours).
#11 "PLMN not allowed"	→	Module may immediately perform a new PLMN selection and may only reattach to the same PLMN after a module reset or the expiration of the T3245 timer if available in EF _CON FIG.
#14 "EPS services not allowed in this PLMN"	→	Module must ignore the PLMN for EPS services and immediately perform a new PLMN selection. It may only reattach to the same PLMN for EPS services after a module reset or the expiration of the T3245 timer if available in EF _NASCONFIG.
#17 "Network failure"	→	Module may try again after expiration of T3411to attach to the same PLMN.A fter 5 unsuccessful attempts, the module shall either wait for expiration ofT3402 and attempt to attach to the same PLMN, or optionally perform a new PLMN selection.
#42 "Severe network failure"	→	Module shall immediately ignore the PLMN+RAT combination for a duration of 2 x T and perform a new PLMN selection.
#19"ESM failure"	→	The Module shall either wait for expiration of T3402 and attem ptto attach to the same PLMN, or optionally perform a new PLMN selection.

T3245 (Timer affer permanent affach reject): between 24 and 48 hours
T3410 (Timeout of Attach Request): 15s
T3411 (Timer after attach failure due to lower fayer failure or '"other EMM causes" :10s
T3402 (Timer after 5 attach failures): 12min
Timer T (Timer between 2 periodic network scans): Operator-specific

Figure 2, High-level module reaction to common NAS Cause Codes
(simplified view only, please refer to [6] for a comprehensive description of all sub-cases)

3.2.5 Reject code handling by Radio Policy Manager (RPM)

To address such scenarios and help protect the mobile network from signalling overload, a set of non-standardised features called the Radio Policy Manager (RPM) have been specified by the GSMA in TS.34 – IoT Device Connection Efficiency Guidelines[1].

Several NAS cause codes are handled by the Radio Policy Manager deployed in the most popular modules and chipsets.

In the case of EMM rejects, this include for instance:

- EMM Permanent rejects (as per TS.34_8.2.2_REQ_006)
  - EMM Reject Cause #3 (Illegal UE)
  - EMM Reject Cause #6 (Illegal ME)
  - EMM Reject Cause #8 (EPS services and non-EPS services not allowed)
  - EMM Reject Cause #7 (EPS services not allowed)
- EMM Temporary rejects (as per TS.34_8.2.2_REQ_011)
  - Reject Cause #17 (Network Failure)
  - Reject Cause #22 (Congestion)
  - Reject Cause #40 (No EPC bearer context activated)
  - Reject Cause #26 (insufficient resources)
  - Reject Cause #30 (request rejected by Serving GW or PDN GW Congestion)
  - Reject Cause #31 (request rejected)

When triggered by those specific codes, the RPM Manager implements specific backoff rules and/or timers that ensures that a device can't trigger too many attaches or PDN connection requests by the module within a specific time window.

Unfortunately, the scope of application of RPM does not cover all reject causes, so that in many situations, RPM eventually does not come into effect.

And as shown in Figure 2, the majority of the NAS cause codes triggered as a result of a standard reject scenario aren't covered by the current RPM scope, in particular:

- #11 "PLMN not allowed"
- #12 "Tracking area not allowed"
- #13 "Roaming not allowed in this tracking area"
- #14 "EPS services not allowed in this PLMN"
- #15 "No suitable cells in tracking area"
- #19 "ESM failure"

As an example, in the case of an EMM Reject Cause #15 (No suitable cells in tracking area) or #11 (PLMN not allowed), no action will ever be triggered by RPM, leaving the IoT application free to reset the communication module and reattempt to attach as often as it wants, leading to a potential signalling storm.

4. Recommended actions

4.1 Radio Network-related actions

4.1.1 Preamble pool size

A previously mentioned in paragraph “Radio cell impact” page 4, the number of random access preambles available for LTE-M devices to use is typically quite small.MNOs need to understand their existing pool sizes and consider whether it is possible to increase the pool size for LTE-M. Note that doing so may take preambles away from 4G LTE customers to use and therefore may not be possible.

Recommendation

— Consider increasing the preamble pool size for LTE-M

4.1.2 Number of random access attempts

Another parameter broadcast in system information is the number of random access retry attempts. Section 5.1.4 of 3GPP TS36.321[4] uses this value to identify random access failure. MNOs need to understand the number of random access retries configured and consider whether it is possible to increase that for LTE-M.

Recommendation

— Consider increasing the number of random access attempts for LTE-M.

4.1.3 Random access backoff Indications

A radio base station upon detecting high random access traffic load can transmit a rejection message with a backoff indicator set. The maximum backoff interval indicator that can be sent for LTE-M amounts to 960 ms. MNOs should investigate whether the base stations they have deployed support this capability and if so, consider enabling it for LTE-M. It’s not a silver bullet but it may improve the situation.

Recommendation

— Consider enabling random access backoff indication for LTE-M.

4.1.4 RF ramping parameters

A side issue with random access congestion is an increase in base station noise levels. That is, devices may start to ramp up output power levels upon unsuccessful random access attempts in the mistaken belief that the base station could not hear the device. MNOs should investigate their LTE-M power ramping parameters and where a less aggressive ramping could improve the situation. Noting that in genuine cases of low signal level, a less aggressive ramping level may create other issues.

Recommendation

— Consider a less aggressive ramping level for LTE-M.

4.1.5 Access Class Barring

Access Class Barring (ACB) is a network feature that have been available since 3GPP Release R9 and as a consequence, it is supported by many radio modules in market.

When ACB is active in a base station and a random access overload condition is detected, the base station will broadcast a parameter in system information which is used an indicator of the probability a device should continue with the random access procedure. That is, if for example a radio base station is sending a value of 80 which represents an 80% probability that the device can proceed with the random access procedure, the radio module in the device will also generate a random number. If the random number generated by the module is less than that broadcast by the radio base station, the random access procedure can proceed. If the value generated by the radio module is greater than that broadcast by the radio base station, the random access attempt should not proceed until a timer interval in seconds determined by the equation (0.7 + 0.6 x *rand*) *x ac-BarringTime*. Where *“rand”* is the random probability calculated by the radio module and *“ac-BarringTime”* is broadcast by the MNO in system information and can take values such as 4, 8, 16, 32, 64, 128, 256 or 512 seconds.

It is possible for the probability broadcast by the base station to be varied inversely with the random access traffic load. That is, if the load does not abate, the probability can be automatically reduced and if the load decreases, the probability can be automatically increased.

As stated, one of the key benefits of ACB is that today, most radio modules support this capability. A downside of ACB is that for some base station types, enabling ACB for LTE-M may also mean that ACB is enabled for LTE as well. This may create a situation where high random access traffic load on LTE-M by triggering ACB, may mean that LTE connections are also impacted where they otherwise have been unimpacted if ACB was not enabled.

The use of ACB is a bit like shutting the gate after the horse has bolted but if carefully managed, it may improve the situation by dampening random access demand.

Recommendation

— Consider enabling ACB

4.1.6 Extended Access Barring

As the name suggests, Extended Access Barring (EAB) extends the capability of ACB. Introduced in 3GPP Release R14, there is a potential some radio modules in the field do not support this feature.

EAB operates by classifying devices into 1 of 10 groups. The grouping determined by information held on the UICC. Under random access overload conditions, the base station broadcasts a ten-bit parameter in system information which identifies which of the ten groups is allowed to continue with the random access procedure. That is, each bit represents a group and if that bit is set, the group is allowed to proceed with the random access procedure. Initially a MNO may start with a single group only, adding in additional groups as random access traffic demand abates.

Unlike ACB, enabling EAB for LTE-M does not necessarily mean that LTE traffic is also impacted.
A challenge some MNOs may face with EAB is group identifiers may not currently exist on in service UICC. This may mean a campaign of over the air UICC updates is required and some MNOs may be reticent to go down that path for fear of bricking a UICC.

Recommendation

— Consider enabling EAB taking into account support levels for the feature on their UICC and in field devices.

4.1.7 Multiple LTE-M Narrowbands

Today, most networks where LTE-M is deployed, use a single narrowband. The term narrowband is confusing and has nothing to do with NB-IoT. It refers to the six resource blocks used for LTE-M.

Where a single LTE-M narrowband is in use, both the random access signalling and user data are carried over the same LTE-M narrowband. This means that under overload conditions, the base station is forced to decide which to favour over the other.

By introducing an additional LTE-M narrowband, it is possible to confine random access signalling to only one of the two LTE-M narrowbands. The other narrowband is then available for carriage of user traffic, easing the priority call the base station needs to make.

Provisioning an additional LTE-M narrowband may raise some traffic concerns for regular LTE traffic sharing the same spectrum. This concern may be amplified where channel bandwidths are small. For example, a 10 MHz channel bandwidth. Depending on the base station type, it is possible to adjust whether LTE-M or LTE has priority over the contended resource blocks and consequently, the impact of additional LTE-M narrowbands may be minimised.

Recommendation

— Consider the impact and benefit of enabling additional LTE-M narrowbands

4.1.8 Barring recalcitrant devices

Since a device needs to perform random access to be able to authenticate itself to the network, It won't be able to know it is barred from connecting to that network until authentication has been attempted.

Therefore, whilst access barring misbehaving services may seem like the answer, it will most likely not solve the problem of random access congestion when this congestion is caused by rogue devices.

Recommendation

— Implement Access Barring of recalcitrant devices only as an intermediary solution whilst elaborating a strategy to counteract the device's rogue behaviour

4.2 Core Network-related actions

4.2.1 Mapping of use case scenarios to reject causes (Diameter code to NAS cause code)

For all actors in the IoT communication chain to be able to efficiently deal with each error case scenarios, it is critical for them to understand which error cause is to be expected by the device module and the IoT Device Application.

Mobile operators should aim to harmonise both the Diameter Reject codes and their associated NAS Cause code mapping so as to ensure a consistent user experience across all networks. This can be to a large extent achieved by fully adhering to the 3GPP and GSMA specifications, and in particular:

- 3GPP TS29.272[3]: Mobility Management Entity (MME) and Serving GPRS Support Node (SGSN) related interfaces based on Diameter protocol, Annex A
- GSMA IR.88[5]: LTE and EPC Roaming Guidelines
- GSMA IR.73[2]: Steering of Roaming Implementation Guidelines

This applies to the Steering of Roaming scenarios, for which various Diameter Reject codes may be selected by the HPLMN depending on the intended impact on the roaming device behaviour (see Table 2 “Diameter reject codes for roaming”)

In this Steering-of-Roaming scenario in particular, a shared understanding across all Roaming partners of the end-to-end reject process and the associated device behaviour could be a first step towards establishing a joint strategy for dealing with harmful devices in roaming scenarios.

Recommendations

- Ensure full compliance of the Diameter Reject to NAS Cause codes mapping to the GSMA Roaming guidelines ([2], [5])
- Be considerate of the impact of your Steering-of-Roaming choices on the IoT Device behaviour

4.3 Application-related actions

4.3.1 Randomised reconnection

Immediately trying to reconnect to the network on loss of connectivity, may not be the best response. Developers need to consider the implications of doing so when devices are deployed at scale. A far better outcome will be achieved by randomising when a device attempts to reestablish connectivity. The time window over which a device randomly attempts to reconnect needs to be increased with the number of devices per cell. The granularity of the randomised times within that interval needs to be as small as possible. For example, 4000 devices randomised over 16 one-minute intervals still means that at exactly every minute, 250 devices will be trying to reconnect. Whereas 4000 devices randomised on a one second basis means that every second approximately 4 devices will be trying to reconnect.

One possible method of determining the time window over which to randomise, is for devices to report back to an IoT device management platform, the identity of the cell. The IoT device management platform will then be able to build up a database of cell identities and the number of devices in each radio cell. Over time as the number of devices increases, the IoT device management platform should send out a revised randomisation time window parameter to each smart meter that increases as the number of devices in that same cell increases. The specific value of the randomisation window size sent out being a factor of number of devices and the randomisation granularity. For example, an organisation might set themselves a target of never exceeding 4 devices per second and therefore as the number of devices increases, the window size parameter is varied to maintain that metric.

Another challenge that may be faced is that despite randomising reconnection attempts, a device may be unsuccessful in reconnecting. This may be due to a range of factors, for example in a forest fire the radio tower may have burnt down. It makes little sense for a device to keep attempting to reconnect in such a situation. When an attempt at reconnecting has failed, a good practise is for the device to back off for exponentially increasing periods of time up to a maximum limit. For example, a device may have a failed reconnect attempt, instead of immediately retrying again, one possibility is to retry again using twice the randomisation window size. Then if a third retry is required, then the randomisation window size is increased by a factor of 4. This increase in randomisation window size might continue until the window size reaches 24 hours and from then on, the device tries once a day at a random time until it is able to reconnect.

Recommendation

- Randomise reconnection attempts over at least 1 second intervals distributed across as wide as possible reconnection time windows.
- Adapt the reconnection window size to reflect the number of devices in each cell and increase the reconnection windows size exponentially if attempts are unsuccessful.

4.3.2 Solicited and unsolicited response codes

Whenever possible, application software should read and interpret response codes sent back by the radio module. Failing to read and interpret response codes and purely relying on time and assumptions is not a sound practice.

Recommendation

- When supported by the IoT Module, IoT Device Application should read and interpret the solicited and unsolicited response codes from the radio module to derive the next best action in the event of a communication failure

4.3.3 Modem reset by loss of end-to-end connectivity

Over the years, the quality of radio modules has improved markedly. It should not be necessary to perform resets on radio modules to recover the radio module because the device has been unable to connect to a network. This is especially true when the reset mechanism of choice is the asynchronous removal of power to the radio module and UICC. For example, continuously powering down radio modem boards every 4 minutes due to an inability for the radio module to reconnect, is not a sound practice. Power cycling the radio module 360 times a day, for days on end, brings with it a potential to do long term damage to both the radio module and the UICC if power is removed when either entity is midway through a critical function.

On this matter, the GSMA TS.34 Guidelines[1] states in requirements TS.34_4.0_REQ_029 and TS.34_4.2_REQ_029 that "The IoT Device Application SHALL check that communication issues to the server are not caused by higher layer communications (like TCP/IP, UDP, ATM...) before starting to reset the communication module or re-establish the RRC Connection."

Typically, a stepped approach that may be considered by IoT Application Developers when attempting to recover from a connectivity loss could look as follows:

- Step 1. Re-establishment of higher layer connectivity, e.g. VPN tunnel, SSH session, etc.,
- Step 2. Re-establishment of the PDN connectivity or PDP context,
- Step 3. Re-attach (data) to the network,
- Step 4. Re-triggering of a plain network selection,
- Step 5. Complete reboot of the device

Recommendation

- Implement a top-down recovery approach across all communication layers in case of an IoT Device Application connectivity failure.
- Only perform a hard reset of radio modules/UICCs to reestablish connectivity in the last resort after all higher layer reconnection attempts have failed.

4.4 Chipset & Module-related actions

4.2.1 Handling of NAS cause codes not covered by the current TS.34 RPM scope

As of today, no mechanism exists that allows mobile operator to fully control how frequently an IoT Device Application may attempt to reattach after an Attach or a PDN Connection Request failure. And even in those scenarios where timers between two retries are specified (such as T3245 for the periodic erasure of the list of Forbidden Tracking Areas), a simple module reboot often suffices for IoT Device Applications to circumvent those protection mechanisms.

Considerations should therefore be made as to how those NAS Cause codes may be more effectively managed by IoT Communication Module, Chipset or SIM Providers, in order to prevent or respond to harmful behaviours by IoT Device Applications.

This may include the implementation of new or stricter requirements in the TS.34 guidelines to address the most frequent reject scenarios currently not covered in the Radio Policy Manager Requirements in its section 8., in particular for the following temporary rejects:

— #11 “PLMN not allowed”
— #12 “Tracking area not allowed”
— #13 “Roaming not allowed in this tracking area”
— #14 “EPS services not allowed in this PLMN”
— #15 “No suitable cells in tracking area”
— #19 “ESM failure”

Alternative investigative paths may include the development of a SIM application triggered by those reject codes that would temporarily disable the associated UICC or select an alternative HPLMN.

Eventually, all module and chipset vendors shall ensure full compliance with the 3GPP specifications, and in particular with the retry procedures specified in sections 5.5.1.2.5 and 5.5.1.2.6 of TS24.301[6] and sections 5.5.1.2.5 and 5.5.1.2.7 of TS 24.501[7].

Recommendations

— Implement the latest GSMA TS.34 recommendations including the newest RPM handling recommendations for permanent and temporary rejects
— Investigate possible enhancements to the RPM functionality to address harmful behaviours by IoT Device Applications upon receipt of frequent temporary reject scenarios.

4.2.2 Unsolicited response codes exposure towards Application Layer

One of the challenges facing application developers is the lack of visibility of what the underlying chipset/module is doing.

For example, in attempting the random access procedure, each time the UE sends a random access preamble, it increments it's "PREAMBLE_TRANSMISSION_COUNTER" by 1. This counter reflecting the number of unsuccessful random access attempts the UE has made. MNOs place a limit on the number of random access attempts by the UE before a problem should be recognised and this is shared with the UE in system information via the "preambleTransMax-CE" parameter.

The 3GPP Technical Specifications TS 36.321 Section 5.1.4 and TS 36.523-1 Section 7.1.2.3.2 each require UE to indicate a random access problem to upper protocol layers when the PREAMBLE_TRANSMISSION_COUNTER exceeds preambleTransMax-CE. However, few if any chipsets/modules provide any indication up to an application level by way of unsolicited result codes. Some chipsets/modules address this situation by aligning with 3GPP Technical Specification TS36.331 Section 5.3.11.3 radio link failure requirements but again, do not provide any indication to the application level by way of unsolicited result codes. In each instance, the application software gets no visibility of any problem and blindly continues to operate with blissful ignorance.

Chipset/module manufacturers should provide unsolicited response codes to advise higher layer software applications of the onset of random access congestion/failure. This could be done by enhancing the AT command layer to communicate more granular information on the actual error code, highlighting the chipset/module is unable to attach and detail the reason for not being able to attach.

Recommendation

— Implement unsolicited response codes to advise higher layer software applications of the onset of random access congestion/failure.

4.5 SIM-related recommendations

4.5.1 SIM OTA Management

As part of any SIM deployment, i.e. xUICC equipped cellular modem, remote lifecycle management of the subscription data in the SIM is typically handled by a SIM over-the-air (OTA) management system operated by the owner of the SIM card or eSIM profile, normally the MNO/MVNO.

This includes procedures for essential management operations like eSIM management, roaming steering, applet configuration, applet functionality and eventual service termination. These processes are particularly important in IoT environments, where devices may have embedded SIMs (eUICCs/iUICCs) and/or are often deployed in hard-to-reach locations and physical access may be impractical.

Since these procedures often depend on mobile terminated (MT) and mobile originated (MO) SMS, it is critical that the application software does not disable modem SMS features, even if the application itself does not utilize SMS. Besides MO/MT SMS, the SIM function commonly sets up a direct MO https session (Pull/Poll) to backend SIM OTA servers for some remote operations, e.g. applet and eSIM download and management, that must not be blocked by any application-level function.
This "Polling" process is also used if SMS is not supported by the network.

Recommendation

— IoT Application Developers to not disable SMS features in modem or bar SMS service in their Customer Portal, even if the application itself does not utilize SMS
— IoT Application Developers to not block https session triggered by SIM functions

4.5.2 SIM OTA Polling

Consider configuring the following SIM https polling mechanisms:

Recommendation

— Avoid excessive network traffic and power drain, especially in battery-constrained IoT devices by choosing a longer default polling period. If the application is able to trigger an SIM OTA poll – then the default period can be longer
— When a SIM OTA poll results in a data transfer, the delay to the next poll shall be short in order to complete the pending SIM OTA transactions.
— When a SIM OTA poll results in "nothing to do", the delay to the next poll shall be starting at the default delay and increased exponentially ("back-off")
— When a SIM OTA poll/pull fails, e.g. "backend not reachable", the number of retries shall be limited and use exponential delay increase in the retry timers ("back-off").
— For large scale deployments in the same power area, consider randomizing SIM ready access after power loss to avoid network signaling overload. To spread out network attach attempts over time, the SIM may start to send a "dummy-IMSI" to the modem and then - after a random power on timer in the SIM application – refresh the modem IMSI to the actual IMSI.
— The application software shall avoid using modem "power-cycle" when retrying network connect, since this will likely cause SIM delay timers to be reset inhibiting any "back-off" behavior, or interfere with power loss mitigations.
— The aforementioned polling configuration parameters shall be configurable OTA.

4.5.3 Prevent network attach by "dummy-IMSI"

It is common practice for the SIM function to permanently or temporarily disable network attach, if needed, by sending a "dummy-IMSI" to the modem as the result of a refresh command. The format of the dummy IMSI is 00101xxxxxxxxxx. Concretely, a SIM OTA system may be used to invalidating the EF_IMSI (6F07) and/or EF_SUPI (4F07) on the SIM, thus preventing the modem from using the subscription identity and attempting a network attach.

The SIM may later update the modem with a working IMSI using the refresh procedure.

Recommendation

— Consider sending a "dummy-IMSI" to the modem to permanently or temporarily disable harmful network attach behaviors by IoT Device Applications

4.5.4 eSIM profile control and fail-over

When a xUICC has eSIM capabilities, an eSIM management system may perform eSIM download, enable and delete remotely OTA. These operations can be remotely triggered in SGP.02 (via SMS) and SGP.32 (via various IP protocols or SMS). In SGP.22 all triggering is done locally/by the user of the device. An eSIM capable xUICC can store and load multiple subscription profiles, that can be enabled (generally one-at-a-time). The SGP.22 Local Profile Assistant (LPA) or SGP.32 IoT Profile assistant (IPA) software function is responsible for executing the eSIM operations, including profile fail-over.

The fail-over procedure checks if the enablement, normally following a download, of a profile leads to a successful connection to the network. If no connection is established, for any reason, the procedure re-enables the profile already loaded on the xUICC that is flagged the "fall-back" profile, if such a profile exists. The "fall-back" profile may be pre-loaded and flagged during production or downloaded. The flagging may also be local and dynamic – controlled by a number of policies.

Recommendation

— It is recommended that one eSIM profile on the xUICC is configured as fallback profile to avoid lost-device scenarios and the following network load from un-manageable endless network attach retries.
— It is recommended that this profile has mobile data and roaming enabled.

4.5.5 SIM Service Termination

In order to prevent network load from unwanted network attach attempts, a termination procedure should be included in the service lifecycle. Particularly when:

— A contract ends.
— Terms are breached.
— A device is compromised or permanently offline.

This is especially relevant for IoT deployments in inaccessible areas, where physically retrieving or replacing a SIM is not feasible.

In SGP.02 eSIM capable xUICCs, termination may be achieved by enabling an invalid profile with a “dummy-IMSI” as explained above. Note that the “fallback” flag must be managed/removed to prevent a “failover” from the invalid profile.

⚠ WARNING: This action is terminal. Once deactivated, the SIM (or eSIM profile) cannot attach to a network unless physically replaced or reprogrammed (if supported by the form factor and access).

In SGP.22 and SGP.32 eSIM capable xUICCs, termination may also be achieved by disabling the current profile. LPA/IPA implemented/embedded in the xUICC (LPAe/IPAe) allows the MNO/MVNO to perform this operation OTA, otherwise access to the eIM or the device/application management system is needed.

The benefit of the eSIM approach is that the application software may locally, via control of the LPA/IPA, re-enable the functional profile in the event of a e.g. “factory reset”. This approach consequently extends the lifetime of devices with embedded xUICCs.

Recommendation

Consider the following SIM service termination options in case of harmful IoT device behavior:

— For SGP.02 eSIM capable xUICCs: enabling an invalid profile with a “dummy-IMSI”
— For SGP.22 eSIM capable xUICCs: disabling the current profile via LPAe/IPAe

5. Recommendations Summary

5.1 Recommendations for Mobile Operators

- Consider increasing the preamble pool size for LTE-M
- Consider increasing the number of random access attempts for LTE-M.
- Consider enabling random access backoff indication for LTE-M.
- Consider a less aggressive ramping level for LTE-M.
- Consider enabling ACB
- Consider enabling EAB taking into account support levels for the feature on their UICC and in field devices.
- Consider the impact and benefit of enabling additional LTE-M narrowbands
- Implement Access Barring of recalcitrant devices only as an intermediary solution whilst elaborating a strategy to counteract the device's rogue behaviour
- Ensure full compliance of the Diameter Reject to NAS Cause codes mapping to the GSMA Roaming guidelines ([2], [5])
- Be considerate of the impact of your Steering-of-Roaming choices on the IoT Device behaviour

5.2 Requirements for IoT Device Application Developers

- Randomise reconnection attempts over at least 1 second intervals distributed across as wide as possible reconnection time windows
- Adapt the reconnection window size to reflect the number of devices in each cell and increase the reconnection windows size exponentially if attempts are unsuccessful
- When supported by the IoT Module, read and interpret the solicited and unsolicited response codes from the radio module to derive the next best action in the event of a communication failure
- Implement a top-down recovery approach across all communication layers in case of an IoT Device Application connectivity failure
- Only perform a hard reset of radio modules/UICCs to reestablish connectivity in the last resort after all higher layer reconnection attempts have failed
- Do not disable SMS features in modem or bar SMS service in their Customer Portal, even if the application itself does not utilize SMS
- Do not block https session triggered by SIM functions

5.3 Requirements for IoT Module/ Chipset Manufacturers

- Implement the latest GSMA TS.34 recommendations including the newest RPM handling recommendations for permanent and temporary rejects
- Investigate possible enhancements to the RPM functionality to address harmful behaviours by IoT Device Applications upon receipt of frequent temporary reject scenarios.
- Provide unsolicited response codes to advise higher layer software applications of the onset of random access congestion/failure.

5.4 Requirements for eUICC/eSIM/ SIM Application Providers

- Avoid excessive network traffic and power drain, especially in battery-constrained IoT devices by choosing a longer default polling period. If the application is able to trigger an SIM OTA poll – then the default period can be longer
- When a SIM OTA poll results in a data transfer, the delay to the next poll shall be short in order to complete the pending SIM OTA transactions.
- When a SIM OTA poll results in "nothing to do", the delay to the next poll shall be starting at the default delay and increased exponentially ("back-off")
- When a SIM OTA poll/pull fails, e.g. "backend not reachable", the number of retries shall be limited and use exponential delay increase in the retry timers ("back-off").
- For large scale deployments in the same power area, consider randomizing SIM ready access after power loss to avoid network signaling overload. To spread out network attach attempts over time, the SIM may start to send a "dummy-IMSI" to the modem and then - after a random power on timer in the SIM application – refresh the modem IMSI to the actual IMSI.
- The application software shall avoid using modem "power-cycle" when retrying network connect, since this will likely cause SIM delay timers to be reset inhibiting any "back-off" behavior, or interfere with power loss mitigations.
- The aforementioned polling configuration parameters shall be configurable OTA.
- Consider sending a "dummy-IMSI" to the modem to permanently or temporarily disable harmful network attach behaviors by IoT Device Applications
- Consider the following SIM service termination options in case of harmful IoT device behavior:
  - For SGP.02 eSIM capable xUICCs: enabling an invalid profile with a "dummy-IMSI"
  - For SGP.22 eSIM capable xUICCs: disabling the current profile via LPAe/IPAe

GSMA Head Office
1 Angel Lane
London
EC4R 3AB
UK

Email: info@gsma.com