5G security recommendations
Package #1

by NGMN Alliance

Version:	**1.0**
Date:	**06-May-2016**
Document Type:	**Final Deliverable (approved)**
Confidentiality Class:	**P - Public**
Authorised Recipients: (for CR documents only)	

Project:	**NGMN – 5G Security**
Editor / Submitter:	**Rémy HAREL, Steve BABBAGE**
Contributors:	**NGMN 5G security group**
Approved by / Date:	**NGMN Board, 6th May 2016**

Commercial Address:
ngmn Ltd.,
Großer Hasenpfad 30 • 60598 Frankfurt • Germany

Phone +49 69/9 07 49 98-04 • Fax +49 69/9 07 49 98-41

Registered Office:
ngmn Ltd.,
Reading Bridge House • George Street • Reading • Berkshire RG1 8LS • UK
Company registered in England and Wales n. 5932387, VAT Number: GB 918713901

Abstract

The purpose of the NGMN 5G security group, which is a sub-group of the NGMN P1 5G Architecture group, is to identify new threats and security issues that may arise with 5G. The work of the group aims at informing the 5G community, and especially SDOs (like 3GPP) of potential problems in 5G, what we should pay attention to.
This working group does not make requirements, just recommendation that people should have in mind when designing 5G networks.

The group will release several documents with different focus; this first package focuses on Improving the Access Network and also identifies DoS attacks scenarios in a 5G context. Some of the identified threats are not specific to 5G but will be emphasized in a 5G context, and thus should be very carefully considered when designing the 5G networks.
Some of the recommendations described in this document might also be applicable to the current generation mobile networks.

Document History

Date	Version	Author	Changes
Sep 09, 2015	V 0.1	Remy HAREL (ORANGE)	Creation of the template
Feb 11, 2016	V 0.2	Remy HAREL (ORANGE)	Draft, with addition of approved cases
March 23, 2016	V 0.3	Remy HAREL (ORANGE)	Addition of New key issues
March 30, 2016	V 0.4	Remy HAREL (ORANGE)	Update of DoS part + addition of Open issue section
March 31, 2016	V 0.5	Remy HAREL (ORANGE)	Update of Open issue section
March 31, 2016	V 0.6	Remy HAREL (ORANGE)	Addition of objection sections
April 04, 2016	V0.7	R. HAREL, S. BABBAGE	Minor corrections and import of Vodafone/Orange objections
April 20, 2016	V0.8	R.HAREL (ORANGE)	Delete all open issues sections for board approval
April 24, 2016	V0.9	Steve Babbage (Vodafone)	Editorial corrections
May 6, 2016	V1.0	Klaus Moschner (NGMN)	Final Updates after Board Approval

REFERENCES

Ref	Document
1	GSMA. Security Accreditation Scheme for UICC Production Standard Version 5.0, 23 April 2015.
2	GSMA. Embedded UICC Protection Profile. Version 1.1. Certified by BSI. 25 August 2015. [USIM PP] (U)SIM Java Card Platform Protection Profile Basic and SCWS Configurations Evolutive Certification Scheme for (U)SIM cards. Certified by ANSSI. 2012
3	Random Access Mechanism for RAN Overload Control in LTE/LTE-A Networks, Communications (ICC), 2015 IEEE International Conference on QoS, Reliability and Modelling Symposium
4	RAN overload control for Machine Type Communications in LTE, (2012) GC'12 Workshop: Second International Workshop on Machine-to-Machine Communications 'Key' to the Future Internet of Things
5	Framework for Cyber-Physical Systems, Release 0.8, September 2015, Cyber Physical Systems Public Working Group
6	Machine-Type Communications: Current Status and Future Perspectives Towards 5G Systems, 2015 IEEE, Communications Magazine
7	Security and privacy in the internet of things: current status and open issues, 2014 International Conference on Privacy and Security in Mobile Systems (PRISMS)
8	Design Challenges and security issues in the internet of Things, 2015 IEEE Region 10 Symposium
9	Toward an Analysis of Secure Issues, Challenges, and Open Problems in the internet of things, 2015 IEEE 11th World Congress on Services
10	NGMN 5G White Paper, version 1.0, Feb. 2015.
11	The internet of Things, an Overview, Internet Society, October 2015
12	RFC7452
13	RFC7397
14	RFC6574
15	http://www.arcep.fr/index.php?id=13131

GLOSSARY

Abbreviation	Meaning
3GPP	3rd Generation Partnership Project
ACB	Access Class Barring
CN	Core Network
DNS	Domain Name Service
DoS	Denial of Service
DDoS	Distributed Denial of Service
E2E	End to End
ECM	Enterprise Content Management
IP	Internet Protocol
MEC	Mobile Edge Computing
QoS	Quality of Service
RAN	Radio Access Network
SDN	Software Defined Network
SON	Self-Organizing Network
UE	User Equipment (from 3GPP nomenclature)
VNF	Virtualized Network Function
VPN	Virtual Private Network

Contents

1 INTRODUCTION

In the next mobile network generation, 5G, a lot of new use cases leading to a massively connected world are expected to emerge. Although this can imply new services and better user experience for the end customer, it is a challenge for mobile network actors to secure this new world.

Some services, or the desired quality of service for some services, impose constraints that may limit the security design of the architecture. For example, requirements may be expressed for very short delay / latency, or very long battery lifetime, or very high availability, or very high service bandwidth. These can sometimes constrain the security design, and necessitate compromise. Typically, in the 3GPP standardization process, such constraints are set before or during the beginning of the security design process, and then security designers have to work within those constraints and minimize any impact on security.

In the design of 5G, where tradeoffs between security and performance are likely to arise, security should be given high priority from the start of the design process. This will encourage innovative thinking to improve tradeoffs and maintain core security principles.

In this document, the NGMN 5G security group focused on ways to increase the security level of the access network. 4G is usually considered as well secured, but the 5G context can bring new threats but also opportunities to enhance the security level. During the studies, it also appears that the very large number of IoT devices could represent numerous flavors of DoS (Denial of Service) attacks, so a part of this document illustrates a few examples of threats.

Cautionary notes:

1) *5G is still in the very early stage of development. There are still many uncertainties on definitions (eg. slicing), on underlying technologies (radio interface, fronthaul and virtualization) and the key architecture choices. The core design concepts are not defined. The details of the E2E architecture and the subsystems architecture (RAN, Fronthaul, CN) remain unknown.*
2) *5G security standardization is under 3GPP SA3 responsibility.*

2 IMPROVE THE ACCESS NETWORK

2.1 Key Issue 1: Flash Network Traffic

2.1.1 Description

As the network capacity and number of UEs grow, the risk increases that large-scale events could cause significant changes in network traffic patterns, either accidental or malicious. At this scale, it is not possible to differentiate the intent of a network surge and so this recommendation covers both scenarios, although its primary focus is to prevent malicious events.

2.1.2 Recommendation

The 5G system must wherever possible minimize large swings in traffic usage, and be resilient to them when they do happen, while maintaining an acceptable level of performance.

2.2 Key Issue 2: Radio interface keys sent between operator entities

2.2.1 Description

In all of 2G, 3G and 4G, it is the case that keys for radio interface encryption (and sometimes integrity) are computed in the home core network – the AuC – and then transmitted to the visited radio network over signalling links such as SS7 or Diameter. This is a clear point of exposure, and it has been demonstrated how keys can leak.

The most direct, and clearly recommended industry approach is for operators to improve SS7 / Diameter security, e.g. by introducing SS7 firewalls. But well-designed key management protocols for 5G could also reduce the threat significantly.

In GSM terms, this is about the leakage of the cipher key K_C when sent between network nodes.

2.2.2 Recommendation

The signalling link between different operators should be better protected.

If possible, design key management for 5G in such a way that radio interface session keys (or keys from which those session keys are deterministically derived) are not sent over SS7 / Diameter / equivalent links or, that such exposure is minimised.

NOTE 1: The mechanism should still allow the visited network to comply with national laws.

NOTE 2: The objective may be that an attacker who is able to observe or request keys sent between network nodes has to act as a long term, active man-in-the-middle in order to exploit that knowledge.

The performance (e.g. latency) impact of any potential solution must of course be considered.

2.3 Key Issue 3: User plane integrity

2.3.1 Description

2G security provides no explicit integrity protection of either user plane data or control plane data. User plane data is (in most countries) encrypted, but this still provides very limited protection against a Man-In-The-Middle attacker changing that data *en route*, because encryption is linear (a stream cipher) and any checksums are also linear. 3G and 4G include cryptographic integrity protection of (some) signalling messages, but still not for user plane data.

If data integrity is needed, it will usually be enforced at the transport or application layer (with additional encryption). In special cases, bearer layer integrity (plus encryption) may represent a lower overhead solution, provided the security end-points can be appropriately aligned with the service end-points. A bearer security API (supported on the device or within the network, or ideally both) might allow a service to request that a particular security configuration be used (e.g. request integrity is switched on, request a particular end-point), and also check what bearer security is in place.

There is also a risk of a session as a whole being hijacked, and used to insert rogue data into a mobile connection (or waste resources carrying the data to the service end-point, where it will be rejected anyway). User plane integrity could help here. Or the existing 3GPP mechanism "Signaling procedure for periodic local authentication" (3GPP TS 33.102, clause 6.4.7, and 3GPP TS 33.401, clause 7.5) could be refined.

So – should 5G add integrity protection to user plane data?

2.3.2 Recommendation

In most cases, the answer is no, it shouldn't. The mobile network (visited or home) is not the best place to verify the integrity of user plane data; when data integrity is needed, it will usually make more sense to do this at the transport or application layer, typically terminating beyond the mobile network.

There is a possible exception for battery constrained IoT devices or low latency 5G devices when user plane integrity is needed. In this case, application level end to end security may involve too much overhead of data transmission in the handshakes, packet headers.

NOTE 1: The visited network should comply with national laws and roaming regulations
NOTE 2: This mechanism may be mandatory to implement but optional to use.

2.4 Key Issue 4: How much should security measures in the network be mandated?

2.4.1 Description:

It was mentioned in the Introduction that there are sometimes service-driven constraints on the security architecture, and typically then also measures to minimize the impact of those constraints. However, these impact-minimizing measures are often not mandated in current cellular specifications: they are only options or recommendations.

Furthermore, there are many security provisions in 3GPP specs and GSMA recommendations that do not affect interoperability (mainly interoperability with UEs), and historically these have been treated as optional to deploy. Examples include security on the DIAMETER-based interconnect network and security on the backhaul link, which are often poorly protected in practice.

Unfortunately, this optionality may lead to moral hazard where operators are unavoidably affected through the actions of other operators, not by their own fault (cf. weaknesses on the SS7 network). Also, it undermines system-level security assumptions e.g. the assumption that a UE can implicitly authenticate an MME because the HSS already did so when handing out a K_{ASME}, or that a UE can implicitly authenticate an eNodeB because the MME already did so when handing out a K_{eNB}. If such authentication steps do not happen, then the 4G key hierarchy cannot achieve one of its design goals, namely protecting customers against false base stations.

It is quite possible to reduce the dependency on the security in the access network, or on the security provided on intra-network interfaces, but it is unlikely that this dependency can be eliminated completely.

NOTE: The issue of mandatory security will apply to many areas of 5G security, not just improvements to the access network.

2.4.2 Recommendation:

Study whether it would be helpful to mandate the use of security measures in the network to a higher extent e.g. by designing the architecture such that security measures must be used, or else the system will not work at all. Study ways to guarantee a high minimum level of security across operator networks.

2.5 Key Issue 5: There are challenges deploying consistent and effective subscriber/device level security policies within and across operator networks

2.5.1 Description

Subscriber level security procedures are implemented within the operator network and possibly on the UE itself and are designed to protect subscribers/devices from security and privacy attacks. Some forms of security and privacy policies take the form of personal firewalls deployed at different physical locations within the operator network. One great challenge today in implementing such procedures is the need to apply and enforce such policies *consistently* for different subscribers in multiple locations (UE, PGW, and Internet Access Point) within an operator network. This is important in 5G scenarios where it is envisioned that densification requirements needed to offer better RF coverage for a larger number of devices will drive the number of network elements in the operator network to increase.

In addition, inconsistent or irreconcilable ingress/egress firewall security policies among roaming partners can prevent some roaming traffic bound for the internet from working correctly, if at all. This is important in 5G scenarios where local breakout scenarios will be much more prevalent. These local break-out scenarios will cause roaming IP traffic to be routed over a visited network where firewall security policies will very likely differ from the same policies deployed in the home network. The reasons motivating the need for local break-out scenarios in 5G are due to the need to reduce latencies and the planed deployment of wireless functionality together with subscriber content at the edge of the network (Mobile Edge Computing).

2.5.2 Recommendation:

The envisioned use of network slices in 5G with the likely realization using virtualization represents a great opportunity to address these challenges in a consistent and effective way. Network slices will provide a level of traffic isolation necessary for implementing effective security policies that will prevent traffic from leaking inside or outsides the defined boundaries of the network slice itself. In addition, virtualization will enable realization of virtual security domains per subscriber in software. Security policies will be implemented within network virtual functions making it easy to implement a consistent and effective deployment of security policies across such domains.

3 DOS ATTACKS AGAINST NETWORK INFRASTRUCTURE

The internet of people is becoming the internet of devices (things). The research community and the industry are preparing for the next (5th) generation telecommunications networks to support a tremendous growth in the number of network connections due to the expected proliferation of many different types of connected devices possessing a wide range of operating capabilities as well as limitations. In the context of such growth, DOS and DDOS attacks originating from such devices are very likely to become a real threat for operators of 5G networks. Attacks against devices controlling critical infrastructure in areas such as energy, transportation, Health, Financial, Telecommunication and many others can have life threatening consequences with tremendous human and capital costs.

It is important to note that today DOS or DDOS attacks originate from machines, physically distributed over possibly very long distances, connected to the internet against operator data centres. In 5G, DOS and DDOS attacks against critical infrastructure will leverage and possibly also target the access network of a cellular operator, and likely originate from specific physical/geographical locations via a very large number of connected devices as envisioned for 5G use cases.

This contribution will focus on Denial of Service attacks in the context of 5G. DOS attacks are designed to exhaust physical and logical resources of the target. In the context of this contribution two different type of attacks are identified:

1. [**Network Infrastructure DOS Attacks**] Attacks designed to *directly* deplete the resources of the network infrastructure which supports 5G users and devices. In this case, the original target of the attack is the operator infrastructure independent of the fact that an infrastructure attack will also indirectly impact all devices and subscribers that are served by the attacked network infrastructure.

2. [**Devices/Users DOS Attacks**] Attacks designed to *directly* deplete logical and physical resources of 5G users and devices. In this case, the original targets of the attacks are the devices and or subscribers independent of the fact that an attack on a large number of subscribers or devices can indirectly impact a large portion of an operator infrastructure.

DOS attacks against network infrastructure would likely target the resources tied to sustaining “connectivity” and “bandwidth” at promised levels of service and focus on the attacks against the following areas:

1. Signalling Plane needed for authentication and support of appropriate connectivity, bandwidth and mobility requirements for 5G users and devices
2. User Plane needed to support two-way device/user data communication.
3. Management Plane needed to support configuration of network elements which support both signalling and user planes
4. Support Systems needed to provision and bill users/devices for operation on the network
5. Radio resources enabling access for many devices/subscribers
6. Logical and physical resources used within the virtualized infrastructure used to support network clouds
7. Others?

DOS attacks against users or devices would likely target at least the following physical resources [7,8,9]:

1. Battery
2. Memory
3. Disk
4. CPU
5. Radio
6. Actuators and Sensors
7. Others?

DOS attacks against users or devices would likely target at least the following logical resources [7,8,9]:

1. Operating Systems
2. Applications
3. Configuration Data
4. Patching Support System
5. Others?

3.1 **Key Issue 1 :** Overload of the signaling plane by a huge number of infected M2M/IOT devices which simultaneously attempt to gain access to the network

3.1.1 Description

While machine to machine use cases are supported today in 2G, 3G and 4G networks, it is envisioned that 5G will support billions of wireless identifiable IoT devices [6]. Many of these low cost M2M devices have several limitations in terms of computational capabilities, energy support, memory capabilities, thin operating systems with limited patching capabilities, with easy to tamper physical structures. This increases the likelihood that these devices can be operating in a compromised environment which will allow coordinated DoS and DDoS attacks against the radio access network.

The attack described in this section seeks to deplete network resources by creating a very large number of *simultaneous network access attempts* in specific geographical locations thus exhausting the local radio resources of the network. Since there aren't sufficient radio resources to simultaneously provide service to every single device attached to the network an attacker can trigger simultaneous, continuous and massive acquisition of radio resources (trying to connect to the network) to cause a signalling plane overload. In addition, these devices can also behave abnormally due to unexpected non-malicious events (e.g. natural disasters like earthquakes, floods, tornados, etc) and produce "flash crowd" type situations which will also cause depletion of the radio resources.

While current LTE Overload controls would prevent all devices (malicious and not) from attempting to access the network (signalling plane), no method exists today which can coordinate specific devices from attempting to access the network. Current overload controls rely on the MME to initiate the "START" and "STOP" overload procedures as needed, however currently the MME cannot select UEs to target for the overload procedures.

3.1.2 Recommendation:

A flood of signalling messages (natural or malicious) carried out by a large number of IOT devices in specific geographical locations should not bring down the signalling plane. At present, network solutions deployed to address flash crowds tend to rely on replication of resources (redundancy) to address potential failures of elements in the network. These solutions while capable of addressing unplanned non-malicious failure of a "fixed" number of network elements are not adequate to address sustained malicious attacks designed to deplete network resources. It becomes critical to be able to differentiate a natural flash crowd from a malicious DOS/DDOS attack.

While normally it would be challenging to distinguish malicious DoS/DDoS attacks versus an accidental flash crowd, the very persistent and consistent network access behaviour of many of these devices (since there is no human-in-the-loop) should allow use of analytical techniques like anomaly detection to be leveraged for such analysis [5]. . The access control decision that needs to be enforced in this scenario would entail in a simple case, denying or restricting access to the network for all new connections while the overload condition persists. In a more sophisticated case, the access control decision would entail recognizing via inference non-standard network access request patterns and selectively denying network access for those specific connections. This approach is subject to the possibility of mistaken identity due to "false positive and false

negative" results in the anomaly detection analysis results. It is important to study the possibility to modify current overload procedures to be applicable to situations where the overload condition is initiated by malicious actors with sufficient control over the infected devices that they do not comply with network initiated overload procedures.

3.2 Key Issue 2 : Overload of the signaling plane by a huge number of infected M2M/IOT devices which transmit intermittently and simultaneously

3.2.1 Description

The attack described in this section occurs when a very large number of compromised devices have already been granted access to the network; and manage to coordinate very short data transfers followed by periods of inactivity on a periodic basis in specific physical locations. Such behaviour would force radio and network resources to be allocated to support the data transfer and to be released to support access to the network by new connections. A malicious attack could then impact the signalling plane by:

- Depleting signalling plane processing and computation resources needed to support allocation and release of needed radio/network resources
- Depleting radio and network resources needed by new devices trying to establish new network connections

3.2.2 Recommendation:

"Chatty" behaviour (natural or malicious) carried out by a large number of IOT devices in specific geographical locations should not bring down the signalling plane. While recommended behaviour for IOT devices should avoid timed/synchronized data transfers (by utilizing random network access timers across devices used in a given geographic area), the very persistent and consistent data transfer behaviour of many of these devices (since there is no human-in-the-loop) should allow use of analytical techniques like anomaly detection to be leveraged for such analysis [5].

It is important to study the possibility to modify current overload procedures to be applicable to situations where the overload condition is initiated by malicious actors and would, as a result, be designed to avoid effective detection by for instance, changing randomly the access times and duration of data transfers. The access control decision that needs to be enforced in this scenario for devices that are believed to be malicious would require identifying via inference non conformant data transfer patterns for specific devices and selectively targeting those devices to be disconnected from the network. This approach is subject to the possibility of mistaken identity due to "false positives and false negative" results in the anomaly detection analysis results. Depending on the level of compromise an infected device may choose to not comply with network initiated overload procedures. Methods should be studied to identify offending device/s and target countermeasures which prevent or limit network access against such device/s.

3.3 Key Issue 3 : Deliberate triggering of network overload mechanisms

3.3.1 Description

Mobile networks may have overload mechanisms that are triggered when overload (deliberate or otherwise) is detected. These overload mechanisms typically throttle or prevent access to resources, temporarily, until the overload has cleared. Typically, all resources in affected cells are throttled or blocked for a period of time.

But then these overload mechanisms are, themselves, denying service. If an attacker can trick the network into thinking that an overload is occurring, and so into triggering the defense mechanisms, then this could be an efficient way to carry out denial of service.

This is an indirect type of attack, where the reaction (triggering of overload procedures) to the perceived overload condition causes a problem (preventing all device access to the cell) worse than the problem that the network is trying to address.

3.3.2 Recommendation:

Design network overload mechanisms (the triggers and the consequent action) in such a way as to reduce this risk. For example:

- Use a series of overload defenses that gradually increase in severity. Rather than immediately activating the most severe defense mechanism, start with the mildest, and increase gradually until the problem goes away.
- Assess proposed overload mechanisms against this risk of deliberate triggering. Verify that severe defense mechanisms can only be deliberately triggered by attacks that would be expensive to carry out.
- If possible, design mechanisms that limit service to the devices that are causing the problem, while still allowing adequate service to others.

3.4 Key Issue 4 : Bulk configuration, typical of 5G scenarios tied to connected devices, highlights new DOS attack vectors against provisioning and configuration systems

3.4.1 Description

It is envisioned that in the future many of the devices deployed in 5G networks will be configured by a variety of provisioning systems, many of which will be owned and operated by 3rd parties instead of traditional cellular operators. Due to the large number of devices that will need to be provisioned, “Bulk Provisioning” will become a de-facto feature supported by these systems, which will allow quick and massive-scale provisioning of such devices. In these bulk provisioning scenarios it would be easy to misconfigure/mis-provision such devices to force erratic/incorrect behaviour. Such attacks can cause major disruption to the systems supported by these devices.

3.4.2 Recommendation:

The very persistent/consistent behaviour of many of these devices (since there is no human-in-the-loop) should allow use of analytical techniques like anomaly detection to be used to detect non conformant device behaviour. The normal behaviour range of a given IOT/M2M device is often well understood (the province of reliability engineering). These features may make anomaly detection and control easier (the province of resilience engineering, especially when the anomalies were the result of someone’s desire to produce adverse effects). IOT/M2M device have comparatively well-defined network dynamics: servers rarely change, the topology is often but not always fixed (i.e. mobile devices), the user population is relatively stable, communication patterns are often regular, and the number of protocols is limited. These parameters can be modelled, and the model of the dynamics of the system can be used to detect a compromised node or identify out-of-norm behaviour and isolate such devices [5].