

Mobile Communication Security Controllers An Evaluation Paper

Keith E. Mayes, Konstantinos Markantonakis
Information Security Group Smart Card Centre
Royal Holloway, University of London
Egham, Surrey England
(keith.mayes, k.markantonakis) @rhul.ac.uk

Abstract

Cellular communication via a traditional mobile handset is a ubiquitous part of modern life and as device technology and network performance continues to advance, it becomes possible for laptop computers, Personal Digital Assistants (PDAs) and even electrical meters to better exploit mobile networks for wireless communication. As the diverse demands for network access and value added services increase, so does the importance of maintaining secure and consistent access controls. A critical and well proven component of the GSM GSM and UMTS security solution is the smart card¹ in the form of the SIM or USIM respectively. However with the enlarged range of communications devices, some manufacturers claim that the hardware selection, chip design, operating system implementation and security concept are different from traditional mobile phones. This has led to a suggestion that types of “Software SIM” should be used as an alternative to the smart card based solution. This paper investigates the suggestion.

¹This has also extended into some regions using variants of CDMA standards where the RUIM is specified.

1 Introduction

Mobile communication and computing technology have evolved at a remarkable pace and it is now possible to propose new functionality and services that even a few years ago would have been dismissed as impossible. Compared to today, the mobile telephony pioneers of the 1980s worked in a much more restricted technical environment and with a different set of design priorities and associated assumptions. However, it is not only the underlying technology that has changed, but also the market's attitude and expectation towards communication and computing, largely fuelled by ubiquitous internet connectivity, the boom in email, on-line purchasing and increasingly the exchange of digital content and sensitive information. Whilst there are many initiatives aimed to offer new services to individual customers, there is growing interest in machine to machine communications and telemetry systems. Even boundaries between business segments are breaking down as the distinction between wireless and fixed communications is increasingly blurred and if we consider modern Near Field Communication (NFC)² we will see mobile phones acting as credit cards, train tickets or smart poster readers.

Whilst many of the changes are positive and exciting, the long standing problem of securing systems and services against criminal mis-use remains and indeed has become more acute as electronic applications and transactions have become complex and remote. What is also evident is that the great technological advances that are enabling modern services, are also being exploited by attackers. In the case of the GSM mobile communications standard (GSM) the attackers have been successfully repelled (with a few avoidable exceptions) by a device known as the Subscriber Identity Module (SIM). The SIM is essentially an application hosted by a specialised microcontroller (optimised for security) that is normally housed within a plug-in format smart card. There were very good reasons for introducing the SIM for GSM and its design has now been elegantly upgraded to an advanced USIM application for UMTS³ third generation networks. There is also a smart card called a RUIM defined in some competing cellular standards⁴. However, despite its proven track record there will always be parties that question the necessity of a separate and removable hardware component to support security and/or whether the SIM is really the best solution for the 21st Century. This paper will address such questions in an attempt to provide guidance to the industry.

The discussion will begin in Section 2 by recapping on the reasons why the SIM was adopted into standards and the main role that it takes in a mobile communications network. Section 3 first considers the trust relationships of the various parties involved in SIM/handset supply, usage and maintenance with respect to traditional and new categories of cellular usage. Section 4 goes on to extract the security fundamentals and critical capabilities that are necessary to carry out the range of trusted/secure operations in the various communication scenarios. As a fundamental reason for using hardware security modules such as the SIM is the ability to resist anticipated security attacks; Section 5 introduces the broad range of attacks that a smart card would be expected to defend against. Section 6 introduces some candidate software SIM solutions and compares them to the conventional SIM. Section 7 considers the Trusted Platform security element as a potential complementary or alternative technology to the SIM smart card. The overall findings are further discussed and summarised in Section 8, leading to the final concluding remarks.

Note that unless there is a need to show a distinction, the term SIM will imply a SIM or USIM application implemented on a conventional smart card platform. Similarly, the term SIM Application will refer to the SIM or USIM functionality, independent of the hardware platform. The term Mobile Equipment (ME) will be used to refer to all types of cellular communication device such as a mobile phone or data modem, unless there is a need to differentiate between devices.

2 Near Field Communication (NFC) is similar to a contactless smart card/RFID interface for mobile phones

3 UMTS is the successor to GSM, initially standardised by ETSI and now by the Third Generation Partnership Project (3GPP)

4 Standards defined by 3GPP2 – relating to IS95/CDMA2000

2 An Overview of the SIM

Before considering the detailed trust issues surrounding the role of the SIM, it is essential to understand why the SIM exists in the first place. In fact, one of the main reasons was to overcome problems caused by weak security mechanisms implemented in early mobile phones. For example, before the GSM digital phone networks appeared in the UK, there was an analog system called Total Access Communications (TACS). For a mobile to be allowed access to the network, it needed to transmit two identifiers; its telephone number (MSISDN) and a unique electronic serial number (ESN). If the MSISDN-ESN pairing matched the network's own records then the mobile was judged legitimate and was allowed access. Unfortunately, there was no algorithm available to provide confidentiality and so the phone transmissions were not encrypted, making it relatively easy for an attacker to eavesdrop the signalling exchanges and discover the MSISDN-ESN pairs. Taking a modern-day analogy for the network access procedure, the phone was acting as a very simple RFID [1]. Basic RFIDs can be observed and copied/cloned onto alternative platforms, however a legitimate RFID should at least resist unauthorised changes to its critical data/ID; so it cannot be used to assume another identity. The TACS phones were meant to share this property, however in practice it proved simple to re-program them as “clones”. Whilst lack of encryption was a system design issue that resulted in eavesdropping of signalling and phone calls, the weak handset security was a major factor in facilitating cloning.

Therefore when the European Technical Standards Institute (ETSI) [2] standardised its digital phone system (GSM) [3], it decided to remove the reliance on the handset as a security component and introduce a more secure solution based on a SIM smart card. There was still some reliance on phone based security controls in the form of network locks, so that a subsidised mobile could not be moved from one network to another without appropriate authorisation. The fact that even today, so many “unauthorised” parties offer to unlock handsets does little to suggest that mobile based security control is the correct approach. However, technology and application requirements have advanced dramatically and so it is reasonable to suggest that the whole issue of security and trust be reconsidered from a complete system perspective.

2.1 The SIM in Operation

As the role and implementation of the SIM is core to the issues discussed within this paper it is first necessary to obtain a basic understanding of its contribution to the system security. The early SIMs were regarded as a combination of smart card and application i.e. there was no logical decoupling between the SIM Application and the smart card hardware/platform. More modern cards are based on the Universal Integrated Circuit Card (UICC) smart card platform and so the SIM is regarded more as an application rather than part of the underlying hardware. However, in real SIM products, the SIM Application is not necessarily a complete abstraction from the hardware, but rather a “special-case” application, exploiting lower level hardware and software functionality within the platform for both efficiency and resistance to attacks on its security. The SIM stores quite a lot of information and supports a range of functionality, but for clarity we will just focus on three critical components that are stored within the SIM and also in the mobile network's Authentication Centre (AuC).

- An Identity (International Mobile Subscriber Identity - IMSI)
- An Algorithm (A3/A8)
- A Secret Key (Ki)

When a ME is switched on and attempts to access a mobile network there is a signalling exchange that transmits the SIM's identity (IMSI) to the AuC. The AuC issues a random number challenge (RAND) to the SIM (using the phone as a dumb pipe) and the SIM uses its authentication algorithm (A3) and secret key (Ki) to compute a result (SRES), which is sent back to the network. The process is illustrated in Figure 1. The AuC (which also knows the Ki and algorithm) carries out a similar operation and if the SIM and network results match, the SIM is legitimate and the ME is allowed to use the network. In parallel with this, the SIM encryption key algorithm (A8) produces a session key (Kc) to cipher/scramble the communications between the ME and network. For performance reasons, the encryption is actually performed in the ME using

algorithms such as A5/1 and A5/2. Unfortunately, this goes against the philosophy of not trusting the ME as we would expect the algorithms to be vulnerable to attack, however mobile network operators minimise the risks to their systems (and brands) by ensuring that the session key has no long-term significance and is refreshed regularly. It is very important to note that whilst the SIM functionality for authentication and key generation is well standardised, the algorithms (A3/A8) are not and so every network could have its own proprietary algorithms. Furthermore because the AuC can identify a SIM from its IMSI, it can not only obtain its secret K_i , but also determine the associated algorithm. There are numerous proprietary algorithms in use and most designs are kept strictly secret.

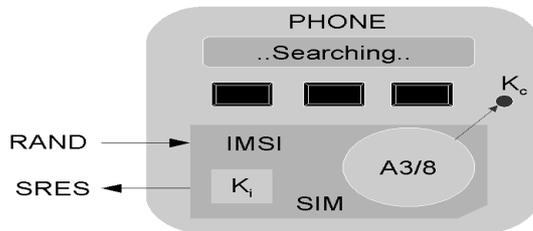


Figure 1: SIM Authentication and Session Key generation

Regardless of the particular algorithm used in the above procedure there are some security weaknesses. The procedure only checks that the SIM is legitimate and not the network, nor does it check that the network challenge is new/fresh. Therefore, the USIM used in 3G networks [4] has an improved exchange that allows the USIM to also test the legitimacy of the messages/challenges received from the network. An example set of algorithms, the *milena* set [5], which may be used for authentication and key generation has also been published. The milena algorithm set was designed by the ETSI Sage group [6] and is based on the Advanced Encryption Standard (AES) [7]. Whilst it is expected that milena will be quite widely used, proprietary algorithms are also likely. A complete description of 3G/UMTS authentication is beyond the scope of this report, however the important changes with respect to 2G/GSM authentication can be appreciated with respect to Figure 2.

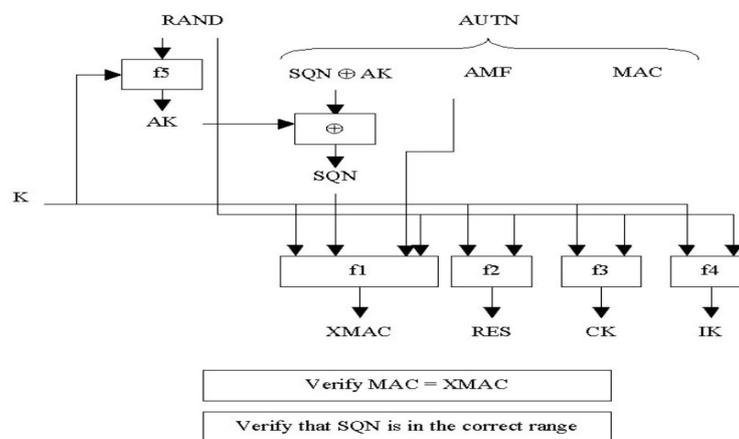


Figure 2: USIM Authentication Overview

There are some changes in nomenclature⁵ and field sizes, but the main design difference is that accompanying the RAND challenge is an authentication token (AUTN). The token includes a Message Authentication Code (MAC) that can be checked by the USIM to determine that the challenge came from the authentic network. It also includes a sequence number (SQN) that can be used to thwart replay attacks and an anonymity key (AK) that can disguise the sequence number within the legitimate challenge. Other enhancements that are beyond the scope of detailed discussion in this paper, include the generation of an Integrity Key (IK) and an authentication management field (AMF). For any new system, 3G/UMTS authentication would be strongly recommended instead of the 2G/GSM method.

⁵ K, CK and RES are similar to K_i , K_c and SRES in GSM authentication

3 Security Analysis

The SIM has been at the heart of the technical measures that underpin the fundamental client-side mobile communication security. However, technology alone does not provide the required system security assurance and there are critical relationships between entities plus associated operational procedures that are necessary to establish and maintain trust. As technology has advanced and the role of mobile communications has expanded, the security requirements go beyond the need to simply safeguard the access to communication bearers. This added complexity not only impacts on technical solutions, but affects the relationship between operational entities. We can begin to examine the roles of these entities with respect to client side trust by considering a range of categories for cellular usage.

3.1 Categories of Cellular Usage

There are many ways that a modern mobile communications system can be exploited and this has led to a number of cellular usage scenarios, supported by a wide range of ME capabilities plus application and service environments. The following categorisation of cellular usage will be used for further discussions.

- Conventional cellular phones (voice/SMS/SIM Toolkit)
 - This category is self explanatory, although we include within it the ability to support SIM toolkit applications i.e. simple and usually menu based applications that are hosted on the SIM itself.
- PDAs and Smart phones
 - These are examples of multi-application MEs, which are more likely than the SIM to host advanced value added applications. These MEs are also regarded as internet-connected, with all the service advantages and potential security risks that this entails.
- PCs and Laptops
 - Computers tend to have a range of communication/connectivity options including ADSL and WiFi as well as cellular. It is therefore most likely that the cellular ME is used as an added modem rather than the computer's application usage being strongly dependent on the MNO. The cellular modem could take several forms including external data cards, USB dongles or an embedded capability of the PC/Laptop.
- Telematics and Metering
 - There is a wide range of possible telematics and metering applications, but considering vehicle telematics or electricity metering, one could imagine a fixed function application using a custom ME reporting to a central application server.
- General Machine to Machine cellular
 - This category is very diverse, but there is an implication that a custom and managed application exists at each “machine” and that the cellular network is used as a logical communications pipe between them. The applications could be hosted in the SIM, ME or a connected device/computer.

Due to their similarities, telemetry and machine to machine MEs will be referred to collectively as “T/M2M” devices. Similarly PDAs and Smart phones will be referred to as “PDA/Smart” devices.

Having introduced an expanded set of cellular communications categories it is now appropriate to introduce some generic definitions for the roles that one would typically find involved in and around a mobile communications system and/or a cellular enabled computer.

3.2 The Roles in Communication Solutions

The following roles would be commonly recognised as present in mobile communications solutions, business and service provision activities. However, it is important to realise that for the SIM and network security viewpoint, we are most concerned about the hacker/attacker who would seek to undermine the system. There is no shortage of potential attackers ranging from academics developing proof-of-concept exploits to organised criminals seeking clones, counterfeit products and sensitive information for financial gain. Similarly, with over 2 billion MEs and SIM cards in circulation and active use, there is no difficulty in

finding devices to attack or transactions to eavesdrop. Attack methods are described in detail within Section 5 and so for now we will focus on the more conventional roles in the mobile communication usage categories.

- Customer/User
 - A typical individual customer is likely to make use of communications services plus value added services that are accessible via his equipment. He is quite likely to change ME whilst keeping the same customer account; normally by transferring his SIM card. He may also sell his old phone and possibly include his pre-pay SIM. He may also change to a new MNO and perhaps keep his old ME. He may register for value added services either with the MNO or a third party service provider. A MNO enabled service might result in a SIM toolkit application being downloaded/enabled on the SIM. If he has a smarter ME he may download applications from the MNO, ME manufacturer or third party service provider.
 - The customer in a metering telemetry environment is likely to be a company such as an electricity supplier. This may require the use of a client-side application in the SIM, ME or computer and a network based application server. Given the likely cost of custom deployment, upgrading the ME is unlikely, but changing MNO en masse is possible for commercial and service performance reasons.
- Mobile Network Operator
 - The owner of the mobile network, the servers and all its deployed SIM cards. Its primary business is to charge customers for use of its network, however virtually all MNOs operate in the value added service domain. They may provide application servers and also corresponding client-side applications for SIMs and MEs. They are able to operate remote management servers for both SIMs and MEs. The large influential MNOs issue customised handsets to their customers.
- SIM Card Manufacturer/Supplier
 - The SIM card manufacturer provides the physical SIM cards to the MNO specification and usually the associated trust services such as initialisation and personalisation. SIM card manufacturers may also see themselves in an application development role and in the provision of operational trust services.
- Mobile Equipment (ME) Manufacturer/Supplier
 - Aside from manufacturing and supplying MEs, the ME manufacturer may see itself in the role of a value added application/service provider. As most suppliers have some means to manage and control the data and applications on their products they may also see themselves in a security or trust services role.
- SIM Client Application Provider
 - A developer/provider of an application hosted on the SIM.
- ME Client Application Provider
 - A developer/provider of an application hosted on the ME.
- ME/SIM Service Provider
 - A provider of a network hosted services matched to applications hosted on the ME or SIM.
- PC Service Provider
 - A provider of a network hosted service matched to applications hosted on a conventional computer.
- PC/OS Manufacturer/Supplier
 - Aside from supplying the computer platform, these suppliers may also see themselves in the role of application/service provider. As most suppliers have proprietary means to manage and control the data/applications on their MEs they may also act in security or trust services roles.

The interaction between the various entities and roles is complex and as there are many MNOs, SIM card manufacturers and ME manufacturers, one cannot cover all possible combinations in this report. However, Figure 3 is an attempt to show the general interaction of the various entities/roles with respect to client-side trust for various cellular usage categories.

An initial observation is that there has always been a very strong trust and security relationship (A) between the MNO, the SIM manufacturer and the users (or at least their SIM cards). This has underpinned every category of GSM cellular usage to date.

In early phone systems there was relatively little complexity as handsets were only capable of basic functionality. As SIM capabilities improved it was possible to have value added services that were implemented as SIM Toolkit (STK) [8] applications. This implied a relationship between the SIM service provider and the MNO (B) who had management control of the SIM card contents and functionality.

There was also a relationship between the STK application developer (SIM application provider) and the SIM card manufacturers (sometimes the same entity) in order to implement the application on delivered SIMs. The reliance on the SIM manufacturer to load the applications has weakened over time due to the introduction of Java Card technology [9] that allows remote loading (or more likely post-manufacture loading) of additional applications.

With the introduction of PDA/Smart devices (case C), ME service providers and ME client application developers/providers work closely together (or are the same entity) to offer services that are less influenced by the MNO. In this scenario the need to establish some level of client-side trust and management may be leveraged from the ME manufacturer e.g. Nokia signing of applications for download to Symbian platforms. This may also serve to strengthen the ME manufacturer relationship with the end user in the area of value added applications and services. The ME manufacturer may not be totally free to control ME trust services as MNOs are major customers for MEs and can in some cases insist on customisations and controls on ME functionality and configuration.

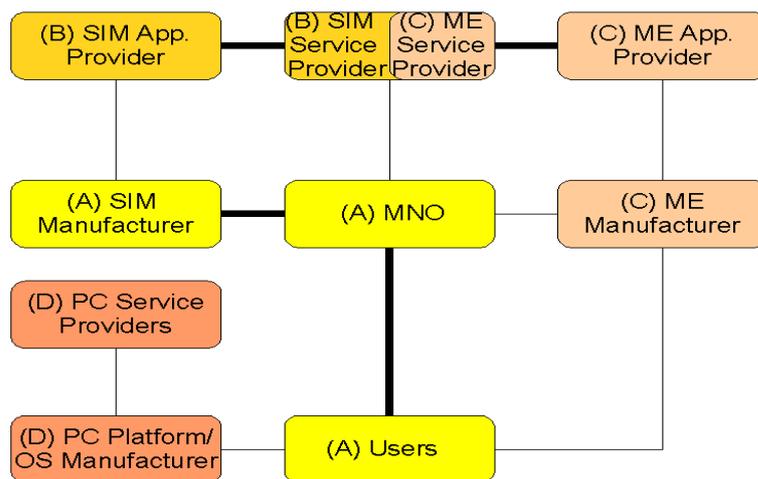


Figure 3: Interaction of Roles in Mobile Service Provision

In the case of cellular enabling PCs/laptops there is much less influence from the MNO and ME manufacturer on client-side trust unless it is established by business/contractual means. The reason is simply that a laptop/PC user may normally work connecting to ADSL/WiFi and so cellular communications is often regarded as another connectivity channel rather than something that would dominate/control service trust and usage.

Telemetry services could in principle fit into any of the options shown in Figure 3, although option (C) is the most likely. The reason is that a fixed function service and custom client application will probably require a customised ME, especially if the accuracy and integrity of the reported measurements are critical. Machine to machine applications would probably suit options (C) and (D) with more emphasis on the deployed applications/equipment and less influence and control from the MNO.

4 Security Fundamentals

So far we have considered trust and security in fairly general terms, but at this stage it is necessary to define some trust items that we will examine further in our cellular usage scenarios. Firstly we will introduce some information security fundamentals;

- Authentication
 - To ensure that entities involved in our trusted solution are legitimate/authentic.
- Confidentiality
 - Information, signals, commands or functionality that are restricted to certain authorised entities must be protected from disclosure/discovery by unauthorised entities.
- Integrity
 - Critical data and applications code should be protected from modification when in storage, operation or during communications/transactions.

These fundamentals can in turn be underpinned by some practical capabilities;

- Cryptographic algorithm(s) plus supporting data for authentication/encryption/integrity
- Secure storage and verification of critical data, with strict access controls
- Secure verification and execution of algorithm(s) and other critical functions
- Secure communication protocols
- Controlled operating environment and isolated “security domains”

The word “Secure” has been used rather freely in the above bullet points and so we should be clear what it means in this context;

The functionality that embodies our security fundamentals has been correctly designed, implemented and tested to strongly resist the anticipated attacks⁶ that may be made against it.

The anticipated attacks will not only be against the design of the functionality, but also against its implementation. This latter point means that the level of security is inextricably linked to a hardware platform and its defensive capabilities. Having established some security fundamentals and supporting practical capabilities, the next step is to consider how they are used in critical operational processes.

4.1 Trust Operations

The concepts described thus far need to be applied to SIM usage and evolution and so it is necessary to consider some relevant processes and operations that rely on security and trust. These entries will be split into two sections; Core SIM Operations, that are fundamental to all forms of mobile communication and Extended Operations that relate to value added services. Although this paper focuses mainly on the Core SIM Operations it is important to be aware of Extended Operations, because if the SIM's capabilities are not sufficient or easily accessible then alternative strategies and technology/management solutions will likely appear.

4.1.1 Core SIM Operations

- Initialisation, Personalisation and Key Management
 - Customised configuration of the solution prior to issue to a customer
- Authentication/ Encryption
- Management of SIM Data and Applications
- Migration
 - Change of MNO
 - Change of ME
 - Change of Computer/Peripheral

⁶ The attacks which are described more fully in the Section 5 include all known logical, physical, side-channel and fault classes.

- Change of Algorithm

4.1.2 Extended Operations/Value Added Services Management

- MNO, ME Manufacturer and third party Value Added Services provision/management
- Near Field Communication Management

4.2 Initialisation, Personalisation and Key Management

The actions under this heading are strongly protected under the trust relationship between the MNO and a few selected SIM card manufacturers; these parties are considered mutually authenticated. Initialisation, prepares the SIM for issue to any of the MNO's customers, however it is a significant stage as the MNO will have shared confidential details of its algorithm, services and data supported by the SIM. It is still the case that some MNOs keep their algorithms secret and whilst they may believe them to have been well designed, they are unlikely to disclose them; as it is not uncommon for proprietary algorithms to fail dramatically once exposed to widespread expert scrutiny. Personalisation is the next critical phase as the keys are generated/loaded to permit authentication/encryption. Normally the keys are also loaded to permit future Over The Air (OTA) [10] management of the SIMs by the MNO and in the case of Java Cards the GlobalPlatform [11] key(s) are loaded to permit application loading/deletion. Various PIN codes are also set at this point to strictly control access to the SIM functionality. Clearly, initialisation, personalisation and the associated key generation and management processes are extremely security sensitive operations and reliant on the integrity of the information loaded onto the smart card. These processes are intended only to be carried out in a secure environment by a party adhering to the highest physical, operational and IT security standards.

4.3 Authentication/Encryption

The basic GSM authentication/encryption support via the SIM was described in Section 2. For it to be secure we must be sure of the algorithm integrity; that it has not been modified and cannot be disclosed by anyone with access to the SIM, nor should it be revealed when in operation. The secret key must not be rewritable, or externally readable and may only be accessed by the algorithm with the condition that the key value cannot be inferred from the operation of the algorithm. To achieve this under anticipated attack conditions requires attention to design and implementation aspects in both software and hardware. Normally, these safeguards are provided by the SIM Card manufacturer (assisted by the specialist chip provider) who would normally also be the algorithm implementor. The session key generated by the SIM and passed to the phone to support confidentiality of radio transmissions is not a strong or long-term secret, but the SIM and phone should ensure that it cannot easily be extracted by an adversary.

4.4 Management of SIM Data and Application

The SIM contains numerous data files [12]. The integrity of the data critically affects the operation of the ME and the services/facilities offered to the customer. The data falls into a number of categories, listed below (and with some examples in brackets);

- System configuration (identity, SMS switching server, language preference, service table)
- User data (telephone numbers, SMS messages)
- Operational data (network lists, fixed dial numbers, customer-care telephone numbers, branding)
- Application data (menus, smart roaming tables, Value Added Services data)

The files have authentication access controls that were set during the initialisation/personalisation processes. Access permissions are controlled locally by a range of PIN codes. The most significant PINs are only known to the MNO (and SIM manufacturer) whereas PIN1 is trusted to the customer for locking the SIM. It is also possible to remotely access the files using an OTA server. This relies on a standardised protocol and another set of keys that have been pre-stored in the SIM during personalisation. The OTA server and keys are normally only available to the MNO, however in principle, some keys may be trusted to other parties for the management of application data on the SIM. SIM Toolkit applications may also be managed in this manner

and the keys permit a secure/confidential channel to be established between the SIM application and the OTA server, independent and additional to any security offered by the basic GSM encryption. In the case of a Java Card supporting Global Platform, it is possible to download/delete SIM applications in the form of Java applets. For this purpose there is also a Card Manager key that is pre-stored on the SIM during personalisation. Clearly the management capability for a modern SIM relies on a significant number of secret keys and PINs.

4.5 Migration

Migration is an important consideration in mobile networks and there are a number of scenarios that deserve consideration.

- Migration to a new MNO is usually achieved simply by SIM replacement. This ensures that the stored data, algorithms, keys, PINS and added functions are exactly as required for the new MNO. It also ensures that the SIM has the correct capabilities in terms of memory, speed, crypto-coprocessor and communications capabilities. There is usually quite a wide range, with some MNOs going for very low-cost limited devices whereas others opt for the more sophisticated and expensive products. Variation may also be evident within SIMs for different customer segments used by a single MNO. There are also significant geographic variations driven by local market conditions and competition. Replacing the complete SIM card also provides assurance that security critical functionality and storage plus the associated attack resistant countermeasures in the underlying hardware and software have been implemented and tested to the MNO's standards.
- Migration to a new ME was one of the reasons that the SIM was introduced as a removable module. Providing that the new ME is not locked to a competing network, plugging in the customer SIM is all that is required. Networks are often obliged by regulation to unlock handsets (for a fee) when requested to do so by a customer, although most are unlocked by unauthorised means.
- Migration to a new PC/Laptop usually means swapping the whole ME + SIM, in the form of a PCMCIA or USB communications adaptor. A new driver may need to be loaded onto the computer. If the laptop has a built in ME then migration would normally be via a SIM and driver swap.
- Although less common, migration of algorithm is possible within mobile networks and indeed, multiple authentication algorithms can be simultaneously supported within the network so that not all SIMs need to be replaced at the same time. At the client level, the migration could be achieved by a SIM swap, however, some networks implement back-up algorithms. These can be switched in, if the normal algorithm is compromised by some unexpected weakness or advance in attack techniques. Because of its emergency-use nature, the back-up algorithm is likely to be a confidential/proprietary design as is the authorisation method to enable it.

4.6 Extended Operations/Value Added Service Management

For value added service management, there are a number of scenarios to consider. Firstly an application may be hosted on the SIM or in the ME. SIM applications are normally controlled by the MNO and tend to be restricted by SIM technology and varying levels of standards support in MEs. On the positive side they benefit from mature SIM standards for authentication, confidentiality, integrity and management. ME applications can benefit from the much greater available processing and user interface resources and potentially may be managed by MNOs, ME Manufacturers and third party Application/Service providers. The deployment/business strategies of these parties may be divergent and the security assurance, control and management requirements can vary for each application.

4.6.1 NFC Management

The prospect of MEs having Near Field Communication (NFC) capability and being able to emulate contactless smart card readers and contactless smart cards is exciting and opens up new opportunities for mobile services. NFC capable MEs introduce the concept of a Security Element (SE) that might be used to host applications in a secure manner. The ownership and control of the SE and indeed whether it is an additional chip or added SIM functionality is still the subject of some debate. It is also feasible to drive the NFC functionality direct from the ME platform, thereby bypassing an additional or SIM hosted SE.

5 Generic Attacks on Smart Cards

Before considering whether the SIM card is any more at risk today compared to when it was launched, it is first necessary to review the range of attacks that have been successfully attempted against smart card solutions. It is a common characteristic of smart card based systems that a critical part of the security solution is in the hand of the user and indeed often in the hands of millions of users. The opportunity for misuse/tampering is enormous and not only from criminals/hackers, but also from the legitimate holders and employees involved with card distribution, sales and operations. It is therefore comforting to note that the most advanced SIM cards are extremely capable of defending themselves, using a sophisticated arsenal of countermeasures against known attacks. It might be imagined that systems would only deploy these state-of-the-art smart cards, however the need to deploy large numbers of cards that often have a limited lifetime, creates pressure on costs. What is actually deployed is the best business compromise, balancing price, risk, functionality and attack resistance. Advances in technology, new attacks and/or increasing the value of protected assets can all upset this balance and so it is prudent to periodically re-evaluate a card's vulnerability to attack. Smart card attacks may be grouped within some general categories that typify the techniques and also the resources and expertise needed by the attackers. Commonly described categories are;

- Logical
- Physical
- Side Channel
- Fault⁷

5.1 Logical Attacks

Logical attacks are not unique to smart cards and are similar to attempts to hack into IT systems. The attack uses the normal interface to the device, but attempts to extract information via repeated guesses, invalid requests and stimulating error conditions etc. Logical attacks exploit things that have been done badly e.g.

- Weak design
- Bad implementation
- Poor testing
- Lack of monitoring/detection

Applied to a smart card, the logical attack only needs a card reader and a PC so there is no cost and/or equipment barrier to such an approach. The basic message formats and protocols used to communicate with smart cards are well standardised [13] and so the attacker can easily construct test messages of correct or intentionally incorrect formats and lengths to try and generate some revealing response.

Weak design is the biggest worry, as it affects not only the information that might be revealed to an attacker, but also its subsequent value and desirability. A great prize would be a global secret held on a legitimate smart card, which if revealed could be used for mass production of counterfeit cards or simply exposure to blackmail. A flawed security algorithm could also be exploited to extract card-specific secret keys, as in a well-reported logical attack [14] against an example GSM authentication algorithm called COMP128-1. Essentially, the attack involved repeated calls (many thousands) to the card authentication command, which eventually revealed sufficient information to allow the secret key (Ki) to be extracted. This should have been practically impossible from a brute-force/trial-and-error standpoint and so the fact that the attack could succeed in a practical time-frame is related to a weakness in the algorithm itself. There were also system design weaknesses in that the authentication command could be called from a PC device that had no provable authority and originally there was no monitoring or detection mechanism for the many authentication attempts. Stronger GSM algorithms exist which are less vulnerable and 3G systems overcome the mutual authentication problem by using a MAC to validate authentication requests. Where COMP128-1 cards are still deployed today they tend to incorporate an authentication counter that terminates the card after too many attempts - albeit with a corresponding reduction in the normal life of the card.

⁷ Fault attacks could be considered as combinations of other categories, but their importance merits separate mention.

It is worth mentioning that a very simple countermeasure is present in most SIMs, which can block logical attacks. The PIN1 lock (see 4.4) prevents unauthorised access of most SIM data unless the user knows the PIN code and a retry mechanism prevents PIN guessing. Because the functionality is simple and underpinned by attack resistant hardware, it is possible for it to be rigorously tested and protected against modification. If the legitimate owner is also the attacker then this measure is undermined, although the most common problem is that users find PIN entry inconvenient and disable the feature when they can.

The best countermeasure against logical attack is quite simple, just design, implement and test things in a rigorous and best-practice manner. For good measure it is also advisable to include monitoring methods to detect and react to attacks in progress. The good news is that modern smart cards bought from reputable manufacturers are generally well designed and tested so that an attacker would be very fortunate to succeed with a logical attack alone. In response to this, more sophisticated attack methods have evolved including physical, side-channel and fault attacks.

5.2 Physical Attacks

It is a false assumption to consider a system secure because it uses a good algorithm. The algorithm alone may provide good logical security, but could be completely undermined by physical tampering. In the case of the smart card there is reliance on the chip to defend itself against intrusive physical attacks that seek direct access to, or modification of memories, buses, CPUs etc., thereby bypassing "logical" security defences.

Whereas logical attacks can be developed or copied by just about anyone, physical attacks normally require a high level of technical expertise and access to sophisticated and expensive laboratory equipment. Almost invariably the attacks require the decapsulation of the smart card chip i.e. its removal from the card in preparation for physical examination and/or modification.

Rendering the source card (or indeed many of them) unusable is not an important issue as physical attacks are most often used for reverse engineering the card implementation. The goal is to discover design information that could be used in some other type of attack or lead directly to the creation of card clones. Although physical probing techniques are tricky because the chip area is so tiny (typically less than 10mm²), the principle is quite simple i.e. attach conductive probes to interesting parts of the circuit. One of the favoured tools for investigating smart card chips was the probe station consisting of a microscope, a high precision mechanical platform and needle-like probes designed to make electrical contact with parts of the chip circuitry. A second-hand probe station can be bought for about €12,000; an example is shown in Fig 4.



Figure 4 Probe Station (Wentworth labs)

As chip technology has advanced, the smart card circuitry has got smaller and smaller, to the point where conventional probe stations are becoming impractical as the target contacts are very small compared to the size of the probe needles. A more modern and expensive tool (approx. €350,000) is the Focussed Ion Beam (FIB). The FIB can be used to examine and/or modify a circuit, and can connect additional circuitry or

simply larger contact pads for access via a conventional probe station.

A physical attack against a chip that has not been specially designed to resist it, will reveal its secrets quite rapidly, however smart cards can incorporate many countermeasure techniques which whilst not guaranteeing protection against physical attack, make it far more difficult for the attacker. A common countermeasure is to introduce a physical barrier that may simply be a slab of tough material or an active current carrying mesh. Removing either type of barrier without destroying the chip requires the type of expensive equipment and high-level expertise normally found in commercial test labs. This is not the end of the defences available to smart cards and indeed the circuit layout and layers may be scrambled to make it difficult to locate the desired attack points. Even if the barriers and layout confusion are eventually mastered there can still be low level encryption methods that prevent direct reading of buses and memory contents. Whilst the attack/investigation is in progress there are also the environmental detectors to worry about. If the chip is exposed to light, extremes of temperature and/or voltage it will trigger a detector and the chip will cease to operate.

For a tiny piece of silicon, the smart card can be astonishingly good at resisting physical attack. However, one must always expect that any security device is only tamper-resistant and not tamper proof as driven by enough motivation, expertise, money and time, a physical attack will succeed. The important question is why would anyone take the trouble? A company-commercial reason is that reverse engineering may permit the creation of counterfeit smart cards. A security reason is the belief that the card contains some secret information or technique that may directly or in combination with other attacks, be used to exploit and or clone other such cards. It should also be appreciated that some researchers would embark on a sophisticated physical attack simply because of the academic challenge. Whilst proof-of-concept attacks are not usually motivated by financial gain, they unfortunately serve to educate other parties that may have malicious intent.

It is worth re-iterating that not all smart cards include all the physical attack countermeasures mentioned above and manufacturers are understandably coy about revealing what their chips will or will not do. It is also worth noting that there have been successful physical attacks that have been very simple and low-cost [15] such as interrupting the power supply during critical processes. Generally, physical attacks are beyond the resources and capabilities of most attackers, whereas the same is not true of side channel attacks.

5.3 Side Channel Attacks

It is very difficult to keep a secret. Usually this statement might apply to some document, secret algorithm or key that is locked away - however it can be applied to the operation of electronic circuits. If a circuit such as a smart card chip is believed to be running an algorithm and using a secret key it may become an attack target. The logical attacker will actively try and trick the chip into revealing its secrets (but should fail), whereas the physical attacker will try and break in and perhaps destroy a few other chips as part of a learning process. What makes the side-channel attacker interesting is that he basically just waits and listens, extracts the secret with a modest amount of equipment and often leaves the original card undamaged. As an analogy, consider the attackers trying to get access to an interesting security lecture. The logical attacker tries to trick his way in, but can't get past the security guard on the door, the physical attacker drives a bulldozer through the wall, whereas the side-channel attacker just listens at the door where the sound of the lecturer's voice "leaks" through the door (the side channel).

The basic principle of side channel analysis is that secret information is always leaking despite the presence of logical and physical attack countermeasures, so the trick is to find the leakage and extract something useful from it. The smart card, in common with most other electronic devices, consists of many logic gates and transistors. An example is shown in FIG 5. As the logic state of the gate changes e.g. from 1 to 0 or 0 to 1 there is a minute surge of electrical current accompanied by a spike of electromagnetic radiation. Therefore, if you can detect the current or radiation changes you can get an idea about the state of the logic gates. Now if for example those gates are part of a register used by the security algorithm then at some stage, the transitions will be caused by the value of the key. To extract the key from the leakage may sound

difficult, but the reality is that until countermeasures were put in place, side channel attacks were effective against even high-end smart cards and they are still valid against other types of non-protected circuitry.

One of the major concerns is that side-channel attacks do not need a lot of expensive equipment and can be recreated with far less expertise than would be needed by a physical attacker. A typical set up for Simple Power Analysis [16] or Differential Power Analysis [17] requires little more than a digital oscilloscope and PC whereas the electromagnetic emission version simply requires an added antenna/amplifier.

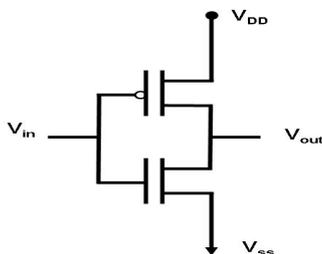


Figure 5: CMOS Switching Circuit

DES was an early target algorithm [18] as it uses multiple rounds of processing in which only a few key bits are used. The side channel analysis is used to reveal key information at each round so the attacker deals with a sequence of small key problems rather than the full DES key. It is of course known that with the right equipment, normal DES can be brute forced [19], so a common strategy is to recommend double or triple key DES. Whilst this is a logical defence it does little to safeguard against a side-channel attack that can simply work its way through the small number of key variations of each round. The critical point to remember is that logically strong algorithms and keys may still be vulnerable to side channel attack unless appropriate tamper-resistant measures are implemented.

Modern high-end smart cards are strongly tamper-resistant and this extends to defences against side channel attack. As attack analysis usually relies on statistical averaging, one approach is to add low-level timing jitter and high-level software variations to prevent alignment of the leakage waveforms. Another defence is to generate artificial noise on the leakage signals to disguise the target information. If the chip hardware is sophisticated enough to support differential switching (rare) then whenever a logic 1 changes to 0 there is a corresponding bit that changes a 0 to a 1, which snuffs out the leakage signal at source.

5.4 Fault Attacks

The classical principle behind a fault attack is that if you can induce a temporary fault into an algorithm the error can be used to reveal secret information such as a key. A necessary pre-requisite is therefore the means to introduce a fault in a non-destructive and controlled manner. This has been practically achieved using spikes on the card power supply and momentary exposure to bright light. The faults can be induced in the processing circuitry or possibly change the state of memory cells. A famous example of this attack class was used against a RSA implementation [20]. Because RSA is computationally intensive it is often implemented in two stages with the results being combined by the Chinese Remainder Theorem (CRT). If you can cause any error in the processing of a bit within one of the stages there is an elegant method that directly extracts the secret key by a simple mathematical calculation.

The ability to introduce a temporary fault in an algorithm is therefore of concern, however it is not the worse thing that can happen. If we recall from Section 4 that “*the level of security is inextricably linked to a hardware platform and its defensive capabilities*”, then the fault might affect the operating system, variables or system state. This type of fault is not necessarily temporary and could have far reaching effects.

If a smart card device is likely to be subject to fault attacks then there are a number of countermeasures that can be used effectively. At the chip level it is possible to add detectors that should trap the fault insertion

attempt e.g. light or voltage glitch detectors. Special purpose operating systems can include checks and redundancy mechanisms to resist such attacks and at the application level the algorithm could be run multiple times and the results compared (although this can have a significant effect on speed and usability).

5.5 Summary and Main Points

It is important to realise that the attacks mentioned in this chapter are not made against smart cards because their security is weak, but rather that they are identifiable and easily accessible security components. Except for very low-end devices, the SIM smart card's resistance to attack is quite admirable. When it was first introduced the SIM's attack resistance provided far more assurance than the available mobile phones, however we should not automatically assume that this is still the case. It is therefore worthwhile to consider various alternative options for SIM implementations with respect to our modern categories of cellular usage.

6 SIM Implementation Options

In this section we will consider a range of implementation options for the SIM with respect to our core security requirements and various categories of cellular usage described in Section 3.

6.1 Pure Software SIM

Our definition of a Pure Software SIM (PSSIM) is a SIM application written in software running on a shared computer processing platform that cannot benefit from any hardware security component. Even before considering the detailed issues it should be clear to most readers that this can only reduce rather than increase existing levels of security protection and so it is important to try and understand why the PSSIM is proposed for some market segments.

6.1.1 Motivation

The root of the interest in the PSSIM lies in cost reduction and logistics. Although typical SIM cards are no longer very expensive (few Euros) the mobile industry deals with very large numbers of them and so even small reductions can result in saving of millions of Euros. With a PSSIM, it is not anticipated that there would be a significant saving to the ME manufacturer (apart from the SIM socket cost) although it may be a little easier to physically construct the product without access for SIM card insertion/removal. The beneficiary is therefore most likely to be the MNO (who typically buys/owns the SIM cards), but of course the MNO has most to lose from compromising the system security. The financial motivation is not only cost savings from smart card chips, but also the removal (or great simplification) of the procurement and distribution channels required to get SIMs from the factories and into mobile devices. The problem is worse for specialist MEs that are effectively embedded modems used in Telemetry/Machine-to-Machine (T/M2M) applications. For example, you may ship cars around Europe that have embedded communications to support local services and telematics and so ideally you would insert the SIM for the country of destination. The destination may not be known during manufacture so adding the SIM cards is a manual post-fit operation that may be costly and difficult in terms of physical access. There are some potential technical and tariffing solutions to this problem, however they put more emphasis on strong security and it would be difficult to sell T/M2M equipment without full co-operation from an international MNO. In the case of general T/M2M the operational environmental conditions may be more extreme than for conventional mobile telephony and so a “special” SIM card may be needed in any case.

6.1.2 Analysis

It is relatively straightforward to write a software program that provides all the core functionality offered by a SIM. If the algorithm, key management and security policies are made available by the MNO then the implementation will also include the main security functions. If our program has been designed, written and tested according to IT best practice then it should resist logical attacks, but alarmingly, none of the other categories described in Section 5. Furthermore because the PSSIM runs on a non-secured shared platform we have additional security concerns. Other applications running on the platform may gain access to the memory used by the PSSIM and use this to monitor/modify critical data and functionality. The other platform applications may also cause (intentionally or accidentally) run-time performance and resource problems for the PSSIM that may compromise the SIM's real-time duties. Furthermore, if the platform has no boot protection there is no guarantee of integrity for any of the platform's operating system and application software. As the platform is also a communications device there are fears of fast-spreading remote attacks that typify PCs, such as viruses, worms and trojans. A successful attack on a PSSIM would seem almost inevitable and that would normally lead to clone devices.

Some MNOs, ill-advisedly, appear prepared to tolerate a level of cloning in their networks – as evidence from continued use of weak authentication algorithms. However, this is dangerous, as attacks can rapidly spiral out of control as particularly easy/lucrative techniques are discovered. PSSIMs would not only simplify attacks, but also provide convenient clone platforms for use with extracted keys. An attack on a

PSSIM using a secret algorithm would be most severe as it would amount to the discovery of a “global secret” which may reveal weaknesses that could undermine all categories of cellular communication.

Aside from discarding almost all of the security measures that have protected the SIM application, the PSSIM also breaks the very important trust linkage between the MNO and the SIM manufacturer. For the PSSIM the MNO may have to share its algorithms, data, policies, secret keys and PINs with ME manufacturers. Whilst SIM manufacturers are set-up as high security companies (and often generate much of the sensitive data for MNOs) this is not normally the role/capability of the ME manufacturer.

It is difficult to have any confidence in the personalisation and lifecycle management (including migration) of the PSSIM and its associated data. For example, personalisation of a conventional SIM is normally carried out in a very secure environment (usually by the SIM manufacturer) and relying on the integrity of the smart card platform. If we are using other parties in a different environment, to configure a platform that cannot guarantee its integrity, or resist simple attacks, we could completely compromise the core security solution.

The situation appears worse if we consider Over The Air management of the PSSIM. Normally the SIM securely stores a number of sensitive OTA keys that are used for the remote management/download of SIM data and applications. Copies of these keys should only be held by the MNO and/or application providers as they can drastically change the functionality of the SIM and its applications. Compromising the OTA keys and/or the OTA functionality, risks the nightmare scenario of a rapidly spreading remote attack or virus.

6.1.3 PSSIM Summary

Despite the temptation to reduce costs and simplify logistics, it is strongly recommended that a PSSIM solution should not be used. It is vulnerable to a wide range of security attacks and undermines proven trust relationships and management processes. PSSIMs would almost certainly result in clones and problems may become fast-spreading once attacks exploit the communications capability of MEs. The use of PSSIMs in some restricted categories of cellular usage is dangerous because revealing secret information such as proprietary algorithms and security policies may become a risk for all categories of usage. This risk is not just from technical attack, but disclosure of this sensitive information to third parties that are not specialist security/trust companies.

6.2 Hardware Shared Security Software SIM Solution (HS-SSIM)

The SIM is not the only application in an ME that may require security assurance and following-on from the discussions on the PSSIM, it is clear that some specialist hardware support is essential. The hardware must protect the particular application, its operational processes and sensitive data, but to do this it must also protect the integrity of the processing platform itself. The conventional SIM card and TPM (see Section 7) approaches are examples of additional and specialist hardware security processor modules that may help protect applications from attack. However, this section focuses on the use of the existing/main ME processor to implement the SIM, which we will refer to as a hardware shared software SIM (HS-SSIM). This differs from the PSSIM because the chosen ME processor has some specialist security enhancements. By way of example, this section will consider the “TrustZone” [21] from Advanced RISC Machines Limited (ARM).

The ARM is a 32-bit RISC processor architecture that is used in a wide range of embedded systems such as mobile phones and PDAs. The ARM architecture is complemented by a number of security extensions under the overall name of “TrustZone”. Key components of the TrustZone architecture are presented and summarised/quoted below:

- a TrustZone CPU that is used to run trusted applications isolated from normal applications, and to access the memory space reserved for trusted applications,
- secure on-chip boot ROM to configure the system,
- on-chip non-volatile or one-time programmable memory for storing device or master keys,
- secure on-chip RAM used to store and run trusted code such as Data Rights Management (DRM) engines and payment agents, or to store sensitive data, such as encryption keys,

- other resources, peripherals, that can be configured to allow access by trusted applications only.

It offers two separate and parallel execution environments that run on the same processor. This is achieved through virtualisation which is responsible for deciding whether the currently executing application should run as normal or secure code. The decision, controls the restrictions (e.g. which peripherals it can access) and privileges (e.g. access to certain memory locations) that apply to the application. As there is only one physical processor, a very close cooperation is required between the hardware and software components, in order to guarantee that the overall architecture behaves as a single well defined system. The measures used to control the execution environment should make TrustZone more secure than the PSSIM solution, but only if the TrustZone software maintains its operational integrity and the underlying hardware functions correctly even when subject to attack.

A core element of the TrustZone functionality is the “Secure Monitor” entity which is responsible for performing the necessary checks and switches between secure and non-secure states. It is up to the processor to enforce the correct data and peripheral access policies depending on the actual execution state, e.g. secure or non-secure. The TrustZone platform also offers a secure boot process using cryptographically signed boot-strap code in ROM. The TrustZone software (TZSW) is at the centre of the platform’s execution environment. It offers the ability to execute native code or interpreted applications. This interpreter is based on the Small Terminal Interoperability Platform (STIP) which is developed by GlobalPlatform [11] and in theory permits applications to run in a protected and sandboxed environment.

6.2.1 Motivation

The motivation to take this approach is similar to that of the PSSIM and for logistics reasons it is most desirable for the T/M2M categories of cellular usage. From a cost minimisation viewpoint it suggests improved security compared with the PSSIM whilst avoiding the cost of an additional security module. There may also be interest in using the method of software upgrades (re-flashing) to also correct/upgrade the SIM functionality.

6.2.2 Analysis

The HS-SSIM is a compromise solution that should offer improved security with respect to a PSSIM, however it may be difficult to achieve tangible assurance that this is actually the case. We are reliant on the HS-SSIM software to control secure/non-secure processing and reliant on the underlying hardware to implement the necessary execution and data storage controls. Unless these components have been evaluated to a known standard (e.g. Common Criteria) or lab-tested by a MNO, one must assume that these components may be vulnerable. The difficulty in trying to achieve a level of assurance is further compounded if software elements may be upgraded (or re-flashed) post-issue, as is often the case in a ME. This upgrade mechanism potentially provides an added security risk and the changes to the software could invalidate previous security evaluations. The hardware could be evaluated against known attacks although resistance is usually a combination of hardware, operating system and application measures. However, the design of the processor is likely to be a compromise between the performance needed for normal use and the measures needed for secure execution. It might be expected that this compromise would lead to less protection and slower execution compared with a dedicated hardware security module using a similar core processor. The literature review reveals that the core of the TrustZone's functionality is not currently promoted (due to its increased size) for smart card products. In fact the core functionality of TrustZone provides secure access to smart cards, suggesting that the technologies are expected to co-exist. Although all software may run in the secure core of the platform, it is advised that only security sensitive code is shielded for trusted execution as the extra checks increase the overall size and complexity of the platform software.

The last point is significant as one of the main motivations of the HS-SSIM was to provide added security without adding an extra chip. If as the TrustZone suggests, there may be added cost for the main processor and its memories and increased power consumption, the benefits of the HS-SSIM compared to adding a specialist hardware security module are eroded. The benefits are not necessarily lost entirely, as an extra

standalone “chip” would probably require more power than the TrustZone increase. Furthermore, flash memory protected by TrustZone should be cheaper than a SIM EEPROM, however if protected flash is judged sufficiently attack-resistant then it could also be used in the dedicated hardware security module.

6.2.3 HS-SSIM Summary

The HS-SSIM should in practice be more secure than the PSSIM because it has some logical features that are aimed to support secure execution. However, from a security assurance viewpoint there may be little difference unless the hardware, platform operating/system environment are evaluated to a known standard and then do not change post-issue. Without this assurance one has to assume that whilst the solution may resist some logical attacks, it will be vulnerable to other techniques. It is also questionable whether adding complexity to the main processor is the best solution and indeed whether it is possible to achieve effective assurance on a device that is shared by non-secured applications. The justification to use the HS-SSIM was to avoid an extra chip and the associated cost, however adding secure execution capability to the main processor appears not without its own costs.

6.3 Standalone HW security SIM solution

Our definition of a Standalone Hardware Security SIM (SH-SIM) is the conventional SIM smart card used in GSM and UMTS communications. It relies on a specialised security microcontroller chip with a security optimised operating system that has an implementation of at least the SIM application and associated (MNO specific) algorithms and data. It is assumed that the whole platform and SIM application have been tested as resistant to all known attacks, to a level of assurance that satisfies the standards/requirements of individual MNOs. Furthermore, the SH-SIM is personalised in a secure environment prior to issue to customers and/or insertion in a ME. In key management terms, security critical keys (and PINs) have been pre-loaded onto the SIM during personalisation and these fields will not be changed thereafter. Post-issue re-personalisation is not possible and migration to a new network requires physical replacement of the SH-SIM. Changing to a new ME simply requires moving the SH-SIM to the new ME. A SH-SIM from a particular MNO might support algorithm migration, but the new algorithm and switching mechanism would have been pre-loaded onto the SH-SIM.

6.3.1 Motivation

The motivation to keep the existing SH-SIM is that it has done a remarkably good job of securing communication in mobile networks. Changing anything to do with the SH-SIM is therefore a risk that could have serious impact on a MNO's business and reputation. Moving from a well-proven security solution to an even better-proven security solution of course has merit. For example, upgrading from a 2G authentication method to the 3G milenage method is well advised as the introduction of mutual authentication and an improved open algorithm design would strengthen the system security.

6.3.2 Analysis

Although there are some arguments to avoid the SH-SIM to reduce costs there are also arguments to keep it. For example, the size of SH-SIM devices varies enormously; whilst a typical device might have a 32kbyte EEPROM, some MNOs are using Mbyte devices and highly advanced Gbyte SIM⁸ devices are available [22]. If MNOs no longer supplied their own physical SIMs, but made use of a Hardware Security Module (HSM) built into a ME, the HSM would either be dimensioned very large (and expensive) to accommodate all MNOs, or would be too small for some. Advanced MNOs might want a HSM with USB interfaces [23] and Single-Wire-Protocol [24] capability to enable Near Field Communication [25] services, however other MNOs may not wish to pay for this. There are also cost factors around crypto-coprocessors that are necessary to support extra services that use public key cryptography. It is therefore important to realise that by providing its own SH-SIM the MNO can always ensure that the cost and capability of the SH-SIM matches its current and planned business requirements.

⁸ It is interesting to note that the traditional limitations of the SIM smart card (small memory, slow interface and restricted CPU) have been overcome by technology advances, albeit at added cost compared to a traditional SIM.

In the SH-SIM the MNO is providing the complete package for its security. It can ensure that the combination of the chip, operating system, applications are not only functionally correct, but have been adequately tested against known attacks. There is sometimes criticism of SH-SIMs in that they claim to be tamper-resistant, but are not evaluated to common criteria standards and so the level of security offered to support cellular usage is unclear. The reason for this is partly due to cost, but also the rapid deployment rates in mobile communication that can mean a SH-SIM lifecycle is shorter than the time needed for a common criteria evaluation. However, while there remains no agreed and practical assurance level for a SH-SIM, it will be far more difficult (than it should be) to argue its superior security credentials compared to say a PSSIM. It would therefore be very useful to have MNOs adhere to a set of industry-wide best-practice criteria for SH-SIM devices (perhaps from the GSM Association [26]) .

Any post-issue changes to the SH-SIM are completely controlled by the MNO and tend to affect data and added applications rather than the core SIM security and the OS/platform. The MNO can also ensure that all critical functionality, keys and data are personalised in a secure environment pre-issue, often by means of a highly trusted third party (e.g. SIM manufacturer).

It should be noted that all the good security properties of the SH-SIM arise from the chip and the associated operational and management processes. The smart card body has almost no useful role in normal operation and most of the card plastic is discarded before insertion into the ME. The body may help in production as standard smart card production machines may be used, which helps to keep costs down. The body also helps by providing portability which allows a SH-SIM to be swapped between MEs.

In principle, if the SH-SIM chip was prepared and personalised as if it were to go into a card body, but was actually soldered onto a ME circuit board and connected via its normal interface then the operational security would be identical to the current SH-SIM arrangement, except that the SH-SIM would not be removable. This could be a problem for conventional telephony as there can be legal requirements that allow a customer to migrate his ME to an alternative MNO, however this might not be the case for T/M2M communications. Embedding the SH-SIM would result in a fixed network ME that would need to be replaced if there was any problem with the ME or the SH-SIM chip/account. Migration might still be possible if it were feasible to place the SH-SIM chip in a small socket. In this case the MNO migration would be a job for a technician. This requirement may not however be unreasonable for T/M2M systems.

6.3.3 SH-SIM Summary

The conventional SH-SIM has proven itself as an effective security module for mobile communications and any change is a potential risk that must be properly understood and justified. One positive example of this is the SH-SIM migration from 2G to 3G security, as this has been rigorously investigated and improves system security with the addition of mutual authentication and an improved open algorithm. The large variation in the types and costs of smart cards used by different MNOs for various market systems suggests that any SH-SIM equivalent provided by the ME risks being over or under specified with corresponding impact on costs and services. An advantage of the SH-SIM is that operators supply the whole package including hardware/OS, applications and data and so can control the level of security assurance and ensure resistance to all known attacks. It would perhaps be better for the industry if more effort was directed towards best-practice criteria for SH-SIMs so comparisons with say PSSIM solutions could be more easily made in future. The conventional preparation and use of the SH-SIM ensures that personalisation is carried out by a trusted party in a secure environment and that the SH-SIM may be subsequently managed by the MNO in a secure manner. The importance of these last points is often overlooked when proposing alternative solutions to the SH-SIM. If a SH-SIM chip is prepared in the normal manner then in principle it could be included on a ME PCB for T/M2M communications and provide the same security as the removable SH-SIM. The disadvantage is that MNO migration may not be possible or require technician services.

7 Trusted Platform

It should be clear from the fore-going discussions that the security of mobile communications has relied to a significant extent on a specialised tamper-resistant microcontroller embedded within the SIM smart card module. Because this small computer platform has been specified, implemented and tested by the MNO and/or its suppliers, the MNO can trust it. Furthermore, as the SIM is owned and managed by the MNO this trust is not eroded during the SIM lifecycle. However, the need for manageable trust in a processing platform is not restricted to mobile communications. The PC world, via the Trusted Computing Group [27]⁹, has driven forward the Trusted Platform Module (TPM) concept. The main motivation behind the development of the TPM specifications was the difficulty in a modern/complex PC to verify whether it is running “uncorrupted” software. If for example the underlying operating system (e.g. Windows or Linux) cannot provide such assurance, there can be little confidence that any applications will perform correctly. Fundamentally, the TPM goal is to prove that the PC is in a state that can be trusted to run applications and process data. Note that in contrast with a SIM, the TPM is not intended to be portable or removable, but rather inextricably bound to one and only one platform (and usually in the form of a chip). It securely stores asymmetric keys which can be used in order to protect sensitive data like other keys, certificates and passwords. The TCG states that “Trust is the expectation that a device will behave in a particular manner for a specific purpose” [28]. The TCG specifications state that a TPM device “should provide at least three basic features: protected capabilities, integrity measurement and integrity reporting.”

- Protected capabilities include specific commands and functionality that hold exclusive permission to specific “shielded” platform resources. These resources could for example be memory locations or registers (for holding sensitive data, keys and integrity measurements) and also management of cryptographic primitives and other keys. The protected capabilities play a crucial role in verifying the correct operation of the platform.
- The TPM should be in a position to reliably collect information that reflects the TPM host's software state (i.e. integrity measurements) and place it in the relevant integrity storage locations. This information will be used for subsequent verification.
- Platform attestation enables the integrity measurements (which reflect the TPM host's software state) to be reliably reported. This then enables a verifier to decide how much trust should be placed in the status of the platform.

7.1 Roots of Trust

To a great extent the required levels of assurance are achieved through the TPM's roots of trust.

- Root of Trust for Measurement (RTM)
- Root of Trust for Storage (RTS)
- Root of Trust for Reporting (RTR)

The RTM is a trusted entity that can generate a reliable integrity measurement for at least one process running on the underlying platform. The TPM specifications define the RTM as “the root in the chain of transitive trust” [28]. The RTM is typically implemented as the normal platform engine controlled by a particular instruction set (the so-called ‘Core Root of Trust for Measurement’ (CRTM)). On a PC, the CRTM may be contained within the BIOS or the BIOS Boot Block (BBB), and is executed by the platform when it is acting as the RTM [42]. The CRTM is the first component to be executed during an authenticated boot process (see Section 7.2).

The RTS is a trusted component that is responsible for providing confidentiality and integrity for stored TPM data, e.g. cryptographic keys and the Platform Configuration Registers (PCRs) used for storing integrity measurements.

⁹ Note the Trusted Computing Platform Alliance (TCPA) was the predecessor to the TCG.

The RTR is a trusted entity responsible for providing various integrity measurements (integrity digests) on information stored in the RTS. Secure storage involves both encryption and sealing. Sealed data is bound to a set of platform measurements (i.e. that define the state of the platform) that must be present in order for data to be decrypted. This will provide the necessary reassurance that the platform will obtain access to “sensitive information”, i.e. the sealed message, only if it exists in a known configuration.

7.2 Authenticated Boot and Secure Storage

During the platform boot process the CRTM and TPM enable each of the components in the system (including hardware and software) to be reliably measured and the resultant measurements to be stored in a set of Platform Configuration Registers (PCR) located in the TPM. The integrity measurements simply involve a SHA-1 digest of the code to be loaded. The TPM is agnostic to the underlying operating systems (OS) or applications and provides no assurance regarding their inherent security. Verification simply reports pre-runtime configuration information and it is up to the OS, applications and external verifiers to judge the trustworthiness of the PC platform and permitted actions.

Providing secure storage functionality requires that the TPM should be tamper resistance and preferably be able to detect and report any tampering attempts. At the same time, secure storage requires the provision of cryptographic functionality (for the integrity and confidentiality) of data. For example, the concept can be further expanded to include the association of keys and data with passwords and/or metrics that depict the platform/software state). The TPM specifications make specific references to the required cryptographic functionality that should be offered, e.g. RSA [29], SHA-1 [30, 31], random number generation [32]. A detailed review of TPM [33] is beyond the scope of this paper, but critical components can be seen in Fig 6.

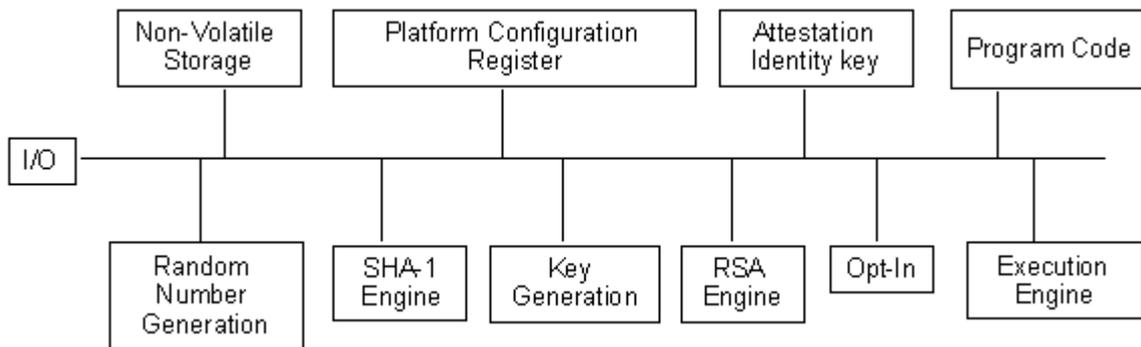


Figure 6: TPM Component Architecture [28]

7.3 Ownership

The concept of TPM ownership is very important. Originally, some concerns were voiced as TPMs may have permitted a third party such as an equipment provider or OS supplier taking control of a PC that was bought/owned by a user. The response was to put the user in control, by requiring the user to “opt-in” and take ownership of the device. The TPM may come with some pre-installed keys and certificates (which can be changed) that will provide the foundation for various platform operations. Opting-in means that the user is guided through a number of steps, e.g. generating further cryptographic keys, and formalising and defining the behaviour of the TPM.

Having, very briefly, examined the core functionality of TPM, it is becoming evident that in principle it could also be utilised by a range of applications requiring secure storage, enhanced authentication and additional platform/application security. Some of the proposed TPM usage scenarios are summarised in [34, 35, 36]. For a general purpose security component to be considered trustworthy by application providers it must clearly prove its security credentials and capabilities. The TCG proposes that the ISO-15408 [37]

Common Criteria security evaluation standard should be utilised, in order to evaluate and certify TCG products and platforms. The TCG is working towards the development of common criteria protection profiles (PP) and guidelines that will assist the security evaluation of TPM devices. Although the TPM device is embedded into the PC motherboard, it could also be used in a variety of other devices¹⁰ such as PDAs and mobile phones. In the next section, we examine the TCG's efforts to apply the TPM to mobile communications through the Mobile Trusted Platform (MTM) specification.

7.4 Mobile Trusted Platform (MTP)

Among the main missions of the TCG is the provision of trusted platform specifications for a number of devices (including MEs) that require information security assurance. The Mobile Trusted Module (MTM) [38] specification and the corresponding MTM reference architecture specifications were published in September 2006 and 2007 respectively. The TCG's definition of trusted computing (in terms of making sure that hardware and software behave as "expected") remains the driving force for the development of the MTM specifications. One of the quoted aims is to complement the existing functionality (of the operating system, hardware/software, SIM/USIM) with additional assurance that will enhance the overall security of the ME.

In common with the TPM, the MTM is responsible for protecting keys and other confidential information in secure storage. It might not be implemented on chip [39] and could even be a virtualisation layer making use of a TPM. The MTM defines two main types of MTM ownership, i.e. the Mobile Local-Owner Trusted Module (MLTM), which is similar to the TPM model and the Mobile Remote-Owner Trusted Module (MRTM). The latter (MRTM) is a significant departure from the user controlled "opt-in" approach and allows remote entities (such as service providers, communication carriers, device manufacturers) to access and control their own space within the platform. In order to maintain clear boundaries between the entities and the operations that are allowed, the concept of "engines" is introduced. Each engine corresponds to an aforementioned owner and it has exclusive control over its data protection mechanisms. Some engines are considered mandatory (critical services) and other discretionary. The specifications state that mandatory engines are owned and controlled by a MRTM whereas the discretionary engines may be owned by a MLTM. There is also the role of Device Owner, which determines the remaining engines in the mandatory domain and determines all engines in the discretionary domain.

The significance of the change in ownership should not be overlooked. The user acceptance for TPM in the PC world is established through the "opt-in" process, however MRTM could see more management control in the hands of third parties. In the case of SIM card control, this has been justified by the MNOs remaining the legal owners, however the ME is normally bought/owned by the user.

Similar to the TPM, the MTM should be able to clearly demonstrate adequate levels of resistance to a wide range of attacks (see Section 5). However, the exact details of the fundamental threats, countermeasures and the planned Common criteria Protection Profile (PP) have not yet been fully finalised. Achieving an acceptable level of trust may prove more difficult for a MTM that uses a software abstraction layer.

7.4.1 MTP Use Cases

A relatively complete list of MTM use case scenarios is defined in [40]. In principle, the MTM can be implemented in many devices with relatively restricted processing capabilities including mobile phones and PDAs. Any application (e.g. Digital Rights Management, ticketing, payment, network access) requiring a level of platform integrity could in principle benefit from the underlying MTM specifications.

The MTM use cases will be influenced by the following three MTM characteristics. Firstly, the specification introduces the secure boot of an MTM engine. This implies that on top of the standard functionality of data protection and platform attestation, the engine must boot into a pre-defined software state or not at all.

¹⁰ <https://www.trustedcomputinggroup.org/specs/>

Secondly, the MTM engine should have the ability to perform runtime checks on the integrity of software components. Finally, the local and remote TPMs essentially implement subsets of the TPM functionality. The TCG proposes these characteristics for MEs, but not for any other type of platform.

7.4.2 Comparison of MTP and UICC

The TCG developed the MTM specifications, in order to address specific needs of the mobile security. The fact that both the TPM and MTM standards are open and widely available encourages to some extent the acceptance of the technology and enables expert peer reviews. However, acceptance will not be sustainable in the long term if the technology does not deliver against its promises or any serious issues or flaws are identified. Fundamentally the underlying hardware (which is the cornerstone of the security assurance) and software specifications must not only be properly defined, but correctly implemented by all suppliers of compatible MEs. Therefore, it remains to be seen whether all real ME products will be able to demonstrate, (through evaluation and operation), long term sustainability against security threats and vulnerabilities. The latter is of particular importance, as many problem issues are discovered only after products are deployed.

In almost all the above issues smart card technology is delivering its aims. This is partly because it is a complete customised and tested package (e.g. chip/OS/application), but also because its functionality is relatively restricted and simply because it has proved itself over many years of practical use. Therefore, smart card design, operational and security requirements are well understood by chip developers, smart card manufacturers and MNOs..

The only reference known to the authors at of the time of writing this article, attempting to categorise the appropriate uses of smart cards and TPM (in terms of machine or user authentication scenarios) is presented in [41] and summarised in Table 1.

User/Machine Authentication Scenarios	Smart card	Trusted Platform Module (TPM)
User ID for virtual private network (VPN) access	Yes	No
User ID for domain logon	Yes	No
User ID for building access	Yes	No
User ID for secured e-mail	Yes	No
Host computer ID for VPN access	No	Yes
Host computer ID for domain access	No	Yes
Host computer ID for attestation (that is, authentication of software applications)	No	Yes

Table 1. Suitable Uses of TPM and Smart Cards[41]

However, although the authors have not verified this analysis¹¹, it does suggest that the two technologies can be considered as complementary. Although the table is not directly aimed at mobile communications there is a split between authentication of the user and the computer/machine. This split is not just related to the security properties of the smart card or TPM, but also the fact that the user account, security and data tends to be personalised separately in a secure environment, independent of the PC/ME that is eventually used.

¹¹ And it is conceivable that the smart card could be used for all scenarios

8 Summary

The role of the traditional SIM smart card has been described as a tamper-resistant module for supporting authentication and encryption in mobile networks. The capability to strongly resist known security attacks is fundamental to its existence and this is underpinned by both the design and implementation of the hardware, operating system and application software. The SIM memory size and crypto functionality for different MNOs (and market sectors) can vary significantly and cost sensitivity means there is no such thing as “one-size-fits-all”. The SIMs are normally pre-personalised in a secure environment before issue to customers and the detailed requirements and related security processes tend to be MNO specific. It is common for there to be a strong trust relationship between the MNO and the SIM manufacturer as the supplier is trusted with and often generates, much of the most sensitive security data associated with SIMs. The personalisation of the SIM is not only vital to normal operation, but also the lifecycle management of the SIM. The authority to make changes to the SIM after issue is underpinned by an ownership model whereby the MNO always retains ownership of the SIM card and facilitated by management keys pre-stored on the SIM during personalisation. The removable nature of the SIM smart card means that a number of migration scenarios, which are sometimes a legal requirement, are supported. A user may migrate to a new ME, but keep an old SIM card, or migrate to a new MNO by swapping that network's SIM card. Occasionally it is possible to migrate to a new account type or even to a back-up security algorithm without changing the SIM card.

The SIM card has proven itself, over a long period of time, to be capable of maintaining the security of mobile communications and so it would take some very compelling reasons to take the considerable risk of changing to an alternative solution. One of the reasons might be to reduce the cost of SIM card supply, meaning not just the card itself, but the whole logistics, storage and distribution process. Another reason is that some newer cellular usage categories may have physical, distribution and environmental constraints that make it awkward to accommodate the normally personalised SIM card.

The two most interesting new cellular categories to consider are PDA/Smart and T/M2M. The former is an example of a high-end/expensive device, becoming virtually a connected and portable PC. Whilst the MNO is in control of communications security and may influence ME features and servers, the ME itself is often under some form of control from the ME manufacturer and perhaps more open to application developers independent of the MNO's influence. The T/M2M is a good contrast as it is not issued to “real users”, may have fixed/custom functionality requiring a specialised ME and is likely to be very cost sensitive.

One might suggest for either of these categories that a Pure Software SIM (PSSIM) could be used, as the logical implementation of a SIM application on most processors would be possible. However, this would be extremely foolhardy as the implementation would be trivial to attack, risking exposure of not only secret keys, but details of confidential information and possibly secret algorithms. Clones would be simple to arrange even on the attacked MEs, dragging security levels down to those of old analog systems. It is therefore strongly recommended that a PSSIM is not considered under any circumstances, but rather solutions that can leverage from some kind of tamper-resistant hardware.

The T/M2M MEs might be considered as candidates for the Hardware Shared Software SIM (HS-SSIM), in order to save the cost of a new hardware security module that may be difficult to insert/remove in practice. Intuitively the HS-SSIM should offer more security than PSSIM, however it may be difficult to provide sufficient assurance to MNOs on security, and also on vendor independence. Normally the MNOs provide the whole SIM “package” i.e. an appropriately sized chip, OS, application and personalised data all tested against known attacks and competitively sourced from multiple suppliers. In the case of the HS-SSIM, the MNO has to fit onto a proprietary third party platform that is a compromise design for supporting both normal and secure processing environments. There is likely to be reliance on critical software to handle the split between operation modes and it is difficult to see how the same levels of attack resistance could be achieved compared to a normal SIM card and indeed some product literature seems to deny support for smart card emulation. A formalised security evaluation would go along way to provide assurance, however it is not

clear if it is at all feasible to even consider this on a general purpose platform, especially if the processor code may need upgrading to support normal operation. The effort and time required for evaluations would probably not be merited and incompatible with the short lifecycle of processor chips used in ME production.

As an alternative to the HS-SSIM the T/M2M could make use of an existing TPM chip. However a T/M2M type of ME is likely to be of limited functionality and so would not be expected to already have a TPM. If one considers an important function of a TPM to be the support for an authenticated/secure boot, a T/M2M device may be installed and permanently powered and so may only boot once in its lifetime. Control over distribution and installation may obviate the need for any special control over the boot function.

It is difficult to identify a cost/logistics argument for investing in an embedded TPM for a T/M2M, rather than a conventional SIM smart card, however it is conceivable that one might exist. The TPM is normally designed as a hardware security module and so in principle is capable of proving its security capabilities in a formal (or widely accepted) evaluation. However, this may not be the case if the MTM version is not fully implemented in the chip, but uses ME proprietary software to interface with a TPM. To progress the general argument further we will assume that sufficient assurance might one day prove possible and that the underlying TPM provides an appropriately dimensioned secure storage and execution environment that could be used for a SIM equivalent. The first observation to make is that the TPM would need to be dimensioned to accommodate the demands of the most demanding MNO and indirectly this may have a cost impact on the less demanding MNOs. A second and far more fundamental observation is that the normal processes and methods for secure SIM card personalisation could not be used. If personalisation is to be carried out in a secure environment then the MNO might need to either use ME manufacturers/suppliers (who are not usually security/trust specialists) or ship the MEs to a trusted party (such as a SIM manufacturer). The risk of the first option and cost of the second would seem to undermine the case for adding a TPM for cost saving and logistics simplification. Personalisation in a non-secure environment would be likely to contradict many MNOs' security policies and would not be recommended due to the risk of disclosing secret keys, sensitive data and secret algorithms/functionality. There is also the question of access to the initial management keys to personalise the platform and the fundamental question of ownership and management rights of a device that is deemed to be owned by the customer. Most migration scenarios would not be supported unless re-personalisation is possible, which MNOs may regard as a major security risk and there is an open question about the transfer of keys and management rights between competing parties.

The PDA/Smart could well have a TPM/MTM installed by default and so the chip cost justification would not be an issue, however the significant problems around personalisation, ownership, management and migration would remain. The TPM could be seen as a complementary technology to the SIM card that could ensure the integrity of the PDA/Smart platform. This would be particularly useful as PDA/Smart devices would likely have internet connectivity and require protection from common perils such viruses and trojans etc. The TPM/MTM would be a useful resource for application providers whose security requirements are less demanding as MNOs and might otherwise offer solutions on completed untrusted platforms. This may move some control from the MNOs towards the owner of the TPM. Who the owner should be is a matter for business debate rather than security analysis.

It would appear that the arguments for providing an alternative SIM implementation on the grounds of reducing costs and simplifying distribution logistics are not clearly compelling for any type of cellular usage. However, there is still the unsolved problem of coping with physical and environmental constraints (particularly in the T/M2M case) that would better suit an embedded chip. One solution is for the SIM suppliers to provide personalised chips, but not the smart card body. These would maintain SIM card functionality and security, but could be embedded directly onto a ME's PCB by a ME manufacturer. This could provide better temperature and physical characteristics and avoid the need for the SIM socket. Aside from the extra controls, management and cost at the ME manufacturer, the MEs would be permanently customised to a particular MNO, which might fragment a manufacturer's stock and production capability. Once deployed, the MEs would be bound to a single MNO/SIM and this lack of migration may not be acceptable to customers or MNOs in the long run. A variant would be to put the SIM chip in a socket (much

smaller than a SIM card socket) so that a common batch of MEs could be subsequently configured for various MNOs. Migration would then be possible, but would probably require technician support which is not necessarily out of the question for T/M2M deployments. In theory one could use the personalised SIM chip approach in any phone although the added migration difficulties suggest that the removable SIM smart card is still generally the more convenient option.

8.1 Value Added Service Management

The situation for value added service management is less clear than the core SIM functionality and there are a number of different scenarios that need to be addressed in a standardised and consistent manner.

- For the MNO, the attraction of a SIM hosted value added service is that the SIM's capabilities and the functionality that it offers are completely specified and controlled by the MNO. This enables the MNO to confidently offer a range of user services as well as background MNO management functionality such as smart roaming support. A disadvantage is that the types of user services, have to date been fairly limited compared to the possibilities offered by modern MEs. Attempts to improve the sophistication of SIM applications have often been frustrated by limited support of the relevant standards by MEs. The MNO can decide to offer SIM security capabilities to ME hosted applications by means of APIs, however considering the range of ME platform types and the many models in use (of varying levels of standards compliance) it is difficult to achieve a ubiquitous solution.
- The ME manufacturer sees the application management problem from another angle. The manufacturer will know its ME models intimately, however there may be little interest in having an "open" application capable of running on MEs from competing manufacturers. If attempting to secure these applications through the SIM, the manufacturer might require a dialogue with all MNOs that use its MEs, to confirm that essential functionality is available. In fact it is very unlikely that all MNOs could guarantee this support as SIM product capabilities are quite variable and so the manufacturer may conclude that the most reliable and predictable strategy is to rely on the capabilities of the ME alone. For example, a developer can currently create an application and send it to Nokia for signature (based on a developer registration process), that is then checked when the application is loaded onto the ME. An obvious disadvantage of this approach is that the personalised and tamper-resistant storage and functional properties of the SIM card are not exploited to secure the applications.
- A third party service provider may have security requirements, however they may not share the ME manufacturer's confidence in the ME platform (and ideally would want the application to run on all MEs). The SIM is promoted by MNOs as a secure platform, however without proof of a best-practice/independent security evaluation, it may be difficult to demonstrate adequate assurance levels to a third party. This problem could in principle be overcome, but there would also need to be more consistency/standardisation regarding the provision of SIM resources and third party access to them. For example a Java Card based SIM could host an added application, but not if the MNO policy was to prevent third party access. Alternatively the MNO could welcome third party access, but then deploys low cost SIMs that have no spare memory to fit extra applications.
- The prospect of MEs having NFC capability and being able to emulate contact-less smart card readers and contactless smart cards has great potential for innovative new services. Some of these services will relate to financial transactions or the exchange of sensitive data and so security support is critical. In current NFC trials the MEs have a separate Security Element (SE) that can be regarded as a Java capable smart card chip embedded in the ME. This again raises difficulties around personalisation and assurance. If a bank provided the SE then it may be satisfied with the security level and control/management of the chip, however what if many banks or other types of application provider all insist on their own security levels and management methods? The MNOs are not keen on the separate SE chip and propose controlling the NFC functionality via the SIM card using Single Wire Protocol. This may satisfy MNOs, but may not be very helpful to other parties that wish to secure their NFC applications and services.

None of the value added service security approaches described above do an ideal job. A possible way forward would be to use the SIM in a limited way that exploits its personalisation, key storage, remote management and security algorithms, without making great demands for memory or processing resources. An application developer would want to be assured that this support would be available from most SIMs and there should be no non-technical barriers (strategy/policy) that would prevent him using it. The applications themselves could sit in the ME providing there was an execution environment that was reasonably secured. If this environment was underpinned by some ME hardware then from an assurance viewpoint the application provider would want a standardised and evaluated chip(s) that is used in most MEs, rather than a range of proprietary solutions.

8.2 Concluding Remarks

From a security and practical viewpoint, there seems little wrong with the long standing practice of using SIM applications implemented as easily replaceable smart card devices, to underpin core security requirements. The alternative approaches that were considered do not clearly or convincingly reduce costs or simplify logistics and come with great risk of compromising current and proven levels of security and assurance. The most promising alternative device is the TPM (and the MTM variant) as it may become a standard feature on many MEs and can in principle, provide some tangible (evaluated) level of security. However, the TPM seems most suited as a complementary solution, rather than a replacement for the SIM. The TPM is well suited to protecting the general integrity of the ME platform and adding some security support to value added applications, particularly in high-end PDA/Smart MEs. An interesting further study would be to determine how the SIM, TPM and other ME security resources could best be used in combination.

There is some justification for doing away with the removable smart card in T/M2M applications due to physical/environmental restrictions. The suggested solution that does not compromise security and still provides means for migration (albeit with technician support), is to provide SIMs in the form of personalised chips for insertion into small sockets on the ME PCBs. It is suggested that SIM and ME manufacturers investigate the practicalities and costs associated with this approach. It is further suggested that this investigation also compares the costs of using an advanced large memory SIM with a correspondingly simplified “dumb” ME.

The secure management of Valued Added Services is not considered to be well handled by any of the individual technologies addressed in this paper. The existing SIM card approach should do more to help in this respect otherwise there will be little choice, but to seek other solutions even if they are less secure, more costly and difficult to manage. An important aspect for application/service providers is to have SIM security capabilities that they can rely on being present, regardless of the particular MNO and ME model. Therefore the requirement is for an industry-wide/standards response rather than ad-hoc solutions from particular MNOs. Specifically, the MNOs could be encouraged to agree an industry-wide best practice guide for the security evaluation of SIM cards or equivalents. For cost and time constraint reasons it may not be practical to carry out independent security evaluations and so it might be sufficient for SIM Vendors to self certify their products against the industry guidelines. Similarly, the MNOs could consider establishing an industry-wide, minimum set of functionality/APIs available to support applications in the ME. To be clear, this would not just be a standardisation exercise, but an initiative to ensure that the SIMs procured by MNOs should always support the minimal APIs. Furthermore, exploitation of the SIM capabilities should not be hindered by complex negotiations with MNOs, but perhaps by a simple registration and signing process, similar to those used for ME application developers.

References

- [1] K. Finkenzeller, RFID Handbook: Radio-Frequency identification fundamentals and applications, Wiley, 1999.
- [2] European Technical Standards Institute (ETSI), <http://www.etsi.org>.
- [3] M. Mouly, M-B Pautet, The GSM System for Mobile Communications, Cell & Sys. Correspondence 1992
- [4] Third Generation Partnership project (3GPP), <http://www.3gpp.org>
- [5] ETSI SAGE Group (originally), 3G Security; Specification of the MILENAGE algorithm set: An example algorithm set for the 3GPP authentication and key generation functions f1, f1*, f2, f3, f4, f5 and f5*; Document 1: General, 3GPP TS 35.205
- [6] Security Algorithms Group of Experts (SAGE), <portal.etsi.org/sage/>
- [7] NIST, Advanced Encryption Standard, FIPS 197, 2001
<http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf>
- [8] 3GPP, Specification of the SIM Application Toolkit for the Subscriber Identity Module - Mobile Equipment (SIM - ME) interface (Release 1999) 3GPP TS 11.14 V8.18.0, 2007-06.
- [9] The Java Card Forum <http://www.JavaCardforum.org/>
- [10] 3GPP, Security mechanisms for the (U)SIM application toolkit; Stage 2 (Release 5) TS 23.048 V5.9.0, 2005-06.
- [11] GlobalPlatform, Global Platform Card Specification, 2006.
- [12] 3GPP, Specification of the Subscriber Identity Module -Mobile Equipment (SIM - ME) interface (Release 1999) TS 11.11 V8.14.0 (2007-06)
- [13] International Standard Organisation, "ISO/IEC 7816, Information technology - Identification cards - Integrated circuit(s) cards with contacts- Part 4 Interindustry commands for interchange", <http://www.iso.org>, 1995.
- [14] David Wagner and Ian Goldberg, "GSM Cloning", ISAAC Berkeley
<http://www.isaac.cs.berkeley.edu/isaac/gsm.html>, 1998.
- [15] Anderson Ross, Kuhn Markus, "Tamper resistance – a cautionary note", second USENIX workshop on electronic Commerce Nov 1996
- [16] Paul Kocher, "Timing Attacks on Implementations of Diffie-Hellman RSA DSS and Other Systems" Advances in Cryptology - CRYPTO '96, LNCS 1109, 104-113, 1996.
- [17] Paul Kocher, Joshua Jaffe and Benjamin Jun, "Differential Power Analysis, Advances in Cryptology - CRYPTO '99, LNCS1666, 388-397, 1999.
- [18] E. Biham, A. Shamir, "Differential Cryptanalysis of DES-like Cryptosystems. Journal of Cryptology", Vol. 4 No. 1, 1991.
- [19] Kumar Sandeep et al, "How to break DES for €8,980", CHES 2006, <http://www.crypto.ruhr-uni-bochum.de>
- [20] Eli Biham, Adi Shamir, " Differential Fault Analysis of Secret Key Cryptosystems", Technicon Computer science dept – Technical report CS0910.revised, 1997
- [21] Tiago Alves and Don Felton. TrustZone: Integrated hardware and software security: Enabling trusted computing in embedded systems. www.arm.com, July 2004.
- [22] Mayes Keith and Markantonakis Konstantinos On the potential of high density smart cards, Elsevier Information Security Technical Report Vol11 No3 2006
- [23] Universal Serial Bus (USB) Forum, <http://www.usb.org>
- [24] ETSI SCP Group , SCP Specifications, <http://docbox.etsi.org/scp/scp/Specs/>
- [25] Near Field Communication (NFC) Forum <http://www.nfc-forum.org/>
- [26] GSM Association, <http://www.gsmworld.com>
- [27] Trusted Computing Group, <http://www.trustedcomputinggroup.org>
- [28] TCG. TCG Specification Architecture Overview. Trusted Computing Group, 1.2 edition, April 2004.
- [29] R. L. Rivest, A. Shamir, and L. M. Adelman. A method for obtaining digital signatures and public key cryptosystems. Technical Report MIT/LCS/TM-82, 1977.
- [30] National Institute of Standards. Secure hash standard. Federal Information Processing Standards (FIPS) 180-1, 1995.
- [31] ISO/IEC. ISO/IEC 10118-3 Information technology – Security techniques – Hash-functions – Part 3:

- Dedicated hash-functions. International Organization for Standardization, <http://www.iso.org>, 2004.
- [32] Berk Sunar, William J. Martin, and Douglas R. Stinson. A provably secure true random number generator with built-in tolerance to active attacks. *IEEE Transactions on Computers*, 56(1):109–119, 2007.
- [33] C J Mitchell, editor. *Trusted Computing*. IEE Press, 2005.
- [34] Roger L. Kay. How to implement trusted computing, a guide to tighter enterprise security. https://www.trustedcomputinggroup.org/news/Industry_Data/Implementing_Truste% d _Computing_RK.pdf. Endpoint Technologies Associates.
- [35] Roger L. Kay. Trusted computing is real and it's here. https://www.trustedcomputinggroup.org/news/Industry_Data/Endpoint_Technologi% es_Associates_TCG_report_Jan_29_2007.pdf.
- [36] Trusted Computing Group. Embedded systems and trusted computing security. https://www.trustedcomputinggroup.org/groups/tpm/embedded_bkgdr_final_sept_1% 4_2005.pdf.
- [37] International Organization for Standardization. ISO/IEC 15408: Information Technology— Security Techniques— Evaluation Criteria for IT Security, 1999.
- [38] Trusted Computer Group, Mobile Trusted Module Specification 1.0. June 2007
- [39] Trusted Computing Group. Mobile Trusted Module Specification faq - general overview. www.trustedcomputinggroup.org, June 2007.
- [40] Mobile Phone Working Group. Use case scenarios v 2.7. https://www.trustedcomputinggroup.org/groups/mobile/Final_use_cases_sept_22_% 2005.pdf, 2005.
- [41] DELL. Securing network-based client computing: User and machine security. Dell's Technology White Papers, 2004.
- [42] Eimear Gallery, "Trusted computing technologies and their use in the provision of high assurance SDR platforms", SDR Technical Conference, Orlando, USA, 13-17 November, 2006.