5G security

Giuseppe Bianchi

Abstract Besides significantly outperforming past generations in terms of capacity and throughput, 5G networks and systems will provide an infrastructure for the support of highly diversified and heterogeneous services. Indeed, many heterogeneous application contexts will be profoundly impacted by 5G systems - to mention a few: Industrial Internet and smart control systems, autonomous vehicles and drones, life-critical e-health and remote surgery, virtual reality and augmented reality, remote diagnostic and preventive maintenance, and so on. Service customization and ultra-rapid deployment will leverage a virtualized network infrastructure, flexibly integrating software-based network functions in both the network core, as well as relocating time critical and low latency processing tasks down to the network edge. Such a diversified and heterogeneous scenario calls for radically new security models, capable to overcome the *"one-size-fits-all"* approach to security that has characterized cellular systems until the latest fourth generation. In this chapter, after a brief excursus on how security has evolved throughout the various generations of cellular systems, we will focus on the security vision in 5G, and on the relevant major approaches and challenges, including the discussion of novel threats to virtualized systems.

1 Security in pre-5G systems

The problem of security in cellular systems has arisen initially to solve a very specific problem: how to authenticate users connecting to the network, and protect the relevant data in transit from attackers able to eavesdrop the radio channel. This activity has been duly addressed during the previous generations of cellular systems, with

Giuseppe Bianchi
Universitá degli studi di Roma Tor Vergata, Dipartimento di Ingegneria Elettronica
Via del Politecnico 1, 00133, Roma
e-mail: giuseppe.bianchi@uniroma2.it

solutions that, although gradually, have today reached a level of protection considered completely satisfactory, to the extent that it is hard to find any breakthrough improvement in this area in the last fifteen years.

The (in)Security of 1G and 2G systems

While the first generation systems did not specify any solution for the protection of communications, GSM (i.e., the second generation) began to explicitly introduce solutions for user authentication and encryption at the radio interface level. However, GSM security solutions have proven to be extremely preliminary and insufficient for many reasons, starting from the total inadequacy of the cryptographic techniques adopted, and the nefarious *"security by obscurity"* strategy employed. Specifically, the cryptographic algorithm adopted in the GSM authentication handshake, later on called COMP-128, had not been validated by the cryptographic community, but had been deliberately secreted, with the idea - a posteriori proven disastrous - that the secrecy of the algorithm itself could increase security. Unfortunately, this was not the case. As soon as, around 1998, the details of the COMP-128 algorithm leaked out, it was a matter of a a few weeks for the cryptographic community to badly break it, and prove its complete inadequacy!

Still, the bad experience with GSM taught the security community a very important twofold lesson. First, it is not nearly easy to prevent leakage when an highly critical algorithm must be implemented by numerous actors (operators, SIM manufacturers, authentication center producers, etc), and when the devices that implement it can be subject to "reverse engineering" of the related software code. Second, it is extremely unlikely that an algorithm not specifically developed by professional cryptographers, nor duly analyzed by the cryptographic community, results to be sufficiently robust.

Although the cryptographic algorithm is the most striking weakness of second generation systems, the security shortcomings of GSM were not limited to this. In particular, GSM did not provide mutual authentication. In fact, in GSM systems, while the user needed to undergo authentication before being permitted to access the network, the opposite was not true, i.e., the user was not provided and means to verify the authenticity of the radio station to which he was attacking to. This feature was perhaps not perceived as critical during the early years when the GSM system was being standardized. But the technological evolution occurred during the late '90s, with the emergence of programmable radio devices (Software-Defined Radio), has made not only possible but even quite cheap attacks based on *"rogue base stations"*, i.e., fictitious radio stations controlled by an attacker, which was thus made able to intercept and tamper with the end users' communications.

Finally, no security solution in the core network part had been standardized in GSM. The encryption on the radio interface terminated in the access network; therefore the information was transported in the clear on the fixed network, with the result that any attacker able to access the transport infrastructure was able to violate the confidentiality and integrity of the data transported.

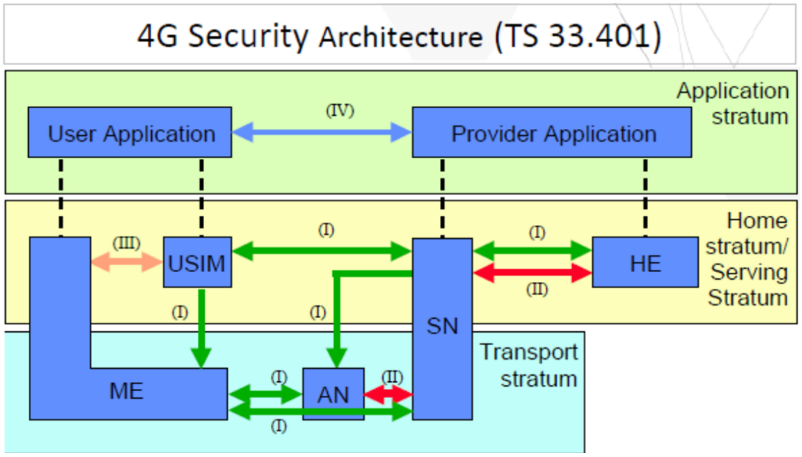

Fig. 1 4G security architecture: roman numbers on top of each arrow specify which security domain is involved for the considered interface.

3G: the security generation.

The next third generation, UMTS, was probably the generation in which the greatest progress has been made on security, both in terms of the quality of the solutions adopted and in terms of interventions in the various sub-systems involved. Firstly, 3G systems have completely abandoned the *"security by obscurity"*, by adopted publicly scrutinized cryptographic algorithms in the AES (Advanced Encryption Standard) family, algorithms which are much safer than previous ones, and in part are still today state of the art. The application of cryptographic techniques has also been significantly improved, both through explicit differentiation of encryption ciphers and relevant keys from data integrity, and through the introduction of privacy and protection features for users against attacks devised to recognize and track the end user position (location privacy). 3G systems have also obviated the problem of rogue base stations, by providing an extremely effective technique of mutual authentication. Finally, 5G systems have duly addressed security in the netwpork core by providing uthentication and protection mechanisms for both data transport in the cose as well as protection of signalling.

Security systematization and 4G.

In line with the progress made in 3G systems, the fourth generation has made several improvements, and has, above all, clearly inserted the theme of *"security by design"*, i.e., by addressing security since the very beginning of the LTE architectural specification phase. To this purpose, the overall 4G security architecture (Figure 1) has been organized in five explicit domains:

I **Network access security**: protection at the level of air interface and secure access to the service by the user;

II **Network domain security**: protect network elements and relevant exchange of data traffic and signalling messages;

III **User domain security**: protection of the mobile terminal and its interfacing with the USIM and device;

IV **Application domain security**: secure communications at the application layer;

V **Visibility and configuration of security**: means to permit to check if (and which) security features are in operation, and how they are configured.

In addition, 4G systems brough about many punctual improvements (and whenever applicable also fixes) in the specific algorithms and techniques employed. These improvements range from better authentication and key management, improved cryptographic algorithms (including support for a new stream cipher called ZUC), end-to-end security, integration with IP security technologies, etc. As such, they are quite technical and incremental; we will see in the next section that a much more radical revision of the security model is expected in 5G systems.

2 5G security: vision

As briefly summarized in the previous section, the past generations of networks and cellular systems have made huge steps in the security sector, both in terms of cryptographic algorithms (starting from 3G systems) as well as in fostering a *"security by design"* approach and a relevant systematic organization of the security tasks into five security domains. Given these advanced results attained by previous generations, a question stands out: should security still play an important role in 5G networks? Or the bulk of the security work has already been done, and 5G security is expected to move along very detailed and quite incremental steps?

Although not officially included in any standard document, several 5G stakeholders and alliances converge upon a 5G security vision revolving around three major driving principles:

- Flexible security;
- Supreme built-in security;
- Automation.

The first item, flexible security, is in our opinion by far the most significant change that 5G systems are called to address. Indeed, it is a direct consequence of the radical change in perspective that characterizes the new generation of wireless networks. While the previous 2/3/4G systems were specified for a well-identified class of users, and had therefore defined a specific and unique family of security and data protection solutions, the emerging 5G systems are born with the aim of supporting extremely diversified vertical services, targeting different types of users, and including services not exclusively dedicated to human users.

It follows that the *"one-size-fits-all"* paradigm, which had characterized security solutions in previous generations, now becomes a questionable approach, which is hardly applied to very diverse services encompassing heterogeneous technologies and terminals with widely different capabilities and service requirements.

The **“Flexible Security”** paradigm promoted by 5G networks is devised to meet such changed needs, by providing diversified security solutions, adaptable to the specific (different) scenarios considered. For example, services with very low latency requirements will need security solutions capable to meet such tight latency constraints. And whenever the need for trade-offs emerge, service deployers should be made able to decide whether to downgrade security requirements or performance ones; in essence, 5G systems mandate for an increased flexibility in the ability of the network to deploy the security technologies and solutions most appropriate to a given scenario.

Some concrete initial steps towards flexible security have been made, with the introduction of a new protocol, *5G EAP-AKA*, specifically meant to support flexible authentication in emerging 5G systems. Indeed, while authentication was based until 4G on a single and well-defined approach, diverse 5G services will have to take into account numerous diversification factors. For instance, which terminal devices or end users do we need to authenticate? And does the device come along with a SIM, or it is a SIM-less terminal? And are there constraints and limitations to take into account, such as energy consumption or computational capacity? And what level of authentication and authorization must we provide to the different types of “mission critical” services? The rationale of EAP-AKA is to act as *“protocol container”*, so as to permit support for more tailored authentication handshakes to be deployed in specific service contexts.

As for the other two points mentioned above, the vision underlying the idea of having a **“built-in security”**, i.e., directly integrated since the early specification stages of the emerging 5G network, is arguably not completely new (4G systems fostered a similar vision), but it remains a crucial and fundamental vision. Moreover, 4G systems have left many challenges still widely open, and there are several specific areas where the level of security of systems based on previous generations can be significantly improved with solutions directly integrated into 5G systems. These areas (at the very least!) include enhanced privacy, security assurance, and the need to increase robustness against cyber-attacks owing to the alarming pace at which malware is evolving on mobile devices.

Finally, in relation to the third point mentioned above - **automation** -, the flexibility of security management must go hand in hand with tools to simplify the management of security itself in the network, and allow a quick adaptation not only of the operation of the network, but also of the security solutions in place, to new emerging services.

3 Network virtualization and softwarization: implications on security

A key feature of 5G systems is the migration towards a virtualized network infrastructure: network functions are no longer provided by dedicated hardware devices or components, but are implemented in software and run inside virtual machines or containers, which in turn reside in cloud architectures or, whenever necessary, are relocated to the edge of the network so as to attain the low latency requirements required by critical 5G applications.

In terms of security, the key advantage inherent in the virtualization of network functions, and in the simplification in the related control and management offered by the *"Software Defined Networking"* paradigm, consists in being able to on-demand deploy security-related network functions, whenever and where they become necessary. Such security-related functions include, but of course are not limited to: traffic analysis modules, firewalls, intrusion detection and prevention systems, Deep Packet Inspection systems, programmable and relocatable network probes for detection of modern sophisticated intrusions (e.g. Advanced Persistent Threats make use of lateral movements), and so on. Their re-implementation as dynamically deployable and reconfigurable virtualized network functions (opposed to their implementation into dedicated components or hardware modules) is expected to foster an unprecedented level of flexibility in the management of security. Indeed, this new dynamic and flexible management paradigm offers a significant opportunity to rethink security functions following a *"security-as-a-service"* paradigm, as well as automate network protection features.

To achieve this goal, however, scientific research is currently still engaged in at least two fundamental challenges. Firstly, traffic analysis or packet inspection modules, made of software running in virtual machines, can hardly achieve performance comparable to those attained by dedicated hardware components, leaving the implementation of scalable and high-performance modules as a top priority. Secondly, a complete integration of security services in the network management and orchestration platforms is still ongoing. Challenges include the identification of platform-agnostic programming interfaces for heterogeneous security and monitoring solutions, as well as compelling visualization of the security status of the network through appropriate visualization techniques.

At the same time, the new paradigm of virtualization of network functions brings about new problems and new challenges in terms of security, given the significant expansion of the attack surface that the softwarization and programmability of network functions entails. On the one hand, the ability to migrate network functions as virtual machines leads to the need to identify (or adapt from the IT world) secure authorization, management, migration and secure attestation/execution of SW images that implement such virtualized network functions. On the other hand, the physical separation of such functions in separate devices is no longer applicable as a security measure, and it is necessary to think of separation and isolation solutions of such functions in a virtualized context, including means to strengthen the isolation for the “slices” that make up the network and for the functions responsible for such segmentation, and means to provide independent security solutions for every “slice” in which the network will be divided. However, as we will discuss specifically in the next section, this activity is only apparently simple, and fundamental problems recently emerging in the processor architectures used in virtualization infrastructures may play havoc with the above isolation requirements.

4 Critical vulnerabilities in modern processors: implications on isolation

The last few months have seen the emergence of a brand new wave of attacks, which leverage fundamental performance optimizations tightly integrated inside the architecture of modern processors. Such attacks are broadly referred to as *“transient execution attacks”*, as they exploit the capability of processors to “anticipate” computation (e.g., via branch prediction and out-of-order execution) so as to speed up performance. The first two attacks of this family, namely *Spectre*[1] and *Meltdown*[2], showed that exception or branch mis-prediction events might leave unauthorized data in the CPU’s micro-architectural state, e.g., in low level caches.

What truly concerns the security community is that *Spectre* and *Meltdown* were not isolated cases, but specific instances of a far more fundamental and general new threat - a Pandora’s Box was opened! As a matter of fact, recent chronicles report that, on August 14th of this year (2018), a new variant of such attacks called *Foreshadow* was announced. Foreshadow had an even more disastrous impact than previous attacks. First, it undermines the security and trust model based on Intel SGX, a hardware-based trusted software attestation considered so far a fundamental component in cloud security. Moreover, as we will discuss in more details later on, a version of Foreshadow pointed out the emergence of critical vulnerabilities in virtualization-based isolation. And such a nightmare is still ongoing: a few days ago, on November 13, 2018, while we were writing this chapter, a systematic analysis of this methodology of attack was published[3]. This analysis allowed not only to discover seven brand new variants, but also showed that the problem is not limited to Intel processors, but also extends to other vendors, namely ARM and AMD. Moreover, this work shows that many of these new variants of attack do not appear mitigated by the security patches issued so far by the processors’ vendors.

Transient execution attacks and virtualization

Our specific concern for 5G systems is that such new wave of transient execution

[1] P. Kocher et. al., *Spectre Attacks: Exploiting Speculative Execution*, 40th IEEE Symposium on Security and Privacy, 2019.

[2] M. Lipp et. al., *Meltdown: Reading Kernel Memory from User Space*, 27th USENIX Security Symposium, 2018

[3] C. Canella et. al., *A systematic evaluation of transient execution attacks and defenses*, eprint arXiv:1811.05441, Nov. 13, 2018, available at https://arxiv.org/abs/1811.05441

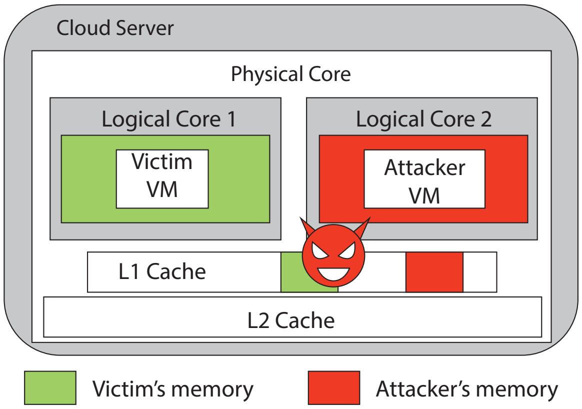

Fig. 2 *Foreshadow-VMM attack: high-level overview*

attacks may play havoc with the desire of virtual infrastructure operators to rely on virtualization for network slicing's isolation and segregation. A first threat to isolation in virtualized systems was concretely shown by the Foreshadow[4] attack. Foreshadow, also known as L1 Terminal Fault, is a speculative execution attack that provides the possibility to completely bypass the virtual memory abstraction, thus providing means to read unauthorized data. For performance reasons, processors use speculative execution during the virtual-to-physical memory address translation. In particular, while the correspondence between virtual and physical addresses is searched in the page table (i.e., during a page table walk), the processor accesses in parallel the L1 data cache. If the logical address has not a mapping to the physical location, the translation process is aborted and a terminal fault rises. However, there is a time period before the retrieve operation in which data are still passed to the cache, even if an access violation occurs, and thus where tailored side-channel methods, similar to those used in the Spectre and Meltdown attacks, may be exploited to gather access to protected information.

There are three different variants of the Foreshadow attack: (i) *Foreshadow-SGX*, the first Foreshadow version, designed to infer data from SGX trusted execution environment; (ii) *Foreshadow-OS/SMM* affecting operating system, kernel memory, and system management memory; (iii) *Foreshadow-VMM* affecting virtual machines (VMs) and hypervisors (VMM). Our specific interest here is on the last variant, Foreshadow-VMM, as it threatens virtualized environments (thus including virtualized network scenarios), by allowing a malicious guest VM to read memory belonging to the VM's hypervisor, as illustrated in Figure 2.

Referring the reader to the original sources for most technical details, in extreme summary the attack relies on the following facts. let us preliminarily recall that the virtual memory exposes to the user is divided in chunks called pages, which are

[4] O. Weisse et. al., *Foreshadow-NG: Breaking the Virtual Memory Abstraction with Transient Out-of-Order Execution*, Technical report, 2018, available at https://foreshadowattack.eu/

then mapped to the physical memory using Page Tables which contains all pairs of $\langle virtual\_address, physical\_address \rangle$. For performance reasons, modern processors maintain a cache of recently translated addresses, and forward the content of the Page Table Entries directly to the cache control logic *while* simultaneously performing relevant checks, thus *including whether an entry is valid*. This implies that, even if a *"not present"* fault is raised, there is a relatively small period of time in which the processor will continue to speculate on the validity of the data stored in the L1 cache (and thus forward such cached data to the relevant processing instructions), until the fault takes effect. In this period of time, a side-channel attack similar in concept to Meltdown can be therefore used to read data from *any arbitrary physical address*, as long as that address is currently loaded in the L1 data cache and a *not present* page table entry is triggered for that address. In practice, such an attack becomes feasible as long as the attacker succeeds in deploying a malicious VM on the very same core of a chosen victim's VM. Indeed, once this is done, since the malicious VM (as any other VM) controls its own virtual-to-physical address translation, and since this information goes directly inside the L1 cache, it suffices to instruct a malicious VM with an offending kernel module[5] to modify the page table entry of its own page table so as to include any desired physical address, pass this table entry to the L1 cache, and trigger a terminal fault.

Implications on 5G systems

The emergence of such a new generation of *transient execution* attacks should loudly alert the 5G community about the potential threats which may affect high-level segmentation and slicing techniques, i.e., techniques which target isolation while remaining agnostic to low level details. As discussed above, even if logically isolated, a virtual machine which shares a same CPU micro-architectural state with a victim's VM (i.e., pinned to the same core) may exploit the shared physical cache to get access to unauthorized information. We therefore believe that supplementary attention should be posed on how to extend current slicing and isolation frameworks with supplementary policies which specify further *physical* micro-isolation requirements - e.g. by preventing that system-critical VMs share the same hardware with user-loaded VMs, or that VMs belonging to two different slides share the same core. This is especially important when deploying segmentation at the network edge, as this is arguably the place where, on one side, low-end commodity hardware will be exploited, and on the other side, non-necessarily trusted computing tasks provided by end users will reside and might therefore attempt to interact with other VMs. Finally, the above described attack has so far been proven only in a scenario comprising virtual machines running on an hypervisor, while Network Function Virtualization is nowadays moving towards more performing approaches, such as *containers* or *unikernels*. Still, we do not nearly expect such technologies to be exempt from such types of attacks.

[5] The feasibility of such a solution was concretely tested by our own group - we are currently looking at simpler (user-space) techniques not even requiring the injection of VMs with a modified kernel.

5 Conclusions

The goal of this chapter was to raise attention on the key aspects and challenges related to 5G security. Rather than giving a detailed, but probably boring, summary of the punctual and/or incremental security improvements planned in the various components and strata of the emerging 5G systems, we preferred to focus on a more conceptual presentation of the new directions that 5G systems are expected to take, and on the novel security issues that the fundamental 5G revolution in terms of infrastructure' softwarization and virtualization is raising.

Obviously, many supplementary security challenges are brought about by *each* of the three specific innovative service scenarios fostered by 5G, namely: Enhanced Mobile Broadband, Ultra-reliable and low latency communications, and Massive Internet of Things (or machine type) communications. For reasons of space we have limited to discuss how the *"Flexible Security"* paradigm promoted by 5G networks is devised to meet the very diverse security needs of such services, as well as other vertical ones. But it is worth to remark that, especially in the case of IoT, the ability of an attacker to gather control of millions of devices (ability proven feasible by Mirai, Bashlite, or Aidra - just to mention some first generation IoT botnets' instances) may lead to new attack scenarios of unprecedented scale and impact (just imagine a massive IoT ransomware taking control of your homes and assets).

Last but not least, the very large number of stakeholders involved in vertical services will require crucial attention to the privacy of the users and the controlled disclosure of sensitive data. We advocate explicit rules and policies devised to govern the way by means of which user profiling activities are conducted, and users' behavioral information is gathered and exploited to provide personalized and customized service models.