Contents

Feature Articles: Communications & Information Security

4 Architectural Approaches to Network and Service Access Authentication
Marcin Dabrowski Piotr Pacyna

17 Analysis of Security Vulnerabilities and Countermeasures of Ethernet Passive Optical Network (EPON)
Di Jin, Stamatios V. Kartalopoulos, Pramode K. Verma

30 Analysis on the Resilience of Key Pre-distribution in Sensor Networks
Shen Yulong Ma Jianfeng Pei Qingqi

38 Hierarchical Spectrum Sharing Networks
Jie Chen, Chuan Han, Chulin Liao, and Shaoqian Li

48 Security Measures Against CBRN Threats: Case Study Olympic Games
P. Stavroulakis, Professor

54 Customized Biometric Architecture for Access Control in Stadiums based on Federated Identities
Christos K. Dimitriadis, Peter Stavroulakis, D. Polemi

63 Simulation of Propagation Loss Models and Antenna Designs for Naval Troposcatter and Tropospheric Duct Communications
Nikos J. Farsaris, Prof. Peter P. Stavroulakis

69 Chaos-based Applications in Secure Optical Communications
A. Argyris, A.Bogris and D. Syvridis

79 PIM Interference Analysis under Multi-band Multi-signal Input in Duplex Indoor Distribution System
Pauling Huang, Wang Baohua, Senior Engineer, China Unicom
ABSTRACT

Authentication is the first step, of central importance, for access control and for security protection in radio access networks. A general model for authentication was adopted from fixed networks and applied to the wireless world. However, the differences in the operational environment between the fixed and the wireless world, heterogeneity of the radio communications systems, new trends in service provisioning, emerging business models and performance requirements raise the need to revisit the original requirements for authentication systems and to come up with schemes that better suit current needs. In this review paper we discuss authentication in single-hop radio access networks by characterizing the current as well as the emerging authentication schemes.

Key words: authentication, access control, wireless networks, next-generation Internet

I. INTRODUCTION

In recent years, the wireless access networks have received broadband capabilities and have become available for residential and institutional users. Broadband radio access networks have been under deployment in enterprises, campuses, public institutions, governmental organizations and in public access networks. The success of the radio access networks is attributed to their high usability, flexibility, cost-efficiency and especially to the unrestricted communications capability. Multiple radio access technologies favor competitiveness, accelerate progress in the field and propel the wireless industry.

Diversity has led to heterogeneity, with the side effect of a burden with inter-working between technologies and systems. Many challenges of inter-working are characteristic of a particular network setup, the technologies in use and the services under deployment. Some of them, however, are common to all installations. Specifically, in all the systems the network infrastructure and the services need to be protected against misuse. Authentication is the first step, of central importance, in network and service access control and in security protection. This review paper studies the current and the emerging approaches to authentication in single-hop radio access networks.

II. SCOPE OF AUTHENTICATION

Authentication allows to prove identity of a subject—a user, a system or a device—by verifying credentials which are presented by the subject. The subject is referred to by means of an identifier—some...
naming convention, such as true name, a pseudonym or, in an indirect way, by means of identifier of a device owned by the subject, which allows for the so-called user- or device authentication schemes. The authentication is performed by a verifier in a unidirectional authentication or by both parties when involved in the mutual authentication process. The subject and the verifier may have come across each other before, but they needn’t necessarily have had a relationship at any time before.

Verification of the credentials submitted in the process is conducted with support of trusted entities, such as e.g. authentication servers, which maintain a binding between the credentials and the subject and are in a position to ascertain at any time validity of the binding.

Authentication serves secure bootstrapping of network attachment, as a part of an access control process aimed to prevent service theft by unauthorized individuals, impersonation of legitimate users also known as the identity theft.

Although a general model for authentication in radio access networks was adopted from fixed copper-wire access and cable networks, and next applied to the wireless world, some methods have been devised specifically for the wireless networks. These new methods are evolving due to significant differences of the operational environment between the fixed and the wireless world and due to changing threat models when new forms of attacks are discovered.

**A. New requirements for the wireless environment**

Successful authentication establishes a trust relationship between the subject and the verifier (also known as the authenticator). The trust can be uni- or bi-directional, symmetric or asymmetric depending on the type and strength of the employed authentication method. It is represented by unidirectional security association(s) which define a ciphersuite and cryptographic keying material used for subsequent protection of communications on the wireless link.

New requirements for authentication in the wireless access networks are mainly related to user mobility, which is defined as the ability to change point of attachment in a network. A physical change of the attachment due to a movement can also imply a logical change of location in the network topology or a change of the network operator, known as intra-domain and inter-domain handoff, respectively. A default requirement of almost any mobility scenario is continuous reachability of a mobile user. The ability to sustain previously established sessions in spite of the movement, preferably with little or no impact on session continuity, is another requirement.

In the new context defined by mobility, a few essential technology requirements need to be taken to account with respect to an authentication system:

1) In systems which are supporting mobility, the users are ‘always on’ the network and may own multi-mode capable devices. With such diversity, the authentication systems need to be particularly robust, scaleable and resilient to attacks. Methods of various complexity must be supported and the system must be extensible for the emerging authentication methods.

2) Mobile users prefer to be involved in a few business, relationships, preferably only with an operator in their home area. Multiple business relationships are either troublesome for users and hence disfavoured or, simply unfeasible because of random roaming patterns.

3) The employed authentication system must not overly depend on the mobility management scheme or type of re-addressing used to support mobility.

4) In open access systems, where multi-access technologies are used, the necessity for frequent re-authentication, likely at every handoff, is required. Also, small cell sizes, typical of some wireless technologies, imply short visiting times and quite frequent intra-domain handoffs. Low protocol overhead and low overall latency of the authentication process is an advantage.

5) Delay sensitive application, in particular conversational services, leave little tolerance margins
for re-authentication during handoff. The requirement for voice call continuity favors low-latency re-authentication and make-before-brake authentication (pre-authentication), so that the latency does not add to handover latency.

New business models are also expected to change service provisioning. In particular:

6) Deregulation of the markets allows for market fragmentation. Pre-established security associations between a customer and the serving network will be rare due to continuous roaming of users.

7) Wireless or cellular network service providers with disproportionate geographical coverage will exist in the future, so that inter-domain handoffs will be more frequent than so far. Authentication should be robust also where roaming agreements between the providers do not exist.

8) Authentication schemes need to support cooperative business models based on delegating some network-related functions and sharing the infrastructure by some providers.

9) In a multi-service environment services are offered by the third party providers. In order to facilitate such schemes a combined network and service authentication is encouraged where the network infrastructure and the service providers are separate, but federated entities.

10) Anonymity and pseudonymity for users need to be supported with un-linkability protection between separate authentications. Authentications should remain un-linkable by other users in the system as well as by the visited access network providers. Nevertheless, tractability of users’ activity, e.g. for lawful interception, should be plausible to comply with local legal regulations.

11) Strong identity management systems maintained by the identity providers, which are capable of partial sharing of personally-identifiable information with third parties, are needed, subject to local law.

New capabilities of radio access technologies influence the way radio access systems are designed, built and operated, so the previous methods for authentication become partly incapable to satisfy the new requirements. On the contrary, existence of large install base of authentication systems used so far, calls for incremental upgrades to build upon the existing infrastructure and experience. Today, new requirements for authentication are diffusing the well established methods.

In the following sections we present recent advances in network access authentication. In order to make the text comprehensive we begin with a revision of legacy methods for better presentation of the differences in the new methods.

III. AUTHENTICATION IN WIRELINE NETWORKS

Extensible Authentication Protocol [1] was introduced in dial-up networks to authenticate a terminal before releasing configuration parameters and an IP address to a terminal. EAP is an authentication framework independent on the lower-layer protocols. It employs different authentication methods for different use cases. The EAP protocol was adopted in broadband access technologies ranging from cable modems to xDSL. Today, it is also used in broadband radio access in WLAN, WiMax, 2G and 3G cellular networks.

EAP relies on data-link layer including point-to-point and shared links so there’s no need to deploy additional protocols. It can also be transported over IP or over higher layer protocols, such as RADIUS. Although EAP is involved in the authentication process, it does not specify any particular authentication scheme. Its strength is in simple syntax, lock-step design that allows for simple implementations and in extensibility. EAP supports the old and the emerging authentication methods without the need to upgrade the protocol itself. There is no need to upgrade Authenticators, since these can operate in a pass-through mode essentially by forwarding messages between the subject and the authentication server.

public key cryptography and certificates to establish a secure communication channel between two entities. On the contrary, the EAP-SIM and EAP-AKA are examples of symmetric cryptography among the authentication methods [7], [8].

The TLS Handshake establishes a shared secret between the client and the server for subsequent protection of the communications channel. TLS is supported with EAP-TLS method which wraps handshake messages in EAP Request – EAP Reponse transactions. With the mandatory exchange of both client and server certificates EAP-TLS sends client’s identity and certificates through a non-secured channel and enforces the use of client side certificates which is sometimes inconvenient.

These drawbacks have been addressed in the two-phase Protected EAP (PEAP) method. In the first phase a secure tunnel is established with TLS handshake over EAP. Unlike in EAP-TLS the client-side certificate is optional here and the client uses temporary Network Access Identifier (NAI) instead of the real name. Once the secure TLS tunnel has been established, other EAP methods can be run over the tunnel [3]. The tunnel protects the subsequent methods so the client can release its identity in a secure way. PEAP allows for TLS session renegotiation, so that after the secure channel is in place, the second TLS handshake can proceed assuring that certificates of both involved parties are protected.

EAP-tunneled TLS (TTLS) has much in common with PEAP. The difference is in the second phase where TTLS remains compatible with the RADIUS protocol.

Flexible Authentication via Secure Tunneling (EAP-FAST) is similar to PEAP in setup of a secure channel for use by less strong methods. Protected Access Credential (PAC), a kind of pre-shared secret, is used to establish a secure channel via TLS. Afterwards a weak authentication method can proceed.

All the above methods fit in the general EAP protocol exchange and look alike, but differ in many respects. Among the differentiating factors there is support for device authentication, user authentication or both. The use of symmetric or asymmetric cryptography and the possibility, or actually the necessity to support client-side certificates is another one. The ability to protect identity of a subject and its credentials in the initial phase is important, too. Key derivation and key management strategy is similar, but somewhat different. Complexity of protocol exchange, the resulting signaling overhead and the induced delay in terms of the round-trip time required to complete the process are different, too.

IV. AUTHENTICATION IN WIRELESS NETWORKS

A. Port-based network access authentication in IEEE 802.1X

In the port-based Network Access Authentication (PBNA), originally specified in the IEEE 802.1X for the wired networks, the supplicant and the authenticator exchange messages in order to verify each other’s identity with help of an Authentication Server, and next generate a shared secret for protection of communications on the access link and to open a port on the Authenticator for unrestricted communications [9].

The use of IEEE 802.1X on the IEEE 802.11 wireless links is different depending, to some extent, on whether the IEEE 802.1X authentication occurs before, or after the Mobile Node association. For example, some concepts, such as e.g. controlled/uncontrolled ports do not apply to IEEE 802.1X pre-authentication. The details can be found in [10]. Further considerations in [10] address protected capability negotiation, filters activation, authentication of control and management frames and comments on 802.1X pre-authentication, some of which have been accounted for during specification of new standards, such as, e.g. 802.11r [15].

B. Native methods in IEEE 802.11

In the infrastructure mode of the IEEE 802.11b a Mobile Node (MN) must first associate with an Access Point (AP) in order to be able to communi-
cate with other nodes within the Basic Service Set (BSS). The association involves Open-System Authentication (OSA) and Shared-Key Authentication (SKA) modes. The OSA authentication is a simple two-way handshake. The station sends a request with its MAC address as an identity to the authenticating station or to the Access Point which responds with the success message.

The Shared-Key Authentication mode comprises of a 4-way handshake to make sure that both the requestor and the authenticating entity (usually an AP) possess the same shared key. The requestor first sends its identity. Next, the authenticating entity responds with a random challenge-text to be encrypted using the secret key by the requestor. The encrypted text is sent back to the authenticating entity for decryption using its own copy of the secret key. If the decrypted text is the same as the one sent earlier, the authenticating entity sends the success message to the requestor.

Both methods are focused on authenticating the station (STA) only.

C. Authentication IN IEEE 802.11i

In IEEE 802.11i WLAN networks strong authentication is achieved with Robust Security Network (RSN), better known as WPA2 [14].

The 802.11i process consists of the initial handshake between the MN and the AP for security capability discovery, then 802.1X exchange between the supplicant, the authenticator and the authenticating server for Master Session Key (MSK) and Pairwise Master Key (PMK) derivation, and finally between the Supplicant and the Authenticator again (the 4 way-handshake) for proof-of-occupation of the MSK and derivation of a Pairwise Transient Key (PTK), as shown in Fig. 2. Successful completion of these three handshake procedures results with mutual authentication between MT and the AP as well as a joint possession of the PMK/PTK by the supplicant and by the authenticator.

The standard also allows for static configuration of a Pre-Shared Key (PSK) on the Supplicant and the Authenticator for use as the PMK to avoid the 802.1X exchange: during re-associations a previously cached PMK can be used immediately to reduce authentication latency of new authentications and to reduce the associated computational overhead.

V. AUTHENTICATION IN IEEE 802.16

The IEEE 802.16e-2005 standard introduces a MAC Security Sublayer [16] which defines Security Associations (SAs), Privacy Key Management (PKMv2) protocol, encapsulation protocol and X.509 certificates for Subscriber Station (SS) and Base Station (BS) mutual authentication. The goal of SS authorization is to create the Authorization SA and the associated Authorization Key (AK). The SS has X.509 certificates with the manufacturer’s and the station’s public keys. These are sent to Base Station (BS) during authorization. BS verifies SS’s certificate using manufacturer’s public key, generates the Authorization Key and sends it to the SS encrypted using SS’s public key. The model assumes trust relation with manufacturers who issue certificates for subscriber stations. It further assumes that the SS’s...
private key is protected from compromising, i.e. hardcoded in the hardware.

After a successful authorization the SS and the BS share the AK key which is then used to create traffic Encryption Keys (TEK) as a part of key provisioning. Because the AK and the TEK have finite lifetimes, they must be periodically regenerated. For that purpose, there are two TEKs: the current and the secondary one. Currently there’s a work-in-progress towards extending this model to support mobile multihop relay (MMR) in IEEE 802.16j.

VI. AUTHENTICATION IN GSM AND UMTS

In GSM, network access is protected with EAP-SIM method based on symmetric encryption/decryption key stored at the same time in the SIM (Subscriber Identity Module) card and in the Authentication Center (AuC) or Home Location Registry (HLR) [7]. The key is used for authentication and derivation of the encryption key. IMSI identifier of a user or a pseudonym can be used to protect the identity. EAP-SIM exhibits vulnerability in roaming scenarios between cellular and WLAN hot-spots, because the protocol lacks mutual authentication allowing for user authentication only. The initial EAP-SIM exchange (SIM handshake) does not include tunnel setup to the AuC. As a result SIM-data is exposed to eavesdropping in prone environment of a hot-spot. The EAP-AKA in UMTS is different by support of mutual authentication of both the user and the network.

VII. MOBILITY WITH AUTHENTICATION AT THE DATA–LINK LAYER

A. Basic Service Set transition in IEEE 802.11i

The four-way handshake in 802.11i is a key management protocol and an important part of the RSNA setup process. The central point of 802.11i is that the 4-way handshake (the mutual proof of possession of the MSK) can be repeated with the same PMK, therefore during re-authentication a check for availability of PMK is performed. With the PMK available from previous invocations, the four-way handshake is immediately started to test the connectivity between the MT and the AP, to select the cipher suite and to generate a new Pairwise Transient session Key (PTK).

Say it short, the RSN has an important built-in capability to generate a new PTK each time a re-association is executed. Therefore, transfer of security context from the old- to the new AP during handoff, and the associated risk of a domino effect in case the previous key was somehow compromised, is not an issue anymore. The new PTK is created and installed in the MAC layer, so a secure channel is available for data transmission. An in-depth security analysis of the IEEE 802.11i 4-way handshake can be found in [14].

B. Fast BSS transition in IEEE 802.11r

The IEEE 802.11r [15] enhances data-link layer mobility of WLAN stations (STA) with Fast BSS transition services referred to as the fast roaming. The main objective of Fast BSS transition is to minimize or eliminate connectivity loss during L2 handover within BSS for better support of delay- and loss sensitive applications, such as voice communications. Fast BSS transition enables a station (STA) to prepare handoff ahead of re-association. The preparation includes re-authentication and derivation of the encryption keys for protection of the new channel. For that, the Fast BSS introduces a new framework for security key derivation and management. The process consists of three phases.

During discovery MT locates and determines the target AP for Fast BSS transition. This phase is out of scope of 802.11r leaving space for use of legacy beacon signal, probe requests/response or for other means, such as, the future IEEE 802.11k or Information Services (IS) of the emerging IEEE 802.21 [17].

Authentication: Fast BSS Transition services provide mechanisms for communications with the target AP prior to re-association. The signalling ex-
change is carried over-the-air or through the existing association with the serving AP, i.e. over-the-distribution-system.

Transition: depending on the handoff algorithm the MT chooses the moment to commit transition to the new AP. During the transition current association is terminated and the new one is established. In summary, the Fast BSS transition is a scheme for re-association with reduced handoff latency and a good security through key re-generation. The procedure is applicable to transitions within the same BSS. It is not suitable to inter-BSS transitions which imply L3 handovers due to change of the Access Router or change of the router interface. The signaling exchanges are shown in Fig. 3.

VIII. RE-AUTHENTICATION AT THE NETWORK LAYER

The common mobility management scheme at the network layer for systems based on IP protocol is the Fast Handover scheme for Mobile IPv6 [18].

The Fast Handover is used in combination with AAA exchange for mutual MT-network authentication at the IP layer. To this point, numerous proposals for joint signaling exchange have been made, most of which aim to reduce the combined latency of Mobile IPv6 location update to a mobility agent, and the latency of the authentication procedure with a AAA server, which are both assumed to be located in the home network. The classical approach to mitigate the latency problem is by simultaneous accomplishment of both procedures as presented in Fig. 4.

The complexity of this mobility scheme is in that the L3 handover is often triggered when the signal strength received by the network interface is low and the handover is imminent. This strategy leaves little time to complete both procedures and introduces the necessity to resume and complete the process after the MT gets connected to the new point of attachment (nAP) and to the corresponding new AR (nAR). This requires some grace period for signaling exchange after a L3 handover.
IX. ROAMING WITH MEDIA-INDEPENDENT PRE-AUTHENTICATION

The recent proposal for media-independent pre-authentication (MPA) is a mobile-assisted handover optimization scheme suitable for mobility management protocols operating at the network or higher layers [19].

MPA aims at low-latency mobility of a Mobile Node achieved by means of pre-authentication. A concept similar to IEEE 802.11i is deployed here at the network layer, with supplementary mechanisms for advance acquisition of IP address from the target network and for an in-advance handover to that network while the Mobile Node is still attached to the current one. So, with MPA, a Mobile Node is able to securely obtain configuration parameters from the target network and to exchange IP datagrams still before it commits a handover.

A security association is set up ahead of handoff with the new AR by means of pre-authentication: a proactive tunnel between the Mobile Node and the nAR for host pre-configuration and for data traffic is established (secure proactive handover). MPA should be capable of supporting inter-domain handovers both between the federated and non-federated domains [19].

The target network and the Mobile Node jointly derive a Pairwise Master Key (PMK), using the MPA Security Association established during pre-authentication [20]. From the PMK, distinct Transient Session Keys (TSKs) are derived for each AP in the target network. Other keys, e.g. for bootstrapping the data-link layer can be derived, too.

The target network may install the keys derived from the PMK and used them for secure associations in points of attachment. The keys may be TSKs or some intermediary keys from which TSKs are derived.

After the Mobile Node chooses the target network and switches to the new point of attachment of its choice, it executes a secure association protocol such as IEEE 802.11i 4-way handshake using the PMK as input in order to establish Pair-wise Transient Keys (PTKs) and Group Transient Keys (GTKs) [20]. The PTK will be used for communications protection at the data-link layer.

X. APPLICATION KEYING

Given the vast number of services, the MN usually has to go through multiple level authentications and authorizations, traditionally based on EAP. In terms of overhead and latency this is not optimal. Some proposals have been recently considered in the IETF [28], which assume one EAP authentication to be sufficient. Based on the resulting keying material, multiple separate master session keys for services can be generated. The notion of Application Master Session Key (AMSK) or Usage Specific Root Key (USRK), a service-specific keying material, is introduced which can be generated from the Extended Master Session Key (EMSK). EMSK is normally derived by an EAP method [29], but not used within the EAP keying framework.

When the initial EAP authentication is completed, AMSKs are created and delivered to the service nodes and to the Mobile Node, so that no separate EAP authentications per service are required.

The approach of Hoakey has recently considered proactive and reactive model for AMSKs distribution. In the proactive one the EAP server distributes AMSKs to service nodes, based on the assumptions that the server knows the services to use and the involved nodes. In the reactive model, a service node queries the server for its AMSK upon service request from the accessing node. The proposal introduces a Key Holder, an entity responsible for caching of keying material.

One particular application of USRK is to derive a root for key hierarchy in handover management. It could help to avoid a complete re-authentication during the handover in wireless access networks and improve scalability of the authentication system.

The root for handover key hierarchy is a special USRK named Reauthentication Root Key (rRK). Next, the hierarchy is built based on the rRK with:
R0-Keys used to derive R1-Keys and delivered to Access Nodes. Finally, the R1-keys are used by a Secure Association Protocol (SAP) to create Transient Session Keys (TSKs) for data protection between the Mobile Nodes and the Access Nodes [29]. The architecture is depicted in Fig. 5.

When the MN roams, instead of running full EAP re-authentication each time, it computes appropriate R1-Key appropriate for the Access Node it moves to.

**XI. PRE-AUTHENTICATION IN HOAKEYP**

In the IETF, two possible scenarios for pre-authentication are studied in the Hoakey working group: direct and indirect pre-authentication. In the first scenario, a Mobile Node pre-authenticates with a target authenticator directly. The so called serving authenticator is unaware of this process. Because the target authenticator and the Mobile Node may possibly be in different subnets, in the direct pre-authentication MN-TA signaling will likely run in network layer using PANA transport [30]. In the second scenario, MN pre-authenticates with the target authenticator with the assistance of the serving authenticator. In this case signaling between the Mobile Node and the serving authenticator may go over the data- or over the network layer; the serving authenticator to target authenticator signaling will probably run at the network layer. This work is in progress.

**XII. ROAMING WITH APPLICATION LAYER AUTHENTICATION**

Transporting users’ credentials from one service provider to another is an important issue pertaining to security at the application layer. One solution to that problem, known as a Single Sign-On (SSO), is addressed by Security Assertion Markup Language (SAML) [21]. When the user roams, SAML assures that security credentials roam with him.

SAML defines three types of assertions, or statements, which cover various types of credentials. Authentication Assertions are issued by a party which has successfully authenticated a user (a principal). Attribute Assertions describe attributes of a user, while Authorization Assertions provide information about permissions user have in a certain security domain.

SAML also introduces SAML Protocols such as Assertion Query and Request Protocol, Authentication Request Protocol and Artifact Protocol. SAML Protocols describe how assertions are requested, but do not provide a transport for SAML. Instead, they introduce Bindings between SAML Protocols and the lower level communication protocols for transport of SAML messages. A general-level component of SAML is a profile, which describes how Assertions, Protocols and Bindings are collected for a particular use case, such as e.g. a Web Browser SSO Profile.

In a Web Browser SSO/Artifact Profile a user roams based on a URL containing an artifact which is a reference to SSO assertion. The artifact allows the new provider to retrieve security assertions for a roaming user from another provider. In Web Browser SSO/POST Profile the assertions are transferred by means of a HTML form which contains a complete assertion, and not by a reference. When a user roams, the serving provider pushes a form with HTTP POST mechanism to the target provider.

In spite of some doubts that federated identity may be inefficient e.g. between systems with essentially different policies, many interesting solutions have been presented, e.g. [22].
XIII. GENERIC AUTHENTICATION IN 3GPP

The Generic Authentication Architecture (GAA) is an authentication and key agreement scheme in 3G systems [31]. The process runs between the user (by means of a handset) and the service provider in the Internet with assistance of the cellular network infrastructure. The solution has been proposed to reduce the burden of managing separate credentials by a user of a cellular phone for different services and to reduce the burden of distributing credentials by the service providers. The GAA provides keying material for a client who is willing to use a service and to service providers who rely on a shared secret key authentication. GAA can also sign certificates for applications which require asymmetric authentication. In that sense GAA is an mutual authentication service between the client and the operators, maintained by the cellular network operator, which allows the client and the service provider to re-use the Authentication and Key Agreement (AKA) of a 3G system.

The service is comprised of two phases: bootstrapping authentication procedure (the common part) and the bootstrapping usage procedure. The common part is whereby the client and the operator execute mutual authentication and agree on the shared session keys. These can next be retrieved by the service provider during bootstrapping usage phase, for direct use between the client and the service provider. Alternatively, the key can be used by the client to authenticate his request for certificate to the operator which is retrieved from the PKI infrastructure. Next, the certificates and the corresponding key pairs are used to authenticate to the client to the service provider.

The actual process is implemented in two modes: GBA_ME and GB_U which differ in the security levels [31].

The advantages of using GAA are clear, provided that the service provider and the user jointly trust the operator. There also exist extensions in 3GPP for the combined use of GAA with Single Sign On service.

XIV. AUTHENTICATION METHODS IN RESEARCH

Many of the methods which are currently under standardization for mobility-enabled Internet originate from previous research activities. The collection of the papers on the topic is hudge and falls beyond the scope of this paper. Here we just want to characterize a few representative, in particular those which employ some use patterns now reflected somehow in the schemes.

A. Proactive Key distribution

Pro-active key distribution was proposed to reduce authentication latency by maintaining information about neighboring APs by means of neighbour graph and advance distribution of keying material to all APs in the graph for future use by EAP-TLS, just in case a MT performs a handover there. The method avoids sharing key material amongst multiple APs, at the cost of substantial load of AS in generation and distribution.

B. Predictive authentication

Predictive Authentication is a family of schemes to accelerate authentication within a Frequent Handoff Region (FHR) determined from past user movement patterns [33]. In this scenario the Authentication Server responds to an authentication request coming from a MT by sending multiple authentication responses to all APs in the FHR.

Predictive authentication can also be achieved by means of security context transfer over a pre-established IPSec tunnel between APs, members of the neighbor graph. This method was specified in IAPP protocol (IEEE 802.11f) and accelerated with proactive caching. Now it is considered unsafe, since not only context, but also a risk can be transferred. The above methods are similar in concept and applicability/they can work in a single administrative domain. Inter-domain handoffs are out of scope, or difficult to maintain in a secure way in predictive authentication.
C. Inter-domain authentication

Some solutions to fast authentication problem have been presented in [27]. They exploit the set of direct and implicit trust relationships that exist between particular elements of Wireless LAN. The first approach is extending IAPP for inter-domain mobility, by setting up pre-existing secure links between APs from separate domains therefore enabling IAPP context transfer between different domains. The important aspect here is that a roaming agreement between the domains must exist. Other approach for optimizing inter-domain authentication is Inter-domain Proactive Key Distribution. Here, the idea is to compute and deliver PMKs to possible handover candidates in an inter-domain environment. Three main aspects have to be taken into account. First, MN location has to be tracked in the home AS. Second, the handover candidates have to be determined, therefore some kind of inter-domain neighbour graph has to be created, and, the computed keys have to be distributed to handover candidates in some way.

D. Roaming key based handovers

Roaming key based handover is a new approach introduced in [26], [27]. It consists of three main components: the Roaming Key (RK), the Context Information (CI), a set of security parameters, and the Security Information (SI). The RK was a new key whose purpose is to provide mutual authentication when handover happens. It can also be used as an encryption key in inter-domain handover until new PTK and PMK are derived.

When the station (STA) associates with an Access Point (AP), CI is distributed in the neighbor graph. If, at some point in time, STA handovers, the serving AP sends the SI to the STA and to the target AP. Now both STA and the target AP possess RK by means of which they can communicate until receiving new PTK and/or PMK.

This approach can address the inter-domain handover problem, but, without going into details, a new disassociation frame would have to be added to IEEE 802.11 standard. Furthermore, inter-domain handover scenario needs prior trust relationship between domains by means of some business agreement.

XV. SUMMARY

Authentication of entities is fundamental in the networking environment. The authentication process, and its manifold subtle issues, have been elaborated by multiple groups and standardization bodies for different network technologies and systems. Most of the approaches to authentication in the context of infrastructure-based networks, have been rooted in the EAP model. Over time, they have been improved when cavities and new types of threats were discovered.

Recent shift to the wireless Internet, forces us to revisit many assumptions about the authentication process. The interest in strong authentication methods is growing as well as acceptance for complex schemes, but with little tolerance for the associated overhead and latency.

With the current uncertainty about the business models, networks and services in next-generation networks there is a need for a flexible, technology-independent authentication framework accommodating requirements typical of wireless systems, and showing good performance under complex mobility schemes with single sign-on capability, not only by means of SAML, but also, for example, in other well established systems such as Kerberos.

XVI. ACKNOWLEDGMENTS

The work described in this paper has been done in IST FP6 Integrated Project DAIDALOS II (IST-2006-026943). The authors wish to thank the anonymous reviewers for their feedback on the paper.
XVII. SELECTED REFERENCES


[24] “IEEE Trial-Use Recommended Practise for


BIOGRAPHIES

Marcin Dabrowski received M.Sc. degree in electronics and telecommunications in 2005 and, next, M.Sc. in computer science in 2007 both from the AGH University of Technology in Krakow, Poland where he is now a Ph.D. student and delivers "Secure Communications Systems“ course. His research activities focus on digital identity management and security.

Marcin Dabrowski takes part in IST Integrated Project Daidalos II. He works as a Project Manager in Telco OSS at Comarch.

Piotr Pacyna received M. Sc. degree in computer sciences and a Ph.D. degree in telecommunications in 2005 from the AGH University of Technology in Krakow, Poland where he is lecturing courses on "Next Generation IP networks", "Distributed Network Environments" and "Secure Communications Systems". His research activities focus on next-generation wireless communications systems, mobility and security. Piotr Pacyna spent sabbatical leave in CNET France Telecom in Paris, France and in University of Carlos III in Madrid, Spain.

Piotr Pacyna has been active in ACTS and IST research projects founded by the European Comission: Broadband Trial Integration (1997-2000), Moby Dick (2001-2003) and in IST Integrated Project Daidalos (2003-2006). In Daidalos II (2006-) he has been a member of the architecture design team.
Analysis of Security Vulnerabilities and Countermeasures of Ethernet Passive Optical Network (EPON)

Di Jin, Stamatios V. Kartalopoulos, Pramode K. Verma
School of Electrical and Computer Engineering
Telecommunications Systems Program
University of Oklahoma - Tulsa
Tulsa, OK 74135, USA
Email: di.jin@ou.edu, kartalopoulos@ou.edu, pverma@ou.edu

ABSTRACT

Ethernet-based Passive Optical Network (EPON) is considered a very promising solution for the first mile problem of the next generation networks. Due to its particular characteristic of shared media structure, EPON suffers many security vulnerabilities. Communication security must be guaranteed when EPON is applied in practice. This paper gives a general introduction to the EPON system, analyzes the potential threats and attacks pertaining to the EPON system, and presents effective countermeasures against these threats and attacks with emphasis on the authentication protocols and key distribution.

Key words: EPON, security vulnerabilities, attacks, countermeasures, authentication protocols, key distribution

I. INTRODUCTION

The world has never stopped requiring more and more bandwidth from networks due to the expansion of numerous new services over the Internet. Among these new services, triple play (voice, data and video) nowadays becomes the basic requirement for people who have access to the Internet from their home. However, due to the bandwidth bottleneck problem of the access network, there was not a good solution to provide such service in a cost-effective and efficient manner. Fiber-to-the-home (FTTH) solution has been accepted as the ultimate solution for the bandwidth bottleneck problem of the access network, nevertheless so far this solution has not been sufficiently cost-effective to be applied. Alternatively, as a very promising solution to the first mile problem of FTTH, EPON has been considered as a very good candidate which can provide Gbps, 10Gbps and even 100Gbps bandwidth recently with cost-effective feature. And actually EPON has been approved as the 802.3ah standard for the “Ethernet in the first mile” by IEEE Standards Board in July 2004 [1].

As EPON is moving into deployment, security issues come to the front with unprecedented attentions of people, and more and more research papers regarding those issues are being published [2-11].
Unlike Local Area Network (LAN) in an enterprise building or a university campus, EPON components are not bounded within a closed area with the physical control of the operator, namely, EPON is an open and shared media network with irrelative users, which enhances the security vulnerabilities and the probability of being attacked. However, as an extremely important property associated with the service agreement, the security of data transmission is crucial in protecting the privacy of the customers and the confidentiality of their communications [12]. In fact, the customers of EPON, who can easily access the network components, could be the most likely attackers. They can eavesdrop to steal sensitive information; intercept and alter data to take advantage of it; impersonate others to send out malicious information; fool the network to get free services; and even break down the whole network. Hence, the security issues are extremely important to EPON future success.

This paper will analyze the security vulnerabilities of EPON system and then present some effective countermeasures to those vulnerabilities. The organization of this paper is as follows: Section I is the background introduction; section II gives the overall introduction to the EPON system; in section III, we will analyze the security vulnerabilities of the EPON system in detail; section IV presents the effective countermeasures to those vulnerabilities and analyzes the advantages of those countermeasures; and section V concludes the whole paper.

II. EPON SYSTEM STRUCTURE OVERVIEW

Basically EPON is a tree topology optical network composed of Optical Line Terminal (OLT), Optical Network Units (ONUs), Optical Distribution Network (ODN) and optical fibers. EPON combines Ethernet with Passive Optical Network (PON), providing low overhead and high transmission efficiency, reducing the amount of optical fibers, transceivers, and switches of the access network, and eliminating the usage of active components on the path, which greatly reduces the cost of equipments, and Operation, Maintenance and Administration (OMA) expenditure. As a result of the above mentioned features, EPON becomes a very cost-effective broadband access network [13]. The schematic diagram of EPON system structure and data transmission is depicted in Figure 1.

In EPON system, OLT is usually located in the Central Office (CO); ODN (usually an optical splitter/coupler) is placed near the neighborhood or the office building; and ONU is often mounted at the premise of the home, in the home, or in the office.
Usually one ONU can serve several end users at the same time, interconnected by cables.

In the downstream direction data transmission, EPON is a point-to-multipoint broadcasting optical network. Ethernet packets transmitted by OLT pass through a 1:N (typically N is a number between 4 and 64) passive optical splitter and reach all the ONUs, and each ONU picks up the packets destined to itself and then passes those packets to its end users. Each ONU is supposed to only pick up the packets belonging to itself and ignore all the other packets, although each ONU can actually receive all the packets transmitted by OLT. Since EPON is a shared media network, and broadcasting is the mechanism used to transmit data in the downstream direction, it is vulnerable to many types of security attacks, which could be very serious to the EPON system.

In the upstream direction, EPON is a multipoint-to-point broadband optical network. The packets transmitted by the end users first arrive at their ONU, and then ONU aggregates all the packets of its end users and transmits these packets to OLT through the optical coupler at a specific time slot assigned by OLT. In the upstream direction, all the packets have the same destination, i.e. OLT, so the upstream direction data transmission is relatively secure compared to the downstream direction in the sense that all the packets transmitted by one ONU can only reach OLT rather than other ONUs. However it is still not secure enough to transmit data in the clear in the upstream direction. By some means, for example, tapping the cables or fibers on the path, attacking ONU, ODN or OLT, and making use of the reflection of the upstream direction data transmission to the downstream direction, the upstream direction data transmission security can still be compromised.

EPON is an Ethernet-protocol-based passive optical network, which uses the standard Ethernet packets to transmit data in the downstream and upstream direction. The format of Ethernet packets is depicted in Figure 2. As showed in Figure 2, the Logic Link Identifier (LLID) which indicates the destination of the packet is placed in the preamble field of the packet; following the preamble is the header of the packet, describing the start-of-packet delimiter, the destination and the source address, and the length of the packet; the payload contains the data to be transmitted; and the Frame Check Sequence (FCS) is attached at the end of the packet to check for the damage of the packets. All the data including user data, management and control information are delivered in normal Ethernet packets.

**III. POTENTIAL THREATS AND ATTACKS IN EPON SYSTEM**

EPON’s security vulnerabilities reside in its openness of the structure and its serving noncooperative users. ODN and ONU are often placed outside of the office or house, hence malicious attackers can easily approach those components physically and launch some attacks. Even if the components are placed inside of the office or house, there might still be some attackers from the inside to attack the EPON system. In addition, the point-to-multipoint broadcasting characteristic and its structural feature actually make it even easier for the attackers to carry out successful attacks.

**A. Eavesdropping**

Eavesdropping is a type of passive attack, which only “listens to” or “copies” the message without altering it. In EPON system, eavesdropping is very easy to be realized in the downstream direction, while relatively difficult in the upstream direction. Due to the broadcasting characteristic of EPON in the downstream direction, each ONU can actually
receive all the packets from OLT, regardless of the destinations of those packets. And the system just simply assumes that each ONU will ignore all the other packets without trying to pick up them after picking up the labeled packets matching its LLID. However when some bad actors are involved in some ONU, that ONU could be instigated to pick up other packets which do not belong to it. In doing so, that ONU can eavesdrop the data transmitted from OLT to other ONUs.

Eavesdropping in the upstream direction is relatively difficult but still possible. Although all the packets transmitted by ONUs have the only destination - OLT, and no ONU will receive any packets from other ONUs, the eavesdropping can still be achieved by the following several means: Eavesdropping the end terminal equipment, for example, the computer that the end user is using; tapping the cables or fibers on the path from the end user’s computer to OLT; attacking the components on the path, for example, ONU and OLT, by looking into their buffers or memories; and taking advantage of the reflection of the upstream direction data transmission to the downstream direction and so on. So we can conclude that it is not secure to transmit data in the clear in the upstream direction, and data must be encrypted in both directions to guarantee the confidentiality of the data transmission.

B. Interception

Different from eavesdropping, interception is another type of attack - active attack, which captures the message, alters it, and then sends the altered message to others. In EPON system, OLT, ODN and ONUs are interconnected by single mode fibers, and ONUs and end users are interconnected by cables. As long as fibers and cables are used for interconnection, there is always a possible way to intercept data by tapping those transmission media. However tapping the transmission media is very different. Sophisticated equipments and operations would be needed, and the attackers must also be smart enough to avoid being detected by the monitoring mechanism of the system and the system operator. Another way to intercept data mentioned in reference [14] is to make use of the special structure of the splitter/coupler (ODN). A special device can be used to connect the unused ports in the splitter/coupler to intercept all the upstream direction data transmission.

C. Impersonation

Impersonation is a type of attack in which the malicious users pretend to be another user to receive sensitive information or send vicious data by forging its identifier (LLID) to the victim identifier. In the network, ONUs and end users each have a unique identifier which is used to identify themselves within that network. With forged identifier, malicious ONUs and end users could pick up the packets which match the forged LLID but actually not belonging to them. Through this operation, malicious attackers can impersonate others to function as if they were the one who is being impersonated. The malicious attackers can do more than that. They can transmit counterfeit and harmful information to give troubles to those victims and even disturb the normal running of the whole network.

D. Denial of Service (DoS)

Denial of service is a type of familiar attack over the Internet, which usually brings on some serious aftermaths. EPON system is also vulnerable to this type of attack [15]. In the upstream direction, ONUs share the upstream bandwidth with each other, OLT allocates bandwidths to ONUs dynamically according to their requests and the available system capacity. Denial of service attacks could be achieved if a malicious user gets the access to the network and keeps asking for a high bandwidth and sending a large amount of garbage traffic, then the normal running of the network would be affected, and network performance would be reduced. And in some worse cases, when a lot of attackers launch the attacks at the same time cooperatively, no other users would be able to get any bandwidth and no services
could be provided to the customers. If the network performance reduces to a severe situation, then the whole network could malfunction, which would cause tremendous loss for both the customers and the service providers. This is the basic idea of the so-called Distributed Denial of Service attack (DDoS) over the Internet.

E. Theft of Service

Theft of service is a type of attack in which the malicious attacker asks for services under the name of the victim, so that the attacker could get those services for free. This attack could happen when a malicious user impersonates his neighbor in the network and transmits and receives packets which are not billed to his account but to his neighbor account. This attack could be realized by inserting the victim’s LLID into the packets actually transmitted by the attacker and intercepting the packets with the victim’s LLID. In doing so, when OLT receives these packets from the upstream direction, the victim will be billed instead of the attacker. Theft of service attack makes use of impersonation and interception attacks, since in the attacking process of theft of service, impersonation and interception techniques will be used. But those attacks have different purposes. The main purpose of theft of service attack is just to get free services and stay away from making any other trouble, however impersonation and interception attacks aim at getting sensitive information from others and sending harmful information to others.

IV. COUNTERMEASURES TO THE VULNERABILITIES IN EPON SYSTEM

In order to counter those potential threats and attacks discussed above, some effective countermeasures must be introduced to the vulnerable EPON system. Basically, encryption of the data and authentication of the communicating parties are two effective methods commonly used to provide data confidentiality, data integrity, user privacy and user authentication [16]. With effective countermeasures, the security attacks of EPON system could be frustrated, and the communication security could be guaranteed.

A. Encryption

Encryption is an effective countermeasure to thwart eavesdropping and interception. Even though the attacker could eavesdrop or intercept the data transmitted by others, the data would remain unintelligible without the possession of the decryption key. It should be stressed here that although the upstream direction data transmission is relatively secure compared to the downstream direction, it is still not secure enough to transmit data in the clear. If the data transmission in the upstream direction is in the clear, then attackers will get plenty of opportunities to carry out many kinds of attacks successfully. So in order to guarantee the communication security, the data encryption of EPON system must be bidirectional.

Traditionally, there are two major categories of cryptography: Symmetric key cryptography and asymmetric key cryptography. Asymmetric key cryptography such as RSA can be used not only for encryption, but also for authentication and key distribution to some extent. However, due to the characteristic of asymmetric key cryptography that it will take much more computational effort for the asymmetric key cryptography to be performed than the symmetric key cryptography, asymmetric key cryptography is more often used for authentication and key distribution rather than for data encryption, especially in the case when encryption speed is a sensitive parameter of the system requirements. And on the other hand, due to the less effort-consuming characteristic of symmetric key cryptography, it is mainly used for data encryption. For EPON system, which acts as the access network, is supposed to be very cost-effective and efficient, so less effort-consuming and less complicated encryption technology should be adopted to meet the low-cost requirement of the broadband access network. As a result, symmetric key cryptography should be applied to data encryption, and asymmetric key cryptography should
be adopted for authentication and key distribution in EPON system.

Recently, with the dramatic increase of computational power that people can achieve, traditional cryptographic technologies (such as symmetric and asymmetric key cryptography) which are based on computational difficulty seem not as secure as before. Fortunately, more advanced cryptographic technology such as quantum cryptography is emerging. Quantum cryptography can provide unconditional security to data transmission based on the unbreakable principles of quantum mechanics, which assures people that in the future we can still communicate securely. We will talk about those traditional and advanced encryption technologies in detail and present some new encryption mechanisms in a subsequent paper, while focus on authentication protocols and key distribution in this paper.

B. Authentication

Authentication is used to identify the creditable users who have paid for the services in the network and to verify the legitimacy of the communicating parties. Authentication can also provide data confidentiality and user privacy in the sense that unauthorized communicating parties will not have the chance to read the data because they can not pass the authentication procedure. It is a good countermeasure to frustrate impersonation, denial of service and theft of service attacks. Together with the encryption technology, basically EPON can keep away from those known attacks listed above, and the security of data transmission can be guaranteed.

In EPON system, the authentication process mainly consists of two parts, ONU/OLT authentication and User/ONU authentication. The purpose of ONU/OLT authentication is to set up trusts between ONU and OLT, so that neither attackers can pretend to be ONU, nor attackers can masquerade OLT. There will not be a third party appearing in the communication between OLT and ONU, and OLT and ONU will be communicating to the one that is supposed to be. The User/ONU authentication is used to guarantee that the user who gets the access to the network is the authorized one, and the ONU is the creditable one without being mimicked by others. The authorized user not only means that he has paid for the service, but also means that he must be the one whom he claims himself to be. Usually, ONU/OLT authentication takes place first, and after the trusts between ONU and OLT have been set up, User/ONU authentication will start. In the following, we are going to present two new authentication protocols for OLT, ONU and end users, offering much better security performance. Before introducing the new authentication protocols, it is necessary to explain the notation defined in the protocols, which is showed in Table I.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>E(K, [M]</td>
<td></td>
</tr>
<tr>
<td>E_PONU/OLT</td>
<td>ONU/User’s expanded password</td>
</tr>
<tr>
<td>ONU_address</td>
<td>ONU’s MAC address or unique identification of product</td>
</tr>
<tr>
<td>N</td>
<td>None</td>
</tr>
<tr>
<td>T</td>
<td>Timestamp</td>
</tr>
<tr>
<td>ID_ONU/OLT</td>
<td>ONU/OLT’ LLID</td>
</tr>
<tr>
<td>ID_User</td>
<td>User’s username</td>
</tr>
<tr>
<td>PR_ONU/OLT/User</td>
<td>ONU/OLT/User’s private key</td>
</tr>
<tr>
<td>PU_ONU/OLT/User</td>
<td>ONU/OLT/User’s public key</td>
</tr>
<tr>
<td>E(PR_ONU, [PU_ONU])</td>
<td>ONU’s certificate</td>
</tr>
<tr>
<td>E(PR_OLT, [PU_OLT])</td>
<td>OLT’s certificate</td>
</tr>
<tr>
<td>E(PR_User, [PU_User])</td>
<td>User’s certificate</td>
</tr>
<tr>
<td>SK_ONU/OLT/User</td>
<td>Session key for ONU/OLT and User-ONU</td>
</tr>
<tr>
<td>NP_ONU/User</td>
<td>ONU/User’s new password</td>
</tr>
</tbody>
</table>
1) ONU/OLT Authentication: The ONU/OLT authentication protocol is showed as follows in Figure 3.

Initially, OLT sends ONU a “GATE” message for “Auto-Discovery”, upon receiving this message, ONU sends OLT a responsive message “E(EPONU, [ONUaddress, N1, T1 || H(M)])” to request for registration. In this message, N1 and T1 are nonce and timestamp, respectively, which are used to prevent from replay attacks; H(M) is the hashed value of the message, which can detect if the message has been changed by others and provide a certain degree of authentication by the fingerprint function of the hashed value to the message; the message is not transmitted in the clear, but encrypted by the expanded password of ONU “EPONU”, since the upstream direction data transmission is not secure in EPON system; and the expanded password is used to enhance the security, due to the shortness of the user password and its vulnerability of being broken. We will talk about the expanded password in detail in the following subsection “Analysis of Authentication Protocols”. After receiving this message, OLT looks into its predefined table in its memory, for the expanded password. Since one OLT may serve 4 to 64 ONUs, on average it will take OLT 2 to 32 times to find the right expanded password to decrypt the message, which is quite acceptable for the infrequently used authentication procedure. After decrypting the message, OLT can verify if the “ONUaddress” matches the expanded password by checking the table in its memory, which also provides a certain degree of authentication in the meantime. Then OLT assigns ONU a LLID and tells ONU its LLID by sending the encrypted message “E(EPONU, [IDONU, IDOLT, ONUaddress, N1, T1, N2, T2 || H(M)])”. As a response, ONU sends OLT another encrypted message “E(EPONU, [IDONU, IDOLT, N2, T2 || H(M)])” indicating that it has received OLT’s message. Till now, the ONU registration process finishes.

ONU/OLT authentication process begins with a “GATE” message sent from OLT to ONU, followed by another message sent from ONU to OLT:
“E(EP_{ONU},\{ID_{ONU},\ ID_{OLT},\ E(PR_{ONU},\ \{PU_{ONU}\}),\ N_3,\ T_3\ ||\ H(M)\})”. In this message, ONU’s certificate “E(PR_{ONU},\ \{PU_{ONU}\})” is included. Only ONU could have made that certificate, because only ONU has its private key; and since OLT has ONU’s public key, it can decrypt and verify ONU’s certificate. The same thing happens to the next message “E(\{PU_{OLT}\},\ \{ID_{OLT}\},\ E(PR_{OLT},\ \{PU_{OLT}\}),\ N_3,\ T_3,\ N_4,\ T_4\ ||\ H(M)\)”**: Only ONU could open it, and only OLT could have made that certificate “E(PR_{OLT},\ \{PU_{OLT}\})”, so that OLT’s legitimacy could be verified. So far the ONU/OLT authentication process ends.

After ONU and OLT authenticate each other, the session key exchange process starts. ONU sends OLT a message “E(\{PU_{OLT}\},\ \{ID_{OLT}\},\ SK_{OLT},\ \{PU_{OLT}\},\ N_3,\ T_3,\ N_4,\ T_4\ ||\ H(M)\)” encrypted by OLT’s public key, informing OLT the newly assigned session key “SK_{OLT}” and ONU’s newly updated user password “NP_{ONU}”. And OLT responds with a message “E(SK_{OLT},\ \{ID_{OLT}\},\ \{ID_{OLT}\},\ N_5,\ T_5,\ N_6,\ T_6\ ||\ H(M)\)” encrypted by the newly assigned session key containing the LLIDs, nonces, and timestamps of the previous message. And later on, OLT and ONU send each other a handshaking message to make acknowledgement by encrypting the “GATE” and “REPORT” message using the newly established session key “SK_{OLT}”. Till now, the whole ONU/OLT authentication process finishes, and from now on ONU and OLT can communicate securely with the newly established session key.

2) User/ONU Authentication: The User/ONU authentication protocol is depicted as follows in Figure 4.

First, the user transmits a “Start” message to ONU, and upon receiving this message, ONU sends back a “Request ID” message to the user to request its ID, which is the username of the user. Then the user responds to ONU by sending a message “E(EP_{User},\ \{ID_{User}\},\ N_1,\ T_1\ ||\ H(M)\)” encrypted by its expanded password. Already having the username and expanded password of the user in its memory, ONU checks the table to search for a match, namely, ONU tries all the expanded passwords of its end users one by one to see if the message could be decrypted by a certain expanded password. If ONU succeeds in doing so, then ONU checks if the username con-

---

Fig. 4. User/ONU Authentication Protocol
tained in the message matches the expanded password, if they match each other, then the user is authenticated in some degree. Since one ONU only serves several end users, it is quite easy for ONU to find the right expanded password. After this, ONU sends a message “E(EPUser, [IDONU, IDUser, N1, T1, N2, T2 || H(M)])” to the user, and the user decrypts this message using its expanded password and gets ONU’s LLID. Then the user sends ONU another message “E(EPUser, [IDONU, IDUser, N2, T2 || H(M)])” as an acknowledgement. So far the user registration process finishes.

Following the user registration process, User/ONU authentication process takes place making use of the asymmetric key cryptography and certificates. The user sends ONU a message “E(EPUser, [IDONU, IDUser, E(PRUser, [PUUser]), N3, T3 || H(M)])” encrypted by the expanded password containing the user’s certificate “E(PRUser, [PUUser])”. Only the user could have made that certificate because he is the only one who has the private key. ONU could verify the user’s legitimacy by decrypting that message using the user’s expanded password and decrypting the certificate by the user’s public key. Then ONU sends its own certificate to the user by the message “E(PUUser, [IDONU, IDUser, E(PRONU, [PUONU]), N4, T4 || H(M)])”, which is encrypted by the user’s public key, so that only the user could open that message. After decrypting the message, the user can verify ONU’s legitimacy by validating ONU’s certificate. At this point, ONU and the user finish authenticating each other.

The next step is to set up a session key between the user and ONU and update the user’s password. First, the user sends a message “E(PUUser, [IDONU, IDUser, SKUser-ONU, NPUser, N5, T5 || H(M)])” to ONU containing the newly assigned session key “SKUser-ONU” and the newly updated user password “NPUser”. Then ONU decrypts the message, gets the newly assigned session key, and responds with a message “E(SKUser-ONU, [IDONU, IDUser, N5, T5, N6, T6 || H(M)])” containing the same nonce, timestamp, and IDs of the previous message. From now on, the user and ONU can communicate securely using symmetric key cryptography with the newly established session key, and the whole User/ONU authentication process ends.

C. Key Distribution

Key distribution is an extremely important topic in communications security. Usually the security performance of a system relies on the security of key distribution to a very high degree. Basically there are three major methods to transfer keys among communicating parties. The first method is primordial, secure but inefficient, which is to transfer keys by human being, the courier. The courier can physically travel from one party to another and deliver the keys. Another method is to use the arbiter to assign keys to the communicating parties. The prerequisite of this method is that the arbiter should be trusted by the communicating parties, and the arbiter should have a secure way to transfer the keys to the communicating parties. The third method is to establish the keys by the communicating parties themselves through an “automated” way. Here through an “automated” way means either through the previously used session key, or public key distribution, or other advanced methods. The third method is the focus of current research.

In this paper, we use the combination of the above mentioned methods. Initially, the manufacturer (“arbiter”) preassigns each OLT, ONU and end user a password, sealed in an envelope. The password is associated with the unique identification of product of OLT, ONU and end user, which will later be used as the composition of their LLID, MAC address and user name. The password is delivered by the installation worker (“courier”) in a sealed envelope, and if the envelope is broken, then a new password will be assigned in another sealed envelope. OLT knows all the passwords of the ONUs it serves from its memory, since all the passwords and IDs are prestored in its memory. And ONU also knows all
the passwords of the end users it serves from its memory, since all the passwords and IDs are also prestored in its memory. When OLT and ONU receive a message encrypted by a certain password, it looks into the memory and tries to find an applicable password to decrypt the message, and later checks if the ID in the decrypted message matches the ID in the memory which is associated with that password. If they match each other, then the authentication of the message sender could be verified, because only that sender has the password and could have put the right ID in the encrypted message. And OLT’s and ONU’s authentication could also be verified, because only OLT and ONU could know the password and decrypt the message. After the registration process and authentication process finish, the two communicating parties will establish a session key for future communications. In session key exchange process, the public key (“automated”) is used to encrypt the message containing the new session key, and after decrypting that message, the other communicating party will get the new session key and later on establish a secure communication between the two parties using symmetric key cryptography. The session key should be changed frequently to enhance the security performance of the communication. After the first successful establishment of the session key, it is very easy to update it simply by sending a message containing the newly updated session key encrypted by the old session key. With frequently changed session keys, the communication security can be guaranteed.

D. Analysis of Authentication Protocols

1) The usage of expanded password: As we mentioned before, predefined and prestored passwords are used initially to set up the registration and authentication process. To be more accurately, those passwords are not exactly the original ones but the expanded ones. The reason why we use the expanded passwords other than the original ones is mainly because longer passwords will provide higher security level. Shorter passwords are easy for people to remember, however trivial for attackers to break. An algorithm that is used to expand the length of the password is involved here, and any currently commonly used algorithm will work. The idea of using expanded password other than the expanding algorithm itself is the main point we are considering here.

With the expanded password, we can encrypt the data transmitted both in the downstream direction and in the upstream direction. This is a big advantage because this solves the problem that during the initial data transmission we used to have to transmit data in the clear. In addition, the usage of expanded password not only avoids the initial data transmission in the clear due to the lack of the initial encryption key, but also provides an enhanced security and authentication level to the data and the communicating parties. As we know, during the initial data transmission, a lot of sensitive information such as the user name, LLID, MAC address, and other OMA information will be involved, and any revealment of those information will significantly help the attackers to carry out successful attacks. For example, with revealed MAC address, the attackers could launch traffic analysis attack; with revealed LLID and user name, the attackers could carry on eavesdropping, interception, impersonation, and other attacks we mentioned above. Hence, solving the problem of initial data transmission in the clear is the very critical step of establishing a high security performance system.

2) The usage of hash function, nonce and timestamp: We use hash function, nonce and timestamp throughout the authentication process. The advantage of using the hashed value of the message and attaching it to the message is that the communicating parties can easily detect any alteration of the message content, which makes the message resistant to interception attacks. The hashed value of a message works as the fingerprint of the message. Suppose the attacker successfully changes the content of the message, then he must change the hashed value of that message accordingly to avoid being detected, however he will find that there is no way to correctly change the hashed value because he is not aware of the
algorithm of the hash function that is used by the sender of that message. Alternatively, the attacker may choose to keep the hashed value unchanged and try to find another message which happens to have the same hashed value, however, unfortunately, the attacker will find it impossible to accomplish due to the one-way function characteristic of the hash function, namely, given the message, it is easy to get the hashed value if one knows the algorithm of the hash function; however, given the hashed value, it is practically infeasible to recover the message even if one knows the algorithm of the hash function. The usage of hash function greatly improves the security level of the message.

Nonces and timestamps are used to thwart replay attacks. Sometimes the attackers may save a previous message and reuse it in a later time to fool others. For example, one bank requests the other bank to transfer one million dollars in a certain time, and later the first bank sends the same request and asks for another transfer of one million dollars. The second bank has no way to tell if the second request is a new request or an old one which has already been transacted, since there is no identifier in the message indicating the series number of the message and the time when the message was sent. The nonces and timestamps can work as the labels of the message and can be used to distinguish different messages in sending time and series number, so that replay attacks will hardly succeed in this case.

3) The usage of certificates, symmetric and asymmetric key cryptography: The certificate is the pivotal component used in the authentication protocols to set up authentication between the two communicating parties. It can be made by encrypting the public key with the private key of the communicating party. The certificate is a creditable symbol to verify the identity of a communicating party, because only that communicating party could have known the private key and made the certificate.

Due to the effort-consuming characteristic of asymmetric key cryptography, it is only used in initial authentication process and session key exchange process. After successfully authenticating each other and establishing the session key, the later updating of the session key is achieved by using symmetric key cryptography, which is much less effort-consuming and much faster than asymmetric key cryptography. And for data encryption in EPON system, symmetric key cryptography is more suitable than asymmetric key cryptography because symmetric key cryptography can provide higher data encryption speed, so we choose symmetric key cryptography instead of asymmetric key cryptography, although asymmetric key cryptography can provide data confidentiality and authentication for the communicating parties at the same time, on which point asymmetric key cryptography can surpass symmetric key cryptography.

V. CONCLUSION

This paper gave us a general introduction to the EPON system, analyzed the security vulnerabilities of the EPON system and pointed out potential threats and attacks in detail, and summarized appropriate countermeasures. It has also presented some novel authentication protocols with enhanced security performance. We hope this paper could benefit the future research on the EPON security topic, and provide some inspiration to those people who are working in this field.

REFERENCES

[4] P. R. M. Inacio, M. Hajduczenia and M. M.


BIOGRAPHY

Di Jin received his B.S. degree and M.S. degree from Dalian University of Technology, Dalian, China, in 2001 and 2005, respectively, both in Electrical Engineering. He is currently a Ph.D. student and a graduate research assistant in the School of Electrical and Computer Engineering at the University of Oklahoma. His research interests include information/network security, cryptography, optical network communications, and intelligent control. He is a graduate student member of IEEE, and a member of IEEE Communications and Information Security Technical Committee.
Stamatios V. Kartalopoulos Ph.D., is the Williams Professor in Telecommunications Networking at the University of Oklahoma. His research emphasis is on optical communication networks (FSO, long haul and FTTH), optical technology including optical metamaterials, and optical communications security and countermeasures including quantum cryptography and quantum key distribution. Prior to this, he was with Bell Laboratories where he defined, led and managed research and development teams in the areas of DWDM networks, SONET/SDH and ATM, Cross-connects, Switching, Transmission and Access systems. He has received the President’s Award and many awards of excellence.

He holds nineteen patents in communications networks, he has published more than hundred research papers, and seven reference textbooks: “Optical Bit Error Rate”, “Next Generation SONET/SDH”, “DWDM: Networks, Systems and Components”, “Introduction to DWDM Technology”, “Fault Detectability in DWDM”, “Understanding SONET/SDH and ATM”, “Understanding Neural Networks and Fuzzy logic”, two additional books in DWDM have been published in Chinese, and he has contributed chapters to other books.

He is an IEEE and has been a Bell Labs Distinguished Lecturer, and he has lectured at international universities, at NASA and conferences. He has been keynote speaker and plenary speaker at major international conferences, has moderated executive forums, has been a panelist of interdisciplinary panels, and has organized symposia, workshops and sessions at major international communications conferences. Dr. Kartalopoulos is an IEEE Fellow, chair and founder of the IEEE Com-Soc Communications & Information Security Technical Committee, member at large of IEEE New Technologies Directions Committee, and he has served as editor-in-chief of IEEE Press, chair of ComSoc Emerging Technologies and of SPCE Technical Committees, area-editor of IEEE Communications Magazine/Optical Communications, member of IEEE PSPB, and VP of IEEE Computational Intelligence Society. He is a member of SPIE and a member of the Editorial Board of China Communications.

Pramode K. Verma Ph.D., is the Professor of School of Electrical and Computer Engineering and the Director of Telecommunications Systems at the University of Oklahoma. Dr. Verma has more than twenty years of leadership experience in the telecommunications industry. In his last position with Lucent Technologies as Managing Director - Business Development, Global Service Providers Business and Business Communications System, his responsibilities included creating strategic alliances and partnerships with leading organizations, and managing the associated P&L. He also held professional and management positions with Lucent Technologies - Bell Laboratories for fifteen years.

Dr. Verma obtained his doctorate in Electrical Engineering from Concordia University in Montreal, Canada in 1970 and an MBA from the Wharton School of Business at the University of Pennsylvania in 1984. He is the author/coauthor of over 75 publications and several books in telecommunications, computer communications and related fields.

He is a past president of the International Council for Computer Communication, a Washington D. C.-based global organization; a senior member of the Institute of Electrical and Electronics Engineers (IEEE); and registered as a Professional Engineer, Province of Ontario, Canada.
Analysis on the Resilience of Key Pre-distribution in Sensor Networks

Shen Yulong\textsuperscript{1,2)} Ma Jianfeng\textsuperscript{1,2)} Pei Qingqi \textsuperscript{2)}

1.School of Computer Science & Technology Xidian University, 
2.Key Laboratory of Computer Networks & Information Security, Ministry of Education, Shaanxi 710071

ABSTRACT

Resilience against node capture is one of the main indicators of the key pre-distribution security in sensor networks. On providing the attack model and the definition of the resilience against node capture of sensor networks, the resilience of basic random key pre-distribution, Q-composite random key pre-distribution and their reinforced schemes are analyzed and compared in depth. Research results show that the size of key pool, the numbers of the keys stored in nodes and the value of Q determine the resilience of random key pre-distribution. The tradeoff between the resilience, security connectivity and costs in sensor networks is presented. These researches lay a foundation on the design of the secure protocol and the algorithm in the specific application environment of sensor networks.

Key Words: sensor networks, resilience, analysis, key pre-distribution

I. INTRODUCTION

Sensor networks have found wide applications in the military and civilian fields, and have become a hot area in information researches [1]. Along with the growing evolvement of the practical process of sensor networks, especially in the military application, the security of sensor networks has become an outstanding problem which is facing a great challenge [2]. In fact, the lack of effective security systems has become the main obstacle of sensor network development.

The technologies available guarantee the sensor network security from the aspects of confidentiality, integrity, non-repudiation, authentication, authorization, and intrusion detection. Sensor nodes are captured by the attackers with ease for they are usually deployed in the enemy areas. As a matter of fact, the resilience against node capture is one of the important security features in sensor networks. This paper researches the resilience of key pre-distribution scheme, balances the resilience, security connectivity and costs, and designs the security protocol and algorithm in the sensor networks.

Sensors are inexpensive, low-power devices which have limited communication, energy, storage, and calculation resources. The conventional key distribution scheme can not meet the needs of sensor networks, for example, the public key cryptosystem can not be applied due to its complicated calculation. The random key pre-distribution becomes the main scheme of sensor network key distribution. Cur-
Currently, the study on the random key pre-distribution focuses on increasing sensor network security connectivity and reducing costs. However, in the case that several nodes are captured by the attackers, the random key pre-distribution is seriously threatened. Increasing the resilience against node capture is one of the main aspects in sensor network random key distribution security. This paper analyzes and compares in depth the resilience of the basic random key pre-distribution, Q-composite random key pre-distribution and its reinforcement. The study shows the tradeoff between the resilience, security connectivity and costs in sensor networks. It is hoped to choose and design a key pre-distribution scheme with greater resilience for a specific application environment and increase the availability, the scalability and the survivability of sensor networks.

II. SENSOR NETWORK KEY PRE-DISTRIBUTION, THREAT MODELS AND RESILIENCE

2.1 Sensor network key pre-distribution

Key distribution is the foundation of the confidentiality, integrity, non-repudiation, authentication, and authorization of sensor networks. However, the random key pre-distribution becomes the main way of sensor network key distribution for it is difficult to predict the deployment location of sensor nodes. Original solution is provided by Basic probabilistic key pre-distribution scheme which relies on probabilistic key sharing among the nodes of a random graph [3]. There are several key reinforcement proposals to strengthen the security of the established link keys, and improve the resilience. Objective is to securely generate a unique link- or path-key by using the established keys, so that the key is not compromised when one or more sensor nodes are captured. One approach is to increase the numbers of key overlap required in shared-key discovery phase. Q-composite random key pre-distribution scheme requires q common keys to establish a link key [4]. Cluster key grouping scheme proposes to divide key-chains into several clusters [5]. Similar mechanism is proposed by Pair-wise key establishment protocol which uses threshold secret sharing for key reinforcement [6]. Pietro et al proposed Cooperative pair-wise key establishment protocol [7]. Key pre-distribution by using deployment knowledge scheme uses location information [8].

2.2 Threat models

All the security systems and algorithms of sensor networks are public, only keys and security materials are confidential. Sensor networks are usually deployed in the unattended and public areas, and they do not have the tamper-proof function, so physical attacks occur easily. The attackers can capture several sensor nodes, and extract the keys and the security materials stored in the captured nodes. It is impossible for the attackers to capture multiple nodes at the same time. The attackers can make use of multiple captured nodes to analyze collaboratively and attack sensor networks.

2.3 Resilience against node capture

Resilience against node capture refers to the probability that the attackers can encrypt the messages transmitted between the uncompromised nodes using the keys and security materials stored in the captured nodes. The lower the resilience, the more difficulty the attackers make use of the security materials stored in the captured nodes to attack the other parts of networks. The more nodes the attackers capture, the more security materials were exposed, the greater the resilience, and the more dangerous the networks.

III. RESILIENCE ANALYSIS ON RANDOM KEY PRE-DISTRIBUTION SCHEME

3.1 Basic random key pre-distribution scheme

Eschenauer et al proposed the basic random key pre-
distribution of sensor networks [3]. This mode consists of three phases, namely key pre-distribution, shared-key discovery, and path-key establishment. First, a large key pool S is generated, and each node chooses k keys from the key pool to constitute the key-chain before the node is deployed. The adjacent nodes with shared-keys can establish security links by means of the shared-keys. In the random key pre-distribution scheme, the probability of key share among two sensor nodes becomes

\[ P = 1 - \frac{((|S| - k)^2)}{|S|(|S| - 2k)!} \]

Since |S| is very large, we use Stirling’s approximation for \( n! \approx \sqrt{2\pi n} n^{n+1/2} e^{-n} \), to simplify the expression of \( P \), and obtain: \( P \approx 1 - \frac{(1 - k/|S|)^{(|S|-k)/2}}{(1 - 2k/|S|)^{(|S|-2k)/2}} \). Its security connectivity is shown in Fig. 1:

Fig. 1 Connectivity between the adjacent nodes in basic random pre-distribution scheme

The probability that a link is compromised, when a sensor node is captured, is \( k/|S| \) which is very high for small key-pools and long key-chains, and produces low resilience. The probability that a given key has not been compromised, when \( x \) sensor nodes are captured, is \((1 - k/|S|)^x\). Therefore the probability that a link between normal sensor nodes is compromised is \( 1 - (1 - k/|S|)^x \). The resilience against node capture of basic random key pre-distribution is \( RES = 1 - (1 - k/|S|)^x \).

The resilience of basic random key pre-distribution is shown in Fig. 2. Under the same conditions, the larger the size of the key pool, the smaller its resilience value; the fewer the keys stored in the captured nodes, the fewer the keys exposed and the smaller the resilience value. But, the larger the size of the key pool, the fewer the keys stored, and the lower the connectivity. The network resilience, the security connectivity and the storage costs of nodes should be taken into full consideration and kept balance when designing the protocol and the algorithm for a particular sensor network application environment.

Suppose that in the sensor network application environment, the critical value of the probability the network links allowed to be compromised is \( C \), so
Fig 3 Captured numbers of nodes allowed by the basic random key pre-distribution

3.2 Q-composite key pre-distribution mode

Chan et al presented Q-composite key pre-distribution to strengthen the security of the established link keys, and improve resilience [3]. Objective is to securely generate a unique link- or path-key by using established keys, so that the key is not compromised when one or more sensor nodes are captured. One approach is to increase the numbers of key overlap required in shared-key discovery phase. Q-composite random key pre-distribution scheme requires $q$ shared-keys to establish a link key. If there exist $q$ shared-keys $k_1, k_2, \ldots, k_q$, between the adjacent nodes $A$ and $B$, $q \geq Q$, link key $k_{A,B}$ is set as hash of all common keys $k_{A,B} = \text{Hash}(k_1||k_2||\ldots||k_q)$. In random key pre-distribution, the probability that any two nodes have exactly $i$ keys in common is

$$P(i) = \binom{|S|}{i} \binom{|S| - i}{2(k-i)} \frac{\binom{k}{i}}{\binom{|S|}{k}}$$

and hence the probability of any two nodes sharing sufficient keys to form a secure connection is

$$P = \sum_{i=q}^{\infty} P(i) = 1 - \sum_{i=q}^{\infty} P(i).$$

We study that the Q-composite random key pre-distribution scheme strengthens the network’s resilience against the node capture. Let the numbers of captured nodes be $x$. Since each node contains $k$ keys, the probability that a given key has not been compromised is $(1 - k/|S|)^x$. The expected fraction of total keys compromised is thus $(1 - k/|S|)^x$. For any communication link between two nodes, if its link key was the hash of $i$ shared-keys, then the probability of that the link being compromised is $(1 - (1 - k/|S|)^x)^i$. The probability of setting up a secure link is $P = \sum_{i=q}^{\infty} P(i)$. Hence, the probability that any secure link setup in the key-setup phase between two uncompromised nodes is compromised when $x$ nodes have been captured is

$$\text{RES} = \sum_{i=q}^{\infty} \left(1 - (1 - k/|S|)^x\right)^i P(i).$$

The resilience value of Q-composite random key pre-distribution is shown in Fig. 4. The resilience in this mode is better than the one in the basic random key pre-distribution mode. However, several shared-keys are required to establish secure connections; as a result, its security connectivity is lower than that in the random key pre-distribution scheme. Under the same conditions, the larger the size of the key pool, the smaller its resilience value; the fewer the keys stored, the fewer the numbers of keys exposed and...
the smaller the resilience value after nodes being captured; the larger the value of $Q$, the more shared-keys are required to establish connections, the smaller the resilience value. But the larger key pools, the fewer keys stored by nodes and the greater $Q$ will decrease the network security connectivity.

Suppose that in the sensor network application environment, the critical value of the probability the network links allowed to be compromised is $C$, hence $\sum_{i=1}^{k} \left(1 - k / |S|\right)^i \frac{P(i)}{P} \leq C$. Fig. 5 shows the maximum captured nodes allowed by $Q$-composite key pre-distribution in the limits of the corresponding critical value. The greater the value $Q$, the more

---

**Fig. 4** Resilience of $Q$-composite random key pre-distribution mode. The size of the key pool $|S|$ is 1000, 2000, 5000, 10000, and 100000 respectively, and the number of keys stored by nodes $k$ is 50, 100, 150, and 200 respectively, the number of common keys $Q$ is 1, 2, and 3, and the number of nodes captured is [0, 200]

**Fig. 5** Captured numbers of nodes allowed by $Q$-composite random key pre-distribution
the nodes allowed to be captured, but the lower the network connectivity.

3.3 Other key pre-distribution schemes

There are several key reinforcement proposals to strengthen security of the established link keys, reduce costs and improve resilience based on the basic random key pre-distribution.

Key pre-distribution by using deployment knowledge scheme [8] uses local information. The deployment knowledge is available a priori. The connectivity of sensor networks is improved. Cluster key grouping scheme [5] proposes to divide the key-chains into several clusters to decrease energy consumption in the shared-key discovery phase. Another solution is given by Pair-wise key establishment protocol [6] which requires every sensor node to have a unique ID which is used as a seed to a PRF. Key IDs for the keys in the key-chain of node SA are generated by $PRF(ID_a)$. Thus, broadcast messages carry only one key ID. Also, storage, which is required to buffer received broadcast message before processing, decreases substantially. But, a sensor node has to execute $PRF(ID)$ for each broadcast message received from a neighbor. These methods only reduce the communication and calculation in the shared-key discovery phase, and they do not increase the resilience against node capture. The resilience is the same as that in the basic random key pre-distribution.

At the cost of increasing communications, storage and calculation expenses, several reinforced key modes are proposed to improve the resilience of sensor network key pre-distribution. In Multi-path key reinforcement scheme [4], node $S_a$ generates $j$ random key updates $rk_i$ and sends them through $j$ disjoint secure paths. $S_b$ can generate reinforced link key $K'_{a,b} = K_{a,b} \oplus rk_i \oplus \cdots \oplus rk_j$ upon receiving all key updates. This approach requires nodes $S_a$ and $S_b$ to send and receive $j$ more messages each of which carries a key update. Moreover, each node on the $j$ disjoint path has to send and receive an extra message. Similar mechanism is proposed by Pair-wise key establishment protocol [6] which uses threshold secret sharing for key reinforcement. $S_a$ generates a secret key $K'_{a,b}$, $j$-1 random shares $sk_i$, and $sk_j = K'_{a,b} \oplus sk_1 \oplus \cdots \oplus sk_{j-1}$. $S_a$ sends the shares through $j$ disjoint secure paths. $S_b$ can recover $K'_{a,b}$ upon receiving all shares. In Cooperative pair-wise key establishment protocol [7], SA first chooses a set $C = \{c_j, c_j, \cdots c_m\}$ of cooperative nodes. A co-operative node provides a hash $HMAC(K_{c,b}, ID_a)$. Reinforced key is then $K'_{a,b} = K_{a,b} \oplus (\oplus_{c_j \in C} HMAC(K_{c,b}, ID_a))$ where $K_{a,b}$ and $K_{c,b}$ are the established link keys. Node $S_a$ shares set $C$ with node $S_b$; therefore, $S_b$ can generate the same key. This approach requires nodes $S_a$ and $S_b$ to send and receive $c$ more messages. Moreover, cooperative nodes have to send and receive two extra messages. In addition to increased communication cost, each cooperative node has to execute $HMAC$ function twice for $S_a$ and $S_b$. The key reinforcement solutions in general increase processing and communication complexity, but provide good resilience in the sense that a compromised key-chain does not directly affect security of any links in the WSN. But, it may be possible for an adversary to recover initial link keys. An adversary can then recover reinforced link keys from the recorded multi-path reinforcement messages when the link keys are compromised.

These reinforced key modes make the link keys between nodes independent of each other through corresponding calculation and collaboration, that is, capturing several nodes has no influence on other links. Consequently, the best resilience follows. However, in the application, increasing network communications and calculation expenses are not permitted. It is necessary to design the sensor network key distribution protocol and algorithm according to the specified application environments and take the resilience, security connectivity, and expenses into full consideration to arrive the best
security and performance balance.

3.4 Comparison of the resilience in the key pre-distribution schemes

The random key pre-distribution scheme is the main way in the sensor network key distribution. The resilience against node capture is the important aspect to consider when designing sensor network security protocol and algorithm. The sensor network key distribution schemes now and their reinforced modes are shown in Table 1:

<table>
<thead>
<tr>
<th>Key pre-distribution schemes</th>
<th>Resilience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic probabilistic[3]</td>
<td>1 - (1 - k/</td>
</tr>
<tr>
<td>Q-composite random[4]</td>
<td>\sum_{i=1}^{p} \left( 1 - \frac{k}{</td>
</tr>
<tr>
<td>using deployment knowledge [8]</td>
<td>1 - (1 - k/</td>
</tr>
<tr>
<td>Cluster key grouping [5]</td>
<td>0</td>
</tr>
<tr>
<td>Pair-wise key establishment [6]</td>
<td>0</td>
</tr>
<tr>
<td>Multi-path key reinforcement [4]</td>
<td>0</td>
</tr>
<tr>
<td>Pair-wise with threshold [6]</td>
<td>0</td>
</tr>
<tr>
<td>Cooperative pair-wise [7]</td>
<td></td>
</tr>
</tbody>
</table>

IV. CONCLUSION AND FUTURE WORK

Since sensor networks are usually deployed in the unattended open areas, the nodes are captured with ease by the attackers. The resilience against node capture is one of the main indicators in measuring sensor network key pre-distribution security. The mode of sensor network attacks against node capture is presented, and its resilience is defined. Through the analysis on and the comparison between the basic key pre-distribution scheme, Q-composite key pre-distribution scheme and random key pre-distribution reinforced scheme, the tradeoff between sensor network resilience, the security connectivity and the costs is proposed. In the key pre-distribution, increasing the resilience means enlarging the key pools, reducing the numbers of keys stored by the nodes, and increasing the amount Q of shared-keys, but all these decrease the network security connectivity. Using the key reinforced mode to increase the key pre-distribution resilience means increasing the network communications and the calculation expenses.

The analysis and evaluation on the sensor network resilience against node capture is a new direction in the study on the sensor network security techniques. In the future we will design the security protocol and algorithm for the specified sensor network application environment, to increase the network reliability, scalability and survivability.

REFERENCES

11th IEEE International Conference on Network Protocols (ICNP’03). 2003:326-335 Atlanta, Georgia, USA.


BIOGRAPHY

Shen Yulong received the BS degree from the School of Science & Technology, Xidian University in 2002 and the MS degree in 2005. He is currently a PhD candidate in the School of Communications Engineering in Xidian University. His research interests include cryptography, wireless sensor networks and security.

Ma Jianfeng received his B. S. degree in mathematics from Shaanxi Normal University (Xi’an) in 1985, and obtained his M. E. and Ph. D. degrees in computer software and communications engineering from Xidian University (Xi’an) in 1988 and 1995 respectively. Since 1995 he has been with Xidian University as a lecturer, associate professor and professor. Prof. Ma is a member of the executive council of the Chinese Cryptology Society and a member of the Ministry of Education Expert Committee for Discipline Development (China). His research interests include information security, coding theory and cryptography.

Pei Qingqi received the BS degree from the School of Electronic Engineering, Xidian University in 1998 and the MS degree from the School of Science & Technology in 2004. He is currently a PhD candidate in the School of Communications Engineering in Xidian University. His research interests include computer security, wireless sensor networks and security.

*Foundation Items: Supported by Educational Innovation Fund for Graduates of Xidian University (Program No.05006); Natural Science Basic Research Plan in Shaanxi Province of China (Program No.2005F28); Xi’an Key Technologies R&D Program(GG06017).
Hierarchical Spectrum Sharing Networks

Jie Chen, Chuan Han, Chulin Liao, and Shaoqian Li
University of Electronic Sci. & Tech. of China

ABSTRACT

A spectrum heterogeneity analysis in the cognitive radio network is conducted in this paper. Subsequently, a spectrum-heterogeneity-based hierarchical spectrum sharing (HSS) network for cognitive radio is proposed. The corresponding method of classifying available spectrums and communication based on the proposed architecture is also presented. Based on the above network architecture, we propose a reference protocol architecture. Research on these protocol function blocks, such as spectrum sensing, spectrum manager, and so on, is conducted. Numerical results show that HSS can provide a considerable extension to available spectrums so that the spectral utility may be further improved.

Key words: cognitive radio, hierarchical spectrum sharing network, communication mechanism, protocol architecture

I. INTRODUCTION

Studies by the Federal Communications Commission (FCC) show that the allocated spectrum is underutilized\(^1\). Temporal and geographical variations in the assigned spectrum utilization range from 15% to 85%. The spectrum usage inefficiency and continuously increasing demand for electromagnetic resources suggests secondary usage of spectrum, which allows opportunistic access to the licensed spectrum or the unlicensed spectrum temporarily available for commercial purposes. This technology is known as cognitive radio that enables networks to use or share spectrum dynamically. Spectrum bands on which secondary users can share with primary users are defined as available bands\(^2\). Primary users refer to licensed users; whereas secondary users are defined as users that opportunistically access the licensed or unlicensed spectrum.

Cognitive capability and reconfigurability are the two main characteristics of the cognitive radio technology\(^3\). The cognitive capability is defined as the capability of real time interaction with its environment. It includes three main steps: spectrum sensing, spectrum analysis, and spectrum decision. Spectrum sensing refers to monitoring the available spectrum bands, capturing their information, and then detecting spectrum holes which are referred to as temporally “unused” spectrum (note that the unused spectrum means spectrum used by secondary users without harmful interference to primary users). Spectrum analysis refers to estimating the characteristics of the spectrum holes that are detected through spectrum sensing. Spectrum decision refers to determining the appropriate spectrum band according to spectrum characteristics and user requirements. Reconfigurability is the capability of adjusting communication parameters for the transmission on the
fly without any modifications to the hardware components.

Since the basic concept of cognitive radio is to opportunistically use available spectrum bands, dynamic use of the spectrum brings about some problems such as interference and spectrum heterogeneity problems. Spectrum heterogeneity refers to spectrum bands availability varying with both location and time. The interference avoidance and spectrum heterogeneity characteristic are essential in implementing cognitive radio networks.

Due to particularity in interference and spectrum heterogeneity problems for cognitive radio, its architecture is different from other architectures in terms of the networking paradigms and architectures. Moreover, because of the spectrum availability fluctuation in both location and time, the spectrum efficiency will not be effectively improved, if the cognitive radio network is deployed as the conventional network architecture. Based on the above consideration, we proposed a novel hierarchical spectrum sharing (HSS) network for the cognitive radio network.

The remaining part of the paper is organized as follows. A thorough survey of existing cognitive radio architectures and problems are presented in Section II. In Section III, a hierarchical spectrum sharing network architecture and the method of classifying available spectrum are proposed. Communication based on this network is described in Section IV. The reference protocol architecture and our research on its protocol function blocks are discussed in Section V. Numerical results are used to evaluate the performance of HSS in Section VI. Finally, we conclude the article in Section VII.

II. MOTIVATION

Cognitive radio network allows secondary users (or CRUs) to make use of licensed spectrum to avoid harmful interference to primary users or unlicensed spectrum. In cognitive radio networks, secondary users (or CRUs) can automatically detect spectrum signatures of primary users, and use or share the spectrum opportunistically. However, dynamic use of the spectrum brings about two challenges. One is avoidance of interference to primary users or other secondary users; the other is how to deal with spectrum heterogeneity.

While coexistence between the primary user and the secondary user (or CRU) and self-coexistence among secondary users is a critical issue for implementing the cognitive radio network, interference is an important factor to be considered. To measure and manage the interference, FCC proposed the interference temperature (IT) model in which secondary users (or CRUs) are permitted to share the licensed spectrum if the interference does not exceed the interference temperature limit\(^4\). The interference temperature limit is defined as an upper bound or cap on the potential RF energy that could be introduced into a certain spectrum band. In most currently available literature, the interference temperature is approximately denoted by the signal-to-noise ratio (SNR) for simplicity. CRUs can make a secondary use of the spectrum by transmitting in a shared spectrum provided that the noise or interference caused by the transmission is below an acceptable threshold of SNR.

Furthermore, the maximum distance at which data can be received without errors is dictated by the acceptable signal-to-noise ratio (SNR) at a particular location. Herein, signal power is attributed to the primary users and everything else is characterized by the noise.

As introduced in Section I, the spectrum heterogeneity problem refers to spectrum availability variations for different secondary users due to location and time differences. More specifically, mobility and traffic variations of primary users also result in available spectrum fluctuation with both location and time at secondary users (or CRUs). In addition, interference constraints and reward obtained on each spectrum band could be different due to non-uniformly partitioned spectrum bands, differences in power constraints and associated technologies.
III. HIERARCHICAL SPECTRUM SHARING NETWORK

Because of cognitive radio spectrum heterogeneity, the available spectrum for one CRU may be unavailable to another CRU, and different transmission power may bring out diversity for available spectrum. Generally speaking, the coverage radius of CRBS is larger than the distance between a CRU and its neighbor, therefore the communication between CRBS and CRU may require higher transmission power than the communication between CRU and its neighbor. Consequently, quite a number of bands which are not suitable for high power transmission between CRBS and CRU may be available for CRUs to transmit signals to their neighbors, especially in a CR wide area network, such as IEEE 802.22 system. Moreover, available bands for the low-power sharing are more plentiful than those for high-power sharing, because cognitive radios with low-power sharing introduce lower interference to primary systems through the transmission power restriction. A network which enables direct communication between CRUs with low-power sharing would improve spectral utilization by exploiting the spectrum heterogeneity.

The HSS network architecture

Based on the above consideration, in order to extend available spectrum bands and gain a better spectral utilization, we propose a hierarchical spectrum sharing (HSS) network, in which different spectrum bands are used for different scenarios. HSS network can be deployed both as an infrastructure network and an ad hoc network. It seems this architecture is similar to the xG in Section II, but there are differences between HSS and xG. The main difference is the distributed behavior under centralized control and the notion of hierarchical spectrum sharing. The distributed behavior under centralized control is described in a latter paper, while hierarchical spectrum sharing will be introduced in the following subsection. The HSS network consists of many cells, and each cell is composed of one CRBS and many CRUs, as shown in Fig.1. Note that the scenario of opportunistic usage of spectrum is also described in Fig.1. The components of HSS network are as follows:

- Cognitive radio users: CRUs with hierarchical spectrum sharing capabilities are allowed to use the licensed or unlicensed spectrum only in an opportunistic manner.
- Cognitive radio base stations: CRBSs are fixed infrastructure components with hierarchical spectrum sharing capabilities. CRBSs provide single hop connection to CRUs. A CRU can access other cells or networks through this connection.

There are two communication modes — the point to multi-point mode and the ad hoc mode in the HSS network. CRUs can either communicate with each other in a multi-hop manner or access the base station. Different communication modes adopt different available spectrum bands. Available spectrum bands in this network are classified into two categories: TYPE I and TYPE II. TYPE I is generally with high power sharing and used for direct communication between CRUs and CRBS, while TYPE II is generally with low power sharing and used for direct communication among CRUs. In other words, different types of available spectrum can be used in different scenarios in HSS networks. The method of classifying available spectrums is described in the next subsection.

---

*Fig.1 Hierarchical spectrum sharing network*
Method for classifying available spectrum bands

In the CR network, whether a spectrum band is available for CRUs depends on the allowable transmission power, channel conditions, environmental noise and PHY techniques such as modulation and coding. Therefore, the method of classifying available spectrum bands is essential in implementation of the HSS network. In our scheme, the criterion of identifying available spectrum is the SNR value, which is restricted by the interference temperature and interference temperature limit. Spectrum band availability is determined according to the interference temperature model, path loss and noise environment. Characteristics of different spectrum bands are derived through cognitive capability, and usable spectrum bands can be obtained consequently in the classification phase which is based on spectrum heterogeneity at different users. The method for classifying available spectrum bands in the HSS network consists of three main steps:

Firstly, the maximum allowable received power of a node (note that the node can be a CRU or a CRBS) on a certain spectrum band is obtained based on the interference temperature model.

Secondly, the actual transmission power at each transmitter is computed using the received power obtained in the previous step.

Thirdly, the SNR at the node is acquired according to the actual transmission power obtained in the second step. Then the node can judge whether a certain spectrum band is available for transmission between this node and its surrounding nodes by comparing the SNR with the minimum SNR for reliable transmission. If the former is higher than the latter, the two nodes can communicate directly on this spectrum band, i.e., this spectrum band is available for the specific link.

Let us use an example to explicitly illustrate the process. Suppose that node 2, node 3, node 4 and node 5 are distributed around node 1, as shown in Fig. 2. For link 1, node 2 is the transmitter while node 1 is the receiver. If SNR of node 1 on a certain spectrum band is higher than the required minimum SNR, it indicates availability of this spectrum band for link 1. But for link 2, where node 1 also acts as receiver, if SNR of node 1 on this spectrum band is lower than the required minimum SNR, this spectrum band is unavailable for link 2. So do link 3 and link 4. That is to say, the availability of a spectrum band for a node depends on the node itself as well as its communication counterparts. Or the spectrum availability is attributed to a certain link.

Through the above spectrum analysis process, we can acquire a set of available spectrum for a certain node. If the counterpart with which the node can communicate directly is a CRBS, this available spectrum band is classified as TYPE I. If the communication counterpart is a CRU, this available spectrum band is classified as TYPE II. Available spectrum bands of the node may be TYPE I, TYPE II, or both. So the node can choose a type of available spectrum bands according to its communication counterparts. In other words, different types of available spectrum can be used in different scenarios. Consequently, available spectrum bands are extended to a certain extent.

IV. COMMUNICATION MECHANISM BASED ON HSS NETWORK

A general communication mechanism is not applicable to HSS networks. Therefore, in this section, we propose a selection rule for the communication
scheme according to the traffic services property. In a hierarchical spectrum sharing network cell, multiple CRUs are managed by a single CRBS which operates as the main controller and access point to the core network. In a certain cell, CRUs are capable of communicating with both CRBSs and other CRUs directly using TYPE I bands and TYPE II bands, respectively. Four different kinds of communication scheme are applicable to HSS due to different traffic services. Traffic services are simply classified into real-time services with strict delay requirements and non-real-time services. The details are listed as follows:

- **Scheme 1**: If the destination node (DN) belongs to the neighbor set of the source nodes (SN), it chooses direct communication that is not necessarily transferred by CRBSs no matter the service is real-time or non-real time. The direct communication between two CRUs reuses TYPE II spectrums in a single-hop way, as showed in Fig.3 (a).

- **Scheme 2**: If a DN does not belong to the neighbor set of the SN but in the same cell as the SN, and the communication between them is not very urgent, they communicate in a multi-hop way without CRBS participation and other CRUs in the same cell can be used as relay nodes. The multi-hop communication among CRUs reuses TYPE II spectrums, as showed in Fig.3 (b).

- **Scheme 3**: There are two cases in this scheme. One case is that the DN is not in the same cell as the SN, such as the traffic that connect to the core network; the other case is that the DN is in the same cell as the SN but not in the neighbor set, and the traffic is urgent. CRUs communicate with CR-BS directly by reusing TYPE I spectrums, as showed in Fig.3 (c).

- **Scheme 4**: In this case, all the CRUs in the cell can be regarded as relay nodes for transferring data on the TYPE II spectrum. The SN constantly searches a route until the end node of the route path is able to directly connect CR-BS with the TYPE I spectrum. This scheme works when the DN is not in the same cell as the SN. The SAP communicates with the CR-

---

**Fig.3 Communication based on HSS**

BS in a multi-hop way, as showed in Fig.3 (d).

**V. REFERENCE PROTOCOL ARCHITECTURE**

In general, the major goals of defining suitable reference architecture for cognitive radio are to guarantee both flexibility and efficiency. Based on the HSS network, we proposed the reference architecture model depicted in Figure 4.

---

**Fig.4 Reference Protocol architecture**

As shown in the figure, new functionalities are required in the proposed HSS network. In summary, the main functions for HSS network are listed as follows:

**Spectrum sensing**

An important requirement of the proposed HSS network is to sense the available spectrum. The spectrum sensing function enables the HSS network to adapt to its environment by detecting available spectrum. The most efficient way to detect available
spectrum is to detect the primary users that are receiving data within the communication range of CRU. Generally, the spectrum sensing techniques can be classified as transmitter detection, cooperative detection, and interference-based detection [3]. Based on research on the existence spectrum sensing techniques, we propose a new distributed sensing scheme by considering the reliability of local spectrum sensing. We quantify the channel condition between the primary user and CRU with a parameter called “credibility”, and the information gathered at CRBS is made up of two parts: decision of each CRU and its associated credibility. To effectively combine these results from different CRUs, we apply Dempster-Shafer’s (D-S) theory of evidence to make a final decision at the CRBS [5]. Simulation results show that significant improvement in detection probability as well as reduction in false alarm rate is achieved by our proposal.

**Spectrum manager**

For the proposed HSS network, the main functions of spectrum manager refer to allocating available spectrum resource to CRUs and managing utilization of spectra. In other words, spectrum manager classifies the available spectrum based on the results of spectrum sensing, and then allocates the different types of spectra to different CRUs based on four communication schemes. Two key goals of spectrum allocation algorithms in HSS network are spectrum utilization and fairness. Specific combinations of these two goals form different utility functions, which tradeoff spectrum utilization and fairness. A parallel allocation algorithm is proposed, which is a modification of CSGC (Color Sensitive Graph Coloring) algorithm [6]. Under the constraint of maximizing system utilization, the parallel algorithm obtains the same allocation matrix as CSGC, while reducing the allocation period, so that it can be adapted to the agile sense requirement of cognitive radio.

**Data transmission**

HSS network is required to optimally use spectrum resource as determined by the spectrum sensing and channel estimation. Orthogonal Frequency Division Multiplex (OFDM) has been the main trend of PHY transmission techniques for data transmission due to its advantages, such as mature techniques, multi-carrier characteristics and high spectrum efficiency. While it is a promising transmission technique for data transmission, we proposed OFDM-based transform domain communication system (TDCS) as a promising signaling transmission scheme in the HSS network.

In our work, we rebuilt the signal model of TDCS, proposed the OFDM-based TDCS, interleaved OFDM-based TDCS, and soft demodulation algorithms of OFDM-based TDCS signal [7][8][9]. Both the analytical and numerical results show that OFDM-based TDCS has low implementation complexity and can operate well in low SNR regions (-20dB in the IEEE 802.22 Profile C channel). More specifically, the interleaved OFDM-based TDCS with hard demodulation serves as a candidate for applications with stringent requirements on implementation complexity and loose requirements on emission power, while OFDM-based TDCS with soft-demodulation serves as a candidate for applications with stringent requirements on emission power and loose requirements on implementation complexity. Together with other inherent characteristics, such as compatibility with common OFDM schemes, low interception probability properties, low data rate and longer coverage than UWB, OFDM-based TDCS is a promising signaling transmission scheme in cognitive radio.

**Channel estimation**

The objective of channel estimation is to acquire the knowledge on the channel impulse response (CIR) in order to accurately detect the received signals. The most popular method to realize channel estimation is based on pilot transmission, which involves periodic transmission of a pilot (training sequence) known to the receiver. Firstly, the receiver obtains the channel state information at the pilots, and then recovers the channel state information at all symbols using the coherent characteristics of the radio channel. Pilot
assisted channel estimation applied to OFDM-based TDCS, semi-blind channel estimation including kalman filter, particle filter and mixture kalman filter, kalman filter channel estimation applied to space time block coding (STBC) system are discussed. We also present how to predict channel capacity according to channel coefficients and how to detect interference in the slow fading channel scenario. Several channel estimation methods based on noise reduction are proposed for OFDM-based TDCS, which performs well when the signal to noise ratio (SNR) is very low. In time domain, time moving average, time forgetting average and the combination of time average and time forgetting are proposed to eliminate the impact of noise by using the slow variation characteristic of slow fading channel. In IDFT transform domain, low-pass filter based method is applied [10].

**Access**

The access function aims to break the spectrum access barrier and enable networks and CRUs to opportunistically access spectrum. In HSS network, an appropriate spectrum etiquette, which is a set of rules regulating access to spectrum and its usage, is proposed to avoid interference and collision. The spectrum sharing scenario is currently characterized by a complete lack of mutual awareness of users with potentially competing needs. And with the increase in complex service requirements, the problem cannot be solved adequately by the previous spectrum etiquette. Based on the above motivation, a cooperative spectrum etiquette for HSS network is proposed [11]. The basic concept of cooperative spectrum etiquette is that users of the spectrum intended to share the related information depending on certain conditions. In our network model, a user can cooperate with other users, thus the behavior of cooperation is equivalent to select less aggressive MAC parameters and is to benefit a user in case of cooperating users. Before a CRU can be serviced by a CBS, it needs to enter the network and negotiate its capabilities with the CBS; this may include many tasks between the CRU and the CBS. More importantly, during this process the CRU needs to ensure that its communication will not cause harmful interference to primary users. The CBS flexibly manages CRU, obtains a reliable spectrum occupancy map of its cells and, if necessary, changes its operating parameters. CRU also has many available ways to report measured information to the CBS.

Further, we proposed one initial link establishment algorithm based on OFDM-based TDCS [12]. This scheme can be used to implement protocols in both centralized and decentralized manners and can potentially solve the common control channel problem in any ad-hoc network. Our idea is based on the difference of spectrum masks at the transmitter and receiver. This type of discrepancy necessitates the exchange of sensing results. To fulfill the exchange procedure, a transmission technique which is of low emission power, a long transmission distance, and robust BER performance in low SNR regions is desired. OFDM-based TDCS is just the perfect candidate.

**Packet scheduling**

In order to let the resource allocation in cognitive radio systems be of the adaptive capability to varying radio resources, we propose an adaptive packet scheduling algorithm, which serves different traffic queues based on the QoS levels of each traffic queue and the variation of available spectrums [13]. Generally, demanding strict QoS guarantees, the real-time traffics have more impacts on the system performance than non-real-time ones. Hence, the proposed algorithm fits for the real-time traffics when there are not enough resources to maintain the QoS level of real-time traffics. Simulation results show that, compared with traditional packet scheduling algorithms, the proposed algorithm provides not only better QoS guarantees for heterogeneous cognitive radio traffics, but also higher system access capacity and spectrum efficiency.

**Power control**

Since the CRU needs to ensure that its communication will not cause harmful interference with primary
users, power control should be developed.

Routing

As mentioned in section III, HSS network can be deployed both as an infrastructure network and an ad hoc network. That is to say, in HSS network with multi-hop communication requirements, novel routing algorithms should be developed.

So far, our research concentrates on spectrum manager, access, spectrum sensing, packet scheduling, and channel estimation. Other issues will be researched in next phase.

VI. NUMERICAL RESULTS

In this section, we present some numerical results to compare the performance of HSS with that of other techniques, which can extend available spectrums, such as Directional Antennas (DA) with beamforming and Power Control (PC).

Consider the following network deployment. In a HSS cell, 48 CRUs are uniformly distributed. The location and coverage of a HSS cell and licensed systems which are licensed from channel CH1 to CH6 is shown in Fig. 5. Occupation time distribution of the six channels is shown in Fig. 6, where high level values mean the specific channel is occupied by the licensed system at the specific time slot. Four scenarios are simulated: the original scheme, DA, a combination of DA and PC, and HSS. The number of available spectrum bands in the above scenarios is shown in Fig. 7.

If a channel band is available for N users, then it will be calculated N times, which is reasonable because N users may share the band with multiple access technologies, such as TDMA and CDMA. As shown in Fig. 6, HSS can provide a considerable extension to available spec-
trums so that the spectral utility may be further improved. More specifically, compared with original network, there are more bands available for users if Directional Antennas technique is adopted, while Power Control contributes little. HSS is the best scheme in extending available spectrums within the discussed schemes.

**VII. CONCLUSION**

CR technology is considered as a solution to the spectrum underutilization problem. In this paper, we present the hierarchical spectrum sharing network, which can be deployed both as an infrastructure network and an ad hoc network. Four kinds of communication scheme can be realized. By adding the ad-hoc mode to a CR cell, CR users are capable of sharing the licensed spectrum on a low power level so that the spectrum efficiency can be further improved. Every node in a certain cell detects its interference temperature and estimates the channel state information (CSI) of every channel. Then it searches the best communication path and adopts four kinds of communication scheme introduced in this novel network. Simulation results in Section VI indicate that HSS can extend the available bands considerably. Based on the HSS architecture, a reference protocol architecture is proposed and subsequent research on function protocol blocks, such as spectrum sensing, spectrum manager and so on, is presented.

Although the available spectrum bands in the HSS network can be extended considerably and spectrum efficiency can be subsequently improved, dynamic usage of the spectrum in the HSS network brings about problems which need in-depth investigation. For instance, the CRU needs to change its operating frequency when current channel conditions become worse or new primary users appear. The criteria of ensuring such changes to work smoothly and timely to minimize the CRU performance degradation have not been thoroughly explored. Furthermore, in a HSS network with multi-hop communication requirements, unique characteristics of the spectrum necessitate new routing algorithms. The work in this article represents a first step for HSS networks. In future work, we will address more issues, such as spectrum handoff and routing.

**REFERENCES**


**BIOGRAPHIES**

**Jie Chen** (jiechen@uestc.edu.cn) has been an assistant professor at Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China, since 2004. She received her B.S.E.E. degree and M.S.E.E. in wireless communication engineering from CQUPT in 1994 and 2002, respectively. She is pursuing her Ph.D. degree at University of Electronic Science and Technology of China. Her current areas of interest include wireless communication system architecture and protocol, radio resource management and cognitive radio networks.

**Chuan Han** received his B.S.E.E. degree from University of Electronic Science and Technology of China (UESTC), Chengdu, in 2004. He is pursuing his M.S.E.E. at UESTC. His research is in the general area of wireless communications, including estimation and detection theory, information theory, wireless networking, cognitive radio PHY issues, and feasibility of cognitive radio.

**Chu Linliao** received his B.S.E.E. degree from University of Electronic Science and Technology of China (UESTC), Chengdu, in 2004. He is pursuing M.S.E.E. and researching at National Key Lab of Communication in UESTC. His current research focuses on broadband wireless access, MAC protocol design and spectrum management in cognitive radio networks.

**Shao Qianli** received his B.S.E degree in communication technology from Northwest Institute of Telecommunication (Xidian University) in 1982 and M. S.E. degree in Communication System from University of Electronic Science and Technology of China (UESTC) in 1984. Now, he is a Professor, Ph.D supervisor, and director of National Key Lab of Communication in UESTC, and he is a member of National High Technology R&D Program (863 Program) Communications Group. His research interests includes wireless information and communication theory, mobile and personal communication, anti-interference technology in wireless communications, spread-spectrum and frequency-hopping technology.
Security Measures Against CBRN Threats: Case Study Olympic Games

P. Stavroulakis, Professor, Technical University of Crete, Greece
pete-tsi@otenet.gr

ABSTRACT

In this paper we present how an serious security relevant event can be taken care of during Olympic Games. Remote healthcare treatment will be given to injured persons during nuclear radiological biological attacks or in the case of physical disasters. The "Emergency-112" wireless telemedicine platform provides the hardware and software infrastructures in order to cope with the most extreme scenarios. A fully autonomous mobile system interconnects a group of injured persons with the expert medical institution. The unlimited capabilities of the system allows the Emergency-112 platform to be used for rescue and surveillance operations regardless the terrain or the weather conditions. A dynamic hybrid system monitors the available spectrum and actively switches between different telecommunication access gateways.

Key words: emergency telemedicine, mobile and satellite comms, bandwidth allocation management, bioterrorism

I. INTRODUCTION

The terrorist attack in the World Trade Center in September the 11th (2001) brought forward numerous outstanding security issues that require special concern. National security agencies analyzed the nature of terrorism and published several reports. The latest developments shows that Terrorists prefer the so called "blind hits" which are extremely well organized therefore result multiple civilian deaths. Each attack is giving a clear political and geostrategical message to certain countries that actively or silently participate in coalition forces around the world. The political and public community impact is huge, unpredicted and frequently creates political instability \cite{1}. These reasons make terrorist attacks to occur frequently during World-class events and Olympic games. Similar cases experienced in Munich and recently in Atlanta in 1996. Risk management strategies interpret and analyze previous attacks in order to develop customized solutions to overtake such events. The Emergency-112 telemedicine platform can be used for countermeasures against nuclear biological radiological or gas attacks. The proposed architecture is capable of all operations with in the mobile unit. Therefore the system can be used for rescue operations during physical catastrophes, that is earthquakes followed by lethal Tsunami waves. Beyond telemedicine and rescue operations the Emergency-112 unit can be used for surveillance and escort services for sea, air or ground based operations.

An intelligent spectrum management technique switches outgoing traffic between fixed and mobile,
broadband or narrowband telecommunication access gateways with regard to the availability. State of the art architecture allows emergency issues, which are encountered in rural and urban environments, to be resolved immediately as if they were regular incidents. The proposed implementation features vital electrophysiological telemetry in the parallel transmission with live video in environmental terrains that lack telecommunications infrastructure.

From 2001 the US Defence Threat Reduction Agency (DTRA) and the Federal Emergency Management Agency (FEMA) spent $1.1 billion in bioterrorism prompt expert computerized systems [2]. Greece in turn spent nearly 500 million Euros in the C4I Olympic security project. Mobile operation and command centers provided computer decision support systems (CDSS) in order to actively participate in the case of biochemical and radiological terrorism in games time. The experience in the field of risk management and in the security preparedness for the Olympics of 2004 makes the Emergency-112 a cost effective, customized, reliable and robust telemedicine solution for rescue and surveillance provision to the public, VIPs and executive members of the IOC and IAAF committees.

II. VULNERABILITY IN TELEMEDICINE SYSTEMS

A very important issue that makes telemedicine systems vulnerable is the operating environment with regard to the ether conditions. The terrain is a critical factor that limits the system capabilities because mountains, lakes, forests, the sea and the metro subway create dead areas that transeption is not possible therefore there isn't coverage as far as GSM, GPRS, G and the satellite are concerned. People who travel by boat or by airplane spend most of the journey with no coverage because when the ship/airplane departs the signal fades after 10 to 20 kilometers in the line of sight. Passengers in the metro suffer signal fading due the thick concrete wall construction e.g. in the Sarin gas attack in the Tokyo subway system in 1995 [3]. The second reason that remote interactivity is limited arisen due to the narrowband capabilities that mobile networks provide. Terrestrial infrastructures behave rather inconvenient due to increased probability for network unavailability from network traffic overloading in the event of physical catastrophes. The scale and the frequency of the events increased dramatically within the last decade. For example remember the earthquakes in eastern Turkey (1998 and 2002), Greece (2001), India and Pakistan (2002), Sri Lanka and Indonesia in December 2004 followed by huge Tsunamis. This results in population confusion and telecommunications and electricity networks break down. This was the case in 2001 Athens earthquake and in New Year's Eve in the millennium. Although the telecommunication networks were all in tact up and running civilians created traffic overloading and bottlenecks due to simultaneous telephone calls and SMS text messages. All mobile operators dropped off every telephone call request and they expunged every SMS message from the incoming and outgoing buffer gateways.

III. THE EMERGENCY-112 TELEMEDICINE FEATURES

The system is capable of prompt and expert medical care improving health care services at understaffed rural areas and out of coverage urban spots such as the metro rail stations. The fields of interest of this paper are Ambulances, Rural Health Centers (RHC), Ships navigating in wide seas, Airplanes in flight and other remote areas of interest that are common examples of possible emergency sites, while critical care telemetry and telemedicine home follow-ups are important issues of telemonitoring. To comply with different growing application fields we created a combined real-time and store and forward facility that consists of a base unit and a telemedicine-mobile unit. This integrated system can be used to:

- Handle emergency cases in ambulances, RHC, ships or airplanes by using the telemedicine unit at the patient - emergency site and the expert's medical consulting at the base unit.
Enhance intensive health care provision by giving a portable base unit to the ICU doctor while the telemedicine unit is incorporated with the ICU’s in-house telemetry system.

Enable home telemonitoring, by installing the telemedicine unit at the patient's home while the base unit remains at the physician's office or hospital.

Provide the hardware and software foundations to produce full laboratory biochemical analysis in outdoors and areas of special interest e.g. the subway.

The mobile unit is composed of rugged modular components that work under extreme weather conditions. The E-112 modular construction supports hot plug in addition to plug and play capabilities.

3.1 Emergency-112 mobile access gateway

The new architecture allows for simultaneous End user terminal operation. The system is composed of the primary unit, which behaves as an access gateway, and group of secondary devices that collect electrophysiological signals, transmit video, produce biochemical and gas analysis. The access gateway connects to a 2 Mbps satellite modem giving real time video streaming in the uplink and the downlink in addition to and biological signals monitoring. Figure 3 shows how the proposed satellite implementation achieves large-scale integration covering wide geographical rural environments that aren't covered from the present implementation.
The server which is embedded in the Emergency-112 primary unit generates multiple port connections in order to broadcast parallel videos, vital biological signals as well as additional information to different stations based on the classification given by the E-112 primary medical crew. The two Megabits per second satellite link provides the physical over the air (OTA) interface that connects the primary unit to the remote administration host.

Figure 4 simulates an underground indoor environment e.g. the metro subway in which groups of patients that are spaced apart but in relatively short distances create a WLAN regardless the terrain, the technology infrastructure or the line of sight. Broadband access in the local loop is achieved through the wireless Ethernet backbone where multiple users connect using the 802.11b/g standard. The E-112 primary unit requires an RJ-45 fast Ethernet plug to be installed in the areas of great concern e.g. departure platforms, the escalators and the exit. This contributes towards the generation of wireless "hotspots" and "hot areas" that provide broadband local access. A scenario like this is not far from reality; assuming a gas attack in the lower levels of the station; the primary E-112 is plugged directly to the Ethernet switch and the personnel that carry the secondary E-112 navigates in areas of high injury concentration. A mobile computer with a "Cepheid bioagent" detector scans the area for large-scale aerosol attacks and reports back to the server [5].

Two different user profiles are created, the administrator access gateway user and the user that transmits data on the fly to the server. Multiple transmissions can have multiple receivers due to the TCP/IP stack that takes over the procedure. The E-112 server performs all network related tasks, that is IP filtering, store and forward, routing, initiation and termination procedures, user access rights and gateway switch over selection.

3.2 Incident classification and priority allocation

End-user in the secondary unit provides information about the injury in order the ambulance crew to rate the severity of the emergency. Heavy injured patients will be classified differently and they will be given the highest priority for guaranteed data transmission. Active directories generate End user profiles so that a full record is maintained during and after the telemedicine treatment. An intelligent technique allows End users to generate alarms in case the patient's condition gets serious. Different levels of alarms update in regular intervals a database that maintains the patient's medical record. Remote physicians will log on to the primary E-112 server and a "push-pull" service will upload the patients profile through a secure multilevel strongly encrypted VPN connection [6]. Secondary users are given bandwidth based on the severity of the injury. An intelligent bandwidth allocation routine running in the primary
server, process parallel video transmissions and alters the bit rate respectively.

### 3.3 Video transmission

Live streaming can use UDP to compensate delays in live video transmission or transmit video over IP/TCP in store and forward for guaranteed delivery. If the primary crew decides that short video clips must be recorded from an injured patient although the emergency is given medium priority the server stores the videos in the hard drive. When the highest priority emergency is cleared then stored video transmission begins if there is remote request.

### 3.4 Picture quality and Bandwidth allocation

Video transmission dissipates most of the system bandwidth; therefore bandwidth saving countermeasures must be developed. The obvious solution is to prohibit parallel video transmissions. To undertake this problem the E-112 server degrades, in real-time, the video picture quality within predetermined limits so that region of interests can be clearly retrieved in remote locations. This technique minimizes video bandwidth consumption allowing for additional video streaming.

### 3.5 Access network switchover

One of the system novelties is the capability to monitor the frequency spectrum for active telecommunication infrastructures. The system regularly scans for active wireless access nodes, if a node is spotted then alerts the server administrator. When the signal becomes strong enough to exceed the minimum signal to noise ratio then a second alert is generated and informs the user that a connection can be achieved. The administrator either activates the line or discards the message however if more than one networks are available the administrator decides which of these networks are most suitable to use. Network selection depends upon the emergency status, if there is a life threatening injury the system decides to activate the satellite modem. If the patient's condition is serious but not critical then terrestrial telecommunication networks are chosen. The level of the emergency denotes the network that is best preferred. The best solution is the most cost effective option in terms of bandwidth availability and tariff charges.

The system uses the primary satellite connection in rural environments however in the case of large cities GSM, GPRS, UMTS, and WLAN connections are going to be used. The server unit performs all the necessary software routines in order to supply the system with the appropriate bandwidth. The server maintains the satellite for the downlink and switch between different uplink platforms. In the case of an emergency in the subway where the satellite is unavailable the RJ-45 Ethernet port is enabled and connects the server to the LAN. The bandwidth can be as much as the Ethernet hub can give.

### IV. TECHNOLOGY CONVERGENCE AND CONTRIBUTION

E-112 is a hybrid system capable of compensating difficulties regardless the geographical location. The system converges existing technologies delivering modular and robust medical services to mobile users in remote locations. The system provides increased immunity against physical and human interactions. The E-112 is a multi operational platform that can be used for medical support, for rescue, surveillance and defense applications such as Anthrax smoke detection and spays monitoring for aerosolised airborne bacterial spores [7]. The system in a later stage will be enhanced with a low power MAC control protocol providing wireless medical multisensor monitoring for wearable products [8].

The E-112 provides the hardware infrastructure to connect to every available public access network and if needed to government TETRA networks (Police and Fire department). The system works in standalone operation or as integral part of a greater turnkey solution. The modular implementation and the technology architecture allow the E-112 unit to operate in 24/7 basis and/or for redundancy purposes during life threatening conditions.
REFERENCES


[4] E. Kyriacou, S. Pavlopoulos, D. Koutsouris, "Multi-purpose HealthCare Telemedicine System with mobile communication link support", http://www.biological-engineering-online.cm/content/2/1/7


BIOGRAPHIES

Peter Stavroulakis received his BS and Ph.D. degrees from New York University in 1969 and 1973 respectively and his MS degree from California Institute of Technology in 1970. He joined Bell Laboratories in 1973 and worked until 1979 when he joined Oakland University in Rochester Michigan as an associate Professor of Engineering. He worked at Oakland University until 1981 when he joined ATT International and subsequently NYNEX International until 1990. In may 1990, he was elected at the Technical University of Crete (TUC) Greece as a full Professor of Electrical Engineering. He is a member of the Editorial Board of the International Journal of Communications systems, the International Journal of Satellite Systems and China Communications and has been a reviewer for many Technical International Journals. His research interests has been focused on the application of various heuristic methods on Telecommunications, including Neural Networks, Fuzzy Systems and Genetic Algorithms and also in the development of new modulation techniques applicable to Mobile and Wireless Systems. His current research has been in the application of Chaos and Interference Reduction Techniques in the design of secure communication systems.

OLYMPIC GAMES AND SECURITY NETWORK EXPERIENCE

Professor Stavroulakis as a technical Director of NYNEX for Europe was responsible for the Design of the Telecom and information Network of the Winter Olympics in France in 1986. He has co-organized an International conference held in CHINA in October 2005 which dealt with the application of Wireless Systems in large Telecom systems and was totally responsible for the Workshop on the Olympics Games Security Networks. He was also responsible for the techno-economic evaluation of the security network used for the Athens Olympics2004 leading a team of professionals. He is also expected to play a major role in the future applications of security networks using TETRA on a European basis because of his recognized expertise in this area. His upcoming book to be published by Springer titled: TETRA: A GLOBAL SECURITY TOOL is expected to be out in the spring of 2007.
Customized Biometric Architecture for Access Control in Stadiums based on Federated Identities

Christos K. Dimitriadis, Peter Stavroulakis, D. Polemi
University of Piraeus, 80 A. Dimitriou, 18534 Piraeus, Greece
cricodc@unipi.gr

ABSTRACT

This paper, presents a system called Athlos2, which implements strong access control for athletic events, enhancing the safety feeling of event spectators. Athlos2 integrates intelligent biometric access control systems and smart cards, under a protocol that complies with the specification of the Liberty Alliance project for federated identities. A pilot version of Athlos2 was deployed and tested in terms of acceptability, information security and performance.

Key words: physical access control, security, biometrics, athletic events

INTRODUCTION

Security, as a factor that influences the sports fan psychology, is of paramount importance for the success of athletic events. Modern biometric technologies provide enhanced security levels by introducing a new dimension in the authentication process called "proof by property". However, the design and deployment of a security architecture incorporating biometric technologies hides many pitfalls, which when underestimated can lead to major security weaknesses [1]. Although biometrics have been deployed in pilot systems for protecting access to athletic events in the past, no integrated solution has been proposed taking into account the related security standards and no complete studies ever proven the benefits of such deployments.

This paper, describes a system called Athlos2 that implements strong access control for athletic events. The system integrates intelligent biometric access control systems and smart cards, under the specifications of the Liberty Alliance project for federated identities. Since security and user acceptance are contradictory factors in such applications, the pilot version of Athlos2 was deployed and tested in terms of these two aspects.

The paper is organized in eight main sections, excluding the introduction and conclusions. The first main section presents the state of the art regarding access control and monitoring in athletic events, while the second presents the underlying technologies of Athlos2. The third and fourth main sections present the application scenarios and the system architecture. Sections five to nine, present the testing environment and the results of acceptance and security evaluation.
II. SECURITY IN ATHLETIC EVENTS

During the recent years, the traditional stadium access control measures, such as ticket checking from security officers, have been enhanced by technological security measures such as ID cards for season-ticket holders, CCTV camera systems, RFID smart card technologies and biometrics.

During the Athens 2004 Olympic Games, almost 70000 security personnel was overseeing the event, with the help of technology. More than 1,100 poles topped with video cameras, speakers and microphones created a distributed net of surveillance posts aimed at locating disturbances quickly [2]. Barcode scanners and ID cards allowed athletes and trainers into the Olympic Village. In Sydney 2000 Olympic Games, a security system integrated with intelligent camera functions was deployed, in order to provide security, surveillance and access control [3]. The systems consisted of the combination of security, CCTV Switcher, Smart Card Access Control and Photo Identification Systems and provided a total solution to monitor and report on all activities. Furthermore, in the Commonwealth 2002 Games in Manchester, a security system protected almost 6,000 athletes and officials representing 72 countries and territories [4]. The system involved the installation of a sophisticated CCTV system that included 79 cameras in the athletics stadium, which enabled Greater Manchester Police to zoom in on every single person in attendance. There were also installed an access control system with intruder alarms, fire alarms and an emergency telephone network in the main stadium. The Millennium Stadium also completed a £2.8 million project, to supply and install systems for crisis management such as fire detection, security and CCTV and PAVA (Public Address Voice Alarm), as well as a system for the distribution of radio, television, data and telephone signals [5].

Biometrics and smart card technology is widely used during athletic events of known stadiums inside the UK. Manchester City Football Stadium, Crystal Palace, England Rugby Supporters, Chelsea and Bolton Wanderers have come up with a high-tech way to profile their sport fans and ccredited persons in an attempt to drive revenue, improve the game environment and provide the greater security in order to better control the flow of crowds from possible crisis situations around and in the stadiums. This was a centralized authenticationsolution, using personalised smart cards. Similar systems have been adopted by the Belgian Football and PSV Eindhoven Stadiums. In the Cricket World Cup (South Africa, 2003) bar coded tickets were deployed, using a two-dimensional barcode, which cannot beduplicated or forged [6]. The system handled 825,000 ticket sales. The bar code allowed for scanning and verification through a sophisticated venue access control system, which in turn generated a customer database holding valuable information on all ticket purchasers. In addition, all stadiums were monitored with CCTV, (eight cameras per event) and had full digital recording facilities.

The various systems that were implemented proved that technology consists of an integral part in the athletic events. Sch systems fulfil the requirements of the organisers, but there are not always effective and efficient in large-scale athletic events, mainly because there is not a tested integrated system for strong authentication. In 2002 World cup in South Korea, all stadiums were monitored with CCTV cameras. A problem occurred causing delays and many fans were unable to enter in time. It was recognized that such incidents wouldn't have occurred if an effective access control system was developed [7].

III. UNDERLYING TECHNOLOGIES

Biometrics is the core technology of Athlos2. The biometric technology has been recognized as a key technology for improving security and trust in different fields of modern society [8]. Biometrics are defined as the automatic use of human physiological or behavioral characteristics to determine or verify an identity [9]. The system conducts a measurement of the features of the user, encodes the data creating a template and compares it against a physical mea-
surement from the user each time accessing the system is attempted. The most widespread biometric technology in today's markets is fingerprint recognition [10]. The sensor's size is conveniently small (area of a few square centimetres, thickness of a few millimetres), enabling easy incorporation into any fixed or mobile terminal and the weight of the sensor is negligible. Reusability on a wide scale is possible through the use of different encodings and undergoes continuous improvement as standardisation is gradually taking effect. Fingerprint recognition systems fit quite well as an integral part of any fixed or mobile terminal. For all the above reasons, fingerprint technologies have achieved the dominant position in the year 2005 in terms of total revenue, achieving approximately 48% of the total biometric market [10]. The biometric component of the system takes into account all relevant aspects including technological, societal and legal issues. More specifically, security, performance, privacy, standardization, scalability, responsibility, interoperability, usability, acceptability and liability issues, were studied, targeting to the development of a biometric component that meets all necessary state-of-the-art specifications. This was accomplished by the exploitation of results of research project, such as FP6-001766 (BIOSEC) "Biometrics and Security".

The smart cards that use contacts are in line with the guidelines determined in the Standard of ISO 7816 Part 1. The reliability of these smart cards has been improved constantly during the previous years, because of the increasing experience in the manufacturing of such cards. On the other hand, the contacts remain one of the more frequent sources of problems in electromagnetic systems. For example, some problems can result from the attrition of contact. Since the contacts, placed in the surface of card are connected immediately with the inputs of the integrated circuit that is incorporated in the card, there is a danger of damage or even destruction of the integrated circuit from the electromagnetic discharges - load of enough thousands of volts is not infrequent. These technical problems are overcome with the contactless (wireless, RFID) smart cards. Apart from its technical advantages, the wireless technology offers also to the issuer and the holder of the card some interesting new applications [11]. For example, the contactless cards do not need to be imported essentially in a card reader, since there are RFID reading systems that function in a distance of up to one meter. This is a big advantage in access control systems where a door or a circular gate should be opened, since the granting of access of an individual can be checked without the requirement of the card to be removed from the wallet or the pocket and to be inserted into the reader. An extensive range of applications for this technology is the public transportation systems, in which a big number of passengers should be identified in a very short time interval. In addition, the wireless technology is suitable in systems that require the deliberate import of the card into a reader, since it is not important how the contactless card will be inserted into the reader. This is in contrast to case of the magnetic or smart cards with contacts, that function only if they are inserted in a consistent way. This freedom of the orientation restrictions simplifies the operation and increases the user acceptance [12]. Apart from the simplicity of use, this solution is attractive because it considerably decreases the danger of vandalism (for example, with the placement of chewing gum or glue in the slot of the reader). Up to now, the wireless cards have been mainly used for the public transportation systems, acting as electronic tickets. These systems currently employ single-use cards, that are cheap to develop. Nevertheless, there is an increasing demand for the incorporation of additional features into the electronic ticket. For this reason, the employment multi-use RFID cards with incorporated microprocessors will be increased in the near future.

IV. APPLICATION SCENARIO

A possible application scenario is the following:

The sports fans are provided with an RFID smart card including their access rights in a specific sta-
These users have to present their official identity documents to identify themselves to a registration authority. When the sports fan purchases a ticket for visiting an athletic event, their access rights are automatically updated for the specific time and space zone that the athletic event takes place. This step also includes a query in a biometric database to establish the uniqueness of the user's claim. If the application for registration is accepted and the user is identified as unique, the system creates a user record with the necessary information, including the unique identifier. Biometric measurements are conducted and a biometric template is created, realizing the biometric enrollment process. The biometric data are securely managed by official procedures based on international standards and best practices such as [18] or the upcoming standards of the ISO JTC1 SC37 on biometrics. An RFID smartcard is personalized and delivered to the user containing the user's biometric template, in symmetrically encrypted form, as well as a key pair (public and private key).

V. PROTOCOL DESCRIPTION

Athlos2 is compliant to the federated identity management principles of the Liberty Alliance project [17]. The purpose of using federated identities is basically system scalability and extensibility, since a central access repository (ticket provider) maybe needed to serve several stadiums. Security is another reason for using federated identities, since the Liberty Alliance specifications are designed in order to preserve the confidentiality of the user's identity, by deploying a secure token and temporal identity exchange mechanism. The following entities are involved:

- **User (or Principal):** The sports fan.
- **The Local Gate Access Controller (LGAC):** A system which controls access at a specific gate of the stadium.
- **Physical Access Server (PAS):** This is a centralized server which controls the gates of the stadium and provides authorization for access. This system acts as a service provider, in Liberty Alliance terms and should request identity services, in order to provide authorization or not to a specific sports fan.

- **Identity Provider (IdP):** A Liberty-enabled entity that manages identity information on behalf of the users and provides assertions of user authentication to a number of PAS. IdP hosts a database with the authorization attributes of the sports fans, which are updated per event. For example, for a specific event at a specific time and at a specific stadium, certain sports fans should have access. These access attributes are held by IdP and provided on demand.

The Liberty Alliance specifications [17], describe that the communication between the entities should be secured, proposing the use of certificates for encrypting and digitally signing the exchanged messages for preserving confidentiality and availability.

Athlos2 realizes a modular and distributed architecture, in order to achieve maximum interoperability, scalability and extensibility. The message sequence chart of Athlos2 is depicted in figure 1.
3. The PAS forwards to IdP, the UID and Suid, through another pre-established mutually authenticated SSL/TLS channel created between the PAS and the IdP.

4. IdP receives the UID and Suid, retrieves the users public key by using UID and validates the signature. If the signature validation is successful, the IdP sends the user's attributes to the PAS. The attributes describe access rights for a specific event at a specific time and for a specific gate of a specific stadium.

5. The PAS receives the attributes and if they correspond to its identity at a specific time, it provides authorization to the LGAC for providing access to the user.

VI. LGAC DESCRIPTION

We focus on the biometric process, presenting the entity that hosts it (LGAC). Figure 2 presents the component diagram of LGAC.

LGAC comprises of the following parts:
- Biometric Sensor: This sensor receives raw biometric data from the user.
- Feature Extractor: The raw biometric data are transformed to one-way biometric templates, by implementing a function that extracts and encodes specific details from the raw biometric data.
- Matching Algorithm: This algorithm matches biometric templates and produces a matching result.
- Application: The application that has control of the overall biometric operation.
- Cryptographic Algorithm: This algorithm is responsible for the cryptographic functions required for establishing a secure channel between LGAC and PAS.
- Network Interface: Responsible for the communication of LGAC and PAS.
- Display: Presents the results of the system operation.
- SmartCard Reader: An interface between the smartcard and LGAC.

As far as the biometric process is concerned, the software implements the functions depicted in figure 3.

The biometric process is described below:
- The raw biometric data are captured by the sensor: A’ in the case of enrolment and A in the case of normal access control.
- The f function implements feature extraction and encoding. It produces T = f(A), where T is the biometric template and A the raw biometric data. This process is also implemented during user's enrolment, producing value T’ by applying function f to A’. Value T’ is securely stored in the smartcard.
- T and T’ are compared, producing a matching result (m).
- D = d(m), where D is a binary value of the decision taken (proceed or not to step 2 of the protocol), by processing the matching result through the use of a decision function d. This function may be configured in order to make the system more or less strict. For example, we could configure d to produce a positive result if m > 60%.

VII. TESTING ENVIRONMENT

A pilot version of Athlos2 was implemented in a
stadium hosting athletic events, including basketball games, athletics and gymnastics. The intelligent access control system is composed of LGAC terminals with embedded fingerprint biometric devices, which are located in the entrances of the stadium, as well as RFID smart card readers placed in the zone borders of the stadium and at the entrances. The terminals communicate with a central server (PAS) through a local area network. Another central server hosts the attribute administration system (IdP) including a database filled with access privileges for sports fans. All servers were located in secure offices, while user enrolment took place after informing the user through the Internet and on-spot, in a secure office by the ticket booth. Since similar systems were described and introduced in the past, the pilot operation focused on the factors that are mostly questioned and doubted - security and user acceptance.

VIII. ACCEPTANCE TESTING

For testing the acceptance and usability of the Athlos2 pilot system, an extended version of the Davis’ Technology Acceptance Model was deployed \cite{13}. TAM contains two dimensions: usefulness (divided into accomplishment and efficiency) an ease of use (divided into learnability, control and mental effort). The extension to the TAM was provided by Amberg et al. \cite{14}. They introduced an Acceptance Model for the Analysis and Design of Innovative Technologies (DART) including dimensions of perceived ease of use, perceived usefulness, perceived network effects and perceived costs.

Based on DART, a survey regarding acceptance and usability of Athlos2 was conducted, focusing on the biometric access control system, taking into account possible privacy consideration of the users. A total of 110 participants, 45% female and 55% male filled the questionnaire, during a 2-month period. Their age varied between 18-65 years. Most participants were familiar with the use of automated systems. The aim of the study was to investigate participants’ acceptance and general attitudes towards biometrics and more generally Athlos2. The questionnaire was answered in three phases: before informing the user regarding the operation of Athlos2, after informing the user and finally after the user was enrolled and tested the system in practice, during an athletic event.

During the first phase, the acceptance of biometrics was relatively high amongst the participants. The overall mean of the attitude was 3.24 measured on a five-point scale (1=negative, 2=quite negative, 3=neutral, 4=quite positive, 5=positive). Similarly, the acceptance of Athlos2 in total was high, with an overall mean of 4.01 measured in the same scale as above. During the second phase, the acceptance of biometrics was even higher amongst the participants. The overall mean of the attitude was 4.14. Similarly, the acceptance of Athlos2 in total was high, with an overall mean of 4.68 measured in the same scale as above. The main reason for this increase in user acceptance, was that the users' privacy concerns, especially regarding the collection and use of biometric data were minimized, after being informed of the operation of the system and especially regarding the fact that the users carry within their smartcards their own biometric data in encoded and encrypted forms, while no storage takes place in any central database. During the last phase, the acceptance and usability of biometrics had an overall mean of 4.43, while the acceptance and usability of Athlos2 have a mean of 4.77 measured in the same scale as above. The participants recognized the benefits of the system and reported that it would increase their level of security while attending an athletic event, without compromising issues, such as usability and privacy.

IX. INFORMATION SECURITY AND PRIVACY ASSESSMENT

Risk analysis was conducted, during the implementation of Athlos2, for evaluating its security level, focusing on the use of biometrics and RFID smart cards, in relation to the users personal biometric data. For this purpose a specialized methodology and
knowledgebase of vulnerabilities, risk and countermeasures for security and privacy was deployed [13]. The vulnerabilities addressed by Athlos2 are described below.

- The utilization of the template in two or more applications with different security levels (i.e. convenience applications and security applications) tends to equalize these security levels, by decreasing the higher security level to the lower one - if a template is compromised in one application, it can be used for gaining access to the other. The biometric algorithm of Athlos2 is custom, producing unique biometric templates hence this vulnerability was addressed.

- Capturing the power consumption of a chip can reveal the software code running on the chip, even the actual command. The application of Simple Power Analysis and Differential Power Analysis techniques is possible to break the matching mechanism of the biometric system or reveal the biometric template stored in smart card. Timing attacks are similar and measure the processing time instead of the power consumption. The RFID smart card had countermeasures implemented against these types of attacks, including low power consumption chips, noise generators and time-neutral code design.

- Poor biometric implementations are vulnerable to spoofing and mimicry attacks. An artificial finger made of commercially available silicon or gelatin, can deceive a fingerprint biometric sensor. This vulnerability is addressed, since vitality detection features were implemented in the fingerprint sensor and the environment was controlled by a guard.

- Poor enrolment, system administration and system use procedures, expose the biometric system. During the enrolment phase, raw biometric data and biometric templates can be compromised and databases can be altered or filled with imprecise user data. Poor system administration procedures, in addition to the above, might lead to altered system configuration files, with decreased False Acceptance Rates, making false acceptance easier, thus security weaker. Similarly, a user might exceed his/her authority, threatening the system.

- This vulnerability was addressed, since enrolment, administration and system use was implemented according to international standards and best practices [18].

- Server based architectures, where the biometric templates are stored centrally inherit the vulnerabilities of such systems. A possible attack can be realized when the impostor inserts his template in the system under someone else's name, or attacks the central database in order to breach the confidentiality or user data. This vulnerability was addressed, since the template was stored in the protected memory of the smart card.

- Data could be captured from a communication channel, between the various components of a biometric system, in order to be replayed at another time for gaining access. This vulnerability was addressed, since the biometric component was limited in a hardware security module, with physical security countermeasures implemented and the environment was controlled by the personnel of the stadium.

- Off-limit power fluctuation or flooding of a biometric sensor with noise data - for example flashing light on an optical sensor, changing the temperature or humidity of the fingerprint sensor, spraying materials on the surface of a sensor or vibrating the sensor outside its limits - might cause the biometric device to fail. Since the corresponding part of the security policy implementation ensured a controlled environment for the biometric devices.

- The residual biometric characteristic of a user on the sensor may be sufficient to allow access to an impostor (e.g. a fingerprint the sensor). The attack is realized on a fingerprint sensor with a residual fingerprint from the previous measurement, by pressing a thin plastic bag of warm water on the sensor, by breathing on the sensor or by using dust with graphite, attaching a tape to the dust and pressing the sensor. This vulnerability was addressed, since the sensor deployed was capacitive and not applicable to these types of attacks. Furthermore the environment is controlled by
Regarding the remainder of the infrastructure, a security study was conducted, including a vulnerability assessment for the network elements, the database, the operating systems, the applications and the servers. All necessary network security controls were deployed, including firewalling and intrusion detection systems, as well as network device hardening and the deployment of secure network protocols. The database security controls were deployed according to best practices, for realizing confidentiality, integrity and availability especially for the user data. Operating system hardening and application level countermeasures were also deployed, implementing a standard security policy. The security policy also covered security organization issues and personnel procedures, being compatible with ISO/IEC 17799:2005: Information technology - Security techniques - Code of practice for information security management.

**X. CONCLUSIONS**

Athlos2 was evaluated in terms of acceptability, security and performance. Acceptability was a very important factor, since the deployment of biometrics usually have a negative impact to the public due to the consideration of privacy issues. The acceptance assessment however, revealed that especially after informing the users regarding the system operation, biometrics were not only accepted by the users but also recognized as a mean to increase security and relief users from the anxiety of incidents during an athletic event. System security was mainly focused on the biometric component of the pilot implementation. A specialized methodology was deployed for assessing the risk of the biometric component of Athlos2 and all necessary countermeasures were developed within the system in order to address all known vulnerabilities. Future work involves a full deployment of the system and the system testing in athletic events of different types.

**REFERENCES**


BIOGRAPHIES

Dr. Christos Dimitriadis, is a researcher at the University of Piraeus, specialized in prevention, detection and response IT security mechanisms. He has been invited by several organizations to provide lectures, including the ITU, US-NIST and several agencies of the European Union. Dr. Dimitriadis has 33 publications in the area of IT security. His research interests include 3G and 4G security architectures, identity management (founding member of the Mobile-Government Study Group - MGSG), honeynets and security protocol design and testing. Dr. Dimitriadis received a diploma of Electrical and Computer Engineering from the University of Patras-Greece, a PhD on IT security from the University of Piraeus-Greece and is a Certified Information Security Manager (CISM) and Certified Information Systems Auditor (CISA) from the Information Systems Audit and Control Association (ISACA).

NOTE

Prof. Peter Stavroulakis’s picture and CV is on page 53.
ABSTRACT

Tropospheric Scattering (Troposcatter) and Tropospheric Ducting are two different mechanisms due to inhomogeneities in the lower part of the Earth’s atmosphere. Their common influence in propagation of microwaves is studied here in order to achieve a feasible naval communication system. Although not new, Troposcatter together with Tropospheric Ducting communications are regaining popularity in current military and civil applications such as isolated islands and oil extraction facilities in the open sea. The innovation proposed here is the application of such systems in naval communications where at least one of the stations is mobile. Several propagation together with fading models are simulated using appropriate programs and scientific simulation packets in order to predict maximum range of such communication systems in each dominant mode of operation. Then appropriate antenna design proposals are given to overcome the high propagation loss and minimize possible interferences.

Key words: naval communications, tropospheric scattering, tropospheric ducting

I. INTRODUCTION – AN OVERVIEW.

Although INMARSAT and other satellite communication networks are the main providers of today’s naval communication, there are certain limitations in operational cost or data security that cannot be met. Especially where sensitive data are involved (e.g. military or telemedicine information), the use of a channel that may be put out of service at provider’s will, is not desirable. Of course alternative long-range communication mechanisms such as HF ionospheric refraction are still operable. Their main
advantage is an extremely long range (4000km with a single hop) but their main disadvantage is their bandwidth, which for an HF channel never exceeds 20KHz.

Another alternative is the use of troposcatter links between ships or between a ship and a land base station (LBS). Although troposcatter systems have been used for a long time in military and civilian communications – especially when long ranges and inability of radio relay installation were involved such as links between islands or off – there is not yet any mobile platform application. The main problems are the high gain antennas and high power transmitters necessary for overcoming the high path losses (usually of the order of 250dB). In this paper it is shown that such limitation may be overcome using purpose designed systems. Alternatively, the value of tropospheric ducts cannot be neglected, since they are very common in the Mediterranean Sea especially in summer.

II. TROPOSPHERIC DUCTS AND SCATTERING.

Propagation in the troposphere for VHF UHF and lower SHF bands (30MHz to10GHz) beyond the line of sight (LOS) is governed mostly by ducting and scattering. Although these mechanisms are different in principle, they both are effects of the refractive index variability caused mostly by the water vapor distribution in the troposphere. Refractivity \( N \) is expressed in terms of pressure \( p \) and water vapor pressure \( e \) in Kelvin degrees \([1,2]\),

\[
N = (n - 1) \cdot 10^6 = \frac{77.6 \cdot \frac{P}{T} - 5.6 \cdot \frac{e}{T} + 3.75 \cdot \frac{e}{T^2}}{T}
\]

In (1) the refraction index \( n = c_o / c = \sqrt{\varepsilon} \), also expresses the propagation velocity of radio waves in the atmosphere, \( \varepsilon \) being the relative dielectric constant. Close to the surface of the Earth, \( n \) is usually 1.00025 – 1.00040. Water vapor pressure is related to relative humidity \( RH \) and saturated vapor pressure with the equations below:

\[
e = \frac{RH}{100} \cdot e_s
\]

\[
e_s = a \cdot \exp\left(\frac{b \cdot (T - 273.14)}{(T - 273.14) + c}\right)
\]

For liquid water: \( a = 6.1121, b = 617.502, c = 240.97 \) and for ice: \( a = 6.1115, b = 22.452, c = 272.55 \).

For almost horizontal radiowave propagation at an altitude \( h \) using the Snell’s law \([1]\) the electromagnetic rays will be curved with curvature radius \( \rho \) :

\[
\frac{1}{\rho} = -\frac{dh}{dN} = -10^{-6} \frac{dN}{dh}
\]

To overcome this complexity an equivalent Earth radius is used:

\[
\frac{1}{R_{eq}} = \frac{1}{R} + \frac{1}{\rho} = \frac{1}{KR}
\]

In standard conditions \( K = 4/3 \). This gives an equivalent Earth radius \( R_{eq} \). Another useful index is the modified refractivity index \( M \):

\[
M = (n - 1 + \frac{h}{R}) \cdot 10^6 = N + 0.157h
\]

Extended theory can be found at [1], [2].

Ducting occurs if \( \frac{dM}{dh} < 0 \) for a tropospheric layer and \( \frac{dM}{dh} > 0 \) for the layer above. This is common over sea, and especially over the Aegean Sea \([3,4]\) the probability of ducting is over 70% in some cases.

This gives the possibility of radio transmissions in the VHF, UHF and lower SHF bands over ranges exceeding by far the radio electric horizon (horizon if equivalent Earth radius is considered).

Even in the absence of ducting conditions random irregularities of the modified refractive index may generate scattering conditions. Although no analytical model has been proposed, numerical models are extensively used in order to estimate the propagation laws. This is the troposcatter case, presented extensively in [5].

Tropospheric duct and scatter simulation programs like AREPS® \([6]\) exist, whereas several empirical models have been also developed by experienced amateurs (e.g. “scatter.exe” B. Atkins – KA1GT), in the second case giving point-to-point estimations. For these reasons, and also for evaluation purposes
another more straightforward method, the Chinese method (described in thoroughly in [5]) for evaluating tropospheric scatter loss, have been provided by author as a Matlab® script in Appendix A.

**III. A TROPOSPHERIC DUCT / SCATTER COMMUNICATION SYSTEM**

For a 1 Mbps capacity system (comparable with 8 new INMARSAT channels) a rough estimation for a robust modulation scheme (AFSK for example) is a 3.2 MHz channel. The received noise is

\[ N = kTB \]  

(7)

For a modest \( C/N \) ratio of 20 (or 13 dB) the received carrier power (also signal power in constant amplitude systems) is:

\[ C = (C/N)_{dB} + N_{dB} = (C/N)_{dB} + 10 \log(kdB) \]  

(8)

Which gives for a system noise temperature of 350 K (made possible by a sensitive high gain first stage amplifier just after the antenna) a received power level of -96 dBm. On the other hand the radio-link budget equation is:

\[ S_{min,dBm} = C_{dBm} = P_{TX,dBW} + G_{T,dB} - 2G_{D,dB} + G_{R,dB} - L_{tot,dB} + 30 \]  

(9)

Where: \( P_{TX,dBW} \): Transmitter power (dBW), \( G_{T,dB}, G_{R,dB} \): transmitter and receiver gains in dB, \( L_{tot,dB} \): total propagation loss in dB.

It has to be noted here that the term \( G_{D} \) is an expression of the decoupling loss and has the meaning that if a pair of antennas are used for a troposscatter link, then the common tropospheric volume is reduced when the gains are increased, and consequently the absolute number of scatterers is decreased (fig.1). It is empirically given by the equation:

\[ G_{D,dB} = 0.07 \exp(0.055(G_{T,dB} + G_{R,dB})) \]  

(10)

Consequently, antenna gains over 55 dB are impractical. It is advisable though, that reasonably high antenna gains are to be used in order to avoid fading spread. For the system discussed above if practical gains at 40 dB and transmitter power about 20 KW are used then the maximum total loss is \( L_{tot,dB} \approx 238 dB \). Then the “Chinese model” of propagation estimates the range of the link. The case presented in Fig. 2 is a land-to-ship communication with antennas at 600 m and 25 m, and carrier frequency of 2200 MHz. Comparing this method to other similar ones, it seems rather pessimistic; but it is good for the design stage.

![Fig.1 Decoupling loss: Scattering volume between angles \( \gamma \) is larger than the volume between angles \( \theta \) (dotted lines – higher gain antennas)](image)

![Fig.2 Troposscatter loss (dB) vs. Range (km) estimation using Chinese method.](image)

In order to estimate its performance, a simulation model has been developed. The critical part is to simulate the fading of a troposscatter link channel that has both slow and fast fading characteristics (Fig.3).
The channel characteristics are as suggested in [7]—a Rayleigh fading channel with a spread factor $SF = T_D D_s \approx 10^{-5}$ ($T_D, D_s$ being the time and Doppler spread respectively). Other characteristics of the system are as described above, plus nonlinearities in the amplifiers. The system is simulated for a 60 second interval, for 4096-bit packets of data. The cumulative probability and probability density diagrams of the error rate are given below (Fig.4 and 5).

These results suggest the use of some form of diversity to overcome fast fading. Spatial diversity is out of discussion for the ship-based station (SBS) but not for the LBS.

Although designed for troposcatter conditions this design is plausible in ducting conditions too [8] if a simple back off mechanism is connected to control high-power stages and receiver thresholds.

V. ANTENNA DESIGNS

The antennas needed in this type of links may be...
either parabolic reflector antennas or phased arrays. As it has been stated thoroughly in [8,9] parabolic reflectors on board a ship pose a significant drag and severe steer ability in adverse weather conditions. It is a good choice though for the ground base stations though.

Although complex, phased arrays are the only way to achieve the desired gain. In order to suppress side lobes (usual in array antennas) the most attractive design appears to be a 200x50 Dolph-Chebyshev phased array with elements placed - $\lambda$/2 apart. This gives 1.4x5.6 degrees of main beam (-3dB angles) and allows a desired limit of side lobe level of -40 dB maximum as it is shown in fig.6. Phased arrays have also the capability of conformal installation on the ships outer frame, and the ease of instantaneously change their main lobe direction.

Phased arrays also are easy to function as monopulse tracking antennas (like a tracking radar) now tracking a static target (the LBS) from a mobile platform (the SBS).

CONCLUSIONS

So far, this proposition of solutions is more than encouraging for possible applications which require high level data security, long range transmission and comparably high link capacity. Problems occurring by fading require a little more effort to be solved but research is underway and the minimization of these problems will eventually be achieved. Further case study of course will required for each ship, communication signal parameters, vessel design and compatibility with existing communication equipment. Comparing it to satellite communications the proposed system is independent from satellites (which in some cases could be unavailable, or monitored by the satellite network operator). It has larger capacity than a HF ionospheric link and high directivity antennas tracking each other may be used to avoid signal interception, or interference. It has longer range than any VHF-UHF conventional point-to-point link. These anticipations and the results so far are encouraging for systems of ducting and troposcatter for naval use.

REFERENCES


Appendix A:The Chinese Method for evaluating a troposcatter link

%Nikos Farsaris 2003-2005
%Matlab 6 code for Chinese method
%based on G.Roda and ITU
%undocumented... contact njf@pathfinder.gr for
detail
clear all;
clc;
RG=6400;
K=4/3;
RE=K*RG;
htx=0.6; %Km
hrx=0.024; %Km
Ns=335;
DN=40;
f=2200;
Xk=[200:5:600];
for i=1:length(Xk)
Xr=Xk(i);
theta=(Xr/RE)-sqrt(2*htx/RE)-sqrt(2*hrx/RE);
%rad
dtx=sqrt(2*RE*htx);
%thtx=-dtx/RE;
drx=sqrt(2*RE*hrx);
%thrx=-drx/RE;
alpha=thtx+Xr/(2*RE)+(htx-hrx)/Xr;
beta=thrx+Xr/(2*RE)-(htx-hrx)/Xr;
dstx=Xr*beta/theta;
dsrx=Xr*alpha/theta;
dshor=Xr-dtx-drx;
Hx=theta*dstx*dsrx/Xr;
LHx=20*log10(5+0.3*Hx)+0.65*Hx; %...
LNs=-0.08*(Ns-320);
A=LHx+LNs;
L50(i)=124.6+30*log10(f)+30*log10(theta)
+10*log10(Xr)+A;
ctrlf=alpha+beta-theta %<10e-12
sigma=8.8+0.11*DN*exp(-3e-3*dshor);
L84(i)=L50(i)+sigma;
Lpm(i)=L50(i)+3.1*sigma;
end
figure(1)
plot(Xk,-L50,'k',Xk,-L84,'k',Xk,-Lpm,'--k',Xk,-
238*ones(size(Xk),'-k')
legend( 'Prop. 50%','Prop. 84%','Prop. 99%','Threshold')
xlabel('Range km')
ylabel('Loss dB and threshold')
grid on

**BIOGRAPHIES**

_Nikos J. Farsaris_ holds a Diploma of Electrical Engineering from the Aristotle’s University of Thessaloniki since 1994 and was specialized in Radio Telecommunications. He is a Ph.D. Candidate in the same university since 2003. From 1995 to 1999 he was a research assistant in the Telecommunication Systems Institute (TSI) of Crete, and a laboratory instructor in the Technical Education Institute of Crete. From 1999 to 2001 he served in the Hellenic Army as a Lieutenant of the Armored Cavalry. From 2001 to 2002 he was an instructor in the Merchant Marine Academy of Crete. Currently alongside his Ph.D. he is an instructor of Technical Education Institute of Crete, and a researcher in its Electromagnetic Radiation Measurement Laboratory. From 2004 to 2005 he was also a technical consultant of the Olympic Games Security Department of the Hellenic Police. He is the author of nine international journal papers, and twenty six refereed conference papers in the areas of Antennas, Tropospheric and Ionospheric Propagation, Radar and Electronic Warfare, Airborne and Satellite Communications, and Electromagnetic Compatibility.
Chaos–based Applications in Secure Optical Communications

A. Argyris, A. Bogris and D. Syvridis
Department of Informatics and Telecommunications, University of Athens, 15784, Greece. argiris@di.uoa.gr, abogris@di.uoa.gr and dsyvridi@di.uoa.gr.

ABSTRACT
A comprehensive study of an all-optical chaotic communication system, including experimental realization, real-world testing and performance characterization through bit-error-rate analysis, is presented. Pseudorandom bit sequences that are effectively encrypted in a broadband carrier produced by a chaotic emitter and sent for transmission are recovered at the receiver side. Bit-error-rate (BER) values as low as 10−7 for 1 Gb/s data rate have been achieved. Different data code lengths and bit-rates at the Gb/s region have been tested. The application of optical transmission using 100km fiber spools in an in-situ experiment and 120km in an installed optical network showed that transmission effects do not act as a considerably deteriorating factor in the final performance of chaos-based optical communication systems.

Key words: optical communications, chaos, encryption, synchronization, security

I. INTRODUCTION
Optical communication systems are now well established in the infrastructure of the global communication nest, providing a huge bandwidth potential for demanding future applications. In order to assure the uppermost possible privacy and security of the interconnected users, encryption methods are also applied to the physical layer of the communication procedure. Various approaches have been essayed over the last decade, including quantum cryptography[1] and chaos-encrypted digital[2], electrical[3] and optical systems[4]. Especially chaotic optical communications is a flourishing contemporary research field, very promising in shielding the security aspects of the future optical networks. The potential of synchronizing coupled non-linear generators has been proved to a great extend[5], including semiconductor laser emitters that operate in the telecommunication wavelengths and that exhibit chaotic dynamics of high complexity[6-10]. Optical feedback[11-17], optical injection[18-20] or optoelectronic feedback[21-23] are some of the typical configurations used to generate a high-dimensional chaotic laser output. The bandwidth of such chaotic carriers may extend even up to tens of gigahertz, making them ideal candidates for high bit-rate message encryption. Experimental observation of chaos synchronization has been reported for all the above systems[23-26].

In recent experiments that involve data communications in chaotic optical systems, encoding and decoding of sinusoidal signals with frequencies up to a few gigahertz has been demonstrated[27]. A 2.5Gb/s non-return-to-zero (NRZ) pseudorandom bit sequence has been referred to be masked in a chaotic
carrier, produced by a 1.3μm DFB diode laser subjected to optoelectronic feedback, and partially recovered \(^{28}\). Also, a similar system was developed by Larger et al. at 1.55μm, who successfully encrypted a 3Gb/s pseudorandom message into a chaotic carrier, while the system’s decoding efficiency was characterized by low BER values of the order of 10^{-9} \(^{29}\); however the non-linear medium in that case was a Mach-Zehnder modulator. Finally, a 1.55μm all-optical communication system with chaotic carriers has been successfully developed and characterized by the authors, with BER measurements at gigabit rates that exhibited promising results \(^{30}\). All the above communication systems have been studied in an in-situ transmitter-receiver configuration, in absence of any transmission medium.

In this work we demonstrate an all-optical gigabit communication system based on chaos encryption that includes transmission medium. Pseudorandom bit sequences that are effectively encrypted into optical chaotic carriers are transmitted over a length of 100km single-mode fiber and are successfully decrypted at the receiver side. BER measurements are presented for different message bit-rates in order to characterize the system’s performance and to identify the extent of the transmission effects induced by the fiber link, assuming various configurations of dispersion management in the transmission path.

II. PRINCIPLE OF OPERATION

The operation principle of the chaos based optical communications systems is schematically depicted in figure 1. In the conventional communications systems an optical oscillator (semiconductor laser), generates a coherent optical carrier on which the information is encoded using one of the many available modulation schemes. On the contrary, in the proposed approach of the chaos based communications the transmitter consists of an optical oscillator forced to operate at the chaotic regime by an external feedback, producing thus an optical carrier with extremely broadened spectrum (in the order on many tens of GHz). The information (typically an on-off keying bit stream) is encoded on this chaotic carrier using different techniques (e.g. a simple yet efficient method is to use an external optical modulator electrically driven by the information bit stream while at its input is coupled the optical chaotic carrier). It is practically impossible to extract this encoded information using conventional techniques like linear filtering, frequency domain analysis or phase space reconstruction, assuming a high complexity in signal carrier and message amplitude that does not exceed a few percent of the amplitude of the chaotic carrier. At the receiver side of the system a second chaotic oscillator is used, as “similar” as possible to that of the transmitter. This “similarity” refers to: (a) semiconductor laser structural, emission (emitting wavelength, slope efficiency, current threshold, etc.) and intrinsic parameters (linewidth enhancement factor, non-linear gain, photon lifetime, etc.), (b) feedback loop characteristics (cavity length, cavity losses, possible non-linearity, etc.) and (c) operating parameters (bias currents, feedback strength, etc.). The above set of hardware-related parameters constitutes “the key” of the encryption procedure.

The extraction process is based on the so called “synchronization process”. In the context of chaotic communications’ terminology, synchronization means that the time evolution of fast fluctuating optical power produced by the chaotic emitter, can be perfectly reproduced by the receiver, provided that both transmitter and receiver chaotic oscillators are “similar” in terms of the above set of parameters. Even minor discrepancies (e.g. only
few percent in the relevant parameters) between the two oscillators can already result in poor synchronization, which in turn leads to poor reproduction of the emitter’s chaotic carrier.

The key issue for efficient message decoding resides in the fact that the receiver synchronizes to the chaotic oscillations of the emitter without being affected by the encoded message. Based on the above considerations, the receiver’s operation can be easily understood. Part of the incoming message with the encoded information is injected into the receiver. Assuming sufficiently good synchronization, the receiver generates at its output a chaotic carrier almost identical to the injected (without the encoded information). Therefore, by subtracting the chaotic carrier from the incoming chaotic signal with the encoded information, the initial information is revealed.

III. EXPERIMENTAL CONFIGURATION

3.1 Transmitter

The experimental setup of an all-optical open-loop chaotic communication system is shown in figure 2. Two DFB lasers from the same wafer with almost identical characteristics have been selected as the transmitter and the receiver lasers. Both lasers operate at current values of 9.6mA and 9.1mA respectively (with threshold current at 8mA) and with proper temperature controlling they emit at 1552.1nm. Their relaxation frequency oscillation is at 3GHz. The chaotic carrier is generated in a 6m optical external cavity formed between the master laser and a digital variable reflector that determines the amount of optical feedback that is sent into the master laser - set in our setup to 2% of the laser’s outpour optical power. A polarization controller inside the cavity is used to adjust the polarization state of the light reflected back from the reflector. A non-return-to-zero pseudorandom message with small amplitude and code length of at least $2^7-1$ is encrypted into the chaotic carrier of the external cavity’s output by externally modulating a Mach-Zehnder LiNbO3 modulator.

3.2 Transmission path

The chaotic carrier with the encrypted message is amplified to gain enough optical power (4mW) and is transmitted through a total length of 100km fiber span, formed by two transmission modules. Each of them consists of 50km single mode fiber (type G.652), a dispersion compensation fiber module that is used to eliminate the chromatic dispersion, an erbium-doped fiber amplifier (EDFA) that is used to compensate the transmission losses and an optical filter that rejects most of the amplified spontaneous emission (ASE) noise of the EDFA. The transmission characteristics of the two modules are presented in detail in table I. Depending on the sequence of the transmission components used in the transmission modules, different dispersion management techniques can be evaluated: the pre-compensation technique, in which the DCF pre-
cedes the SMF (figure 3a) and the post-compensation technique, in which the DCF follows the SMF (figure 3b).

3.3 Receiver

At the receiver’s side, the transmitted output is unidirectionally injected into the slave laser, in order to force the latter to synchronize and reproduce the emitter’s chaotic waveform. The optical power of the injected signal into the receiver’s laser diode is set to around 0.4mW. Lower values of optical injection power prove to be insufficient to force the receiver to synchronize satisfactorily, while higher values of injection power lead to reproduction not only of the chaotic carrier but of the message also. The use of a polarization controller in the injection path is critical, since the most efficient reproduction of the chaotic carrier by the receiver can be achieved only for an appropriate polarization state. The chaotic waveforms of the transmitter and the receiver are driven through a 50/50 coupler to two fast photodetectors that convert the optical input into electronic signal. The photoreceiver used to collect the optical signal emitted by the receiver adds a $\pi$ -phase shift to the electrical output related to the optical one. Consequently, by combining with a microwave coupler the two electrical chaotic signals - the transmitter’s output and the inverted receiver’s output - an effective subtraction is actually carried out. In the transmitter’s optical path an optical variable attenuator is used to achieve equal optical power between the two outputs, while a variable optical delay line in the receiver’s optical path determines temporal alignment of both signal waveforms. The subtraction product is the amplified message, along with the residual frequency components of chaotic carrier which are finally rejected by an electrical filter of the appropriate bandwidth.

IV. SYSTEM OPTIMIZATION

The first attempt to use the all-optical approach for an optical communication system based on chaotic carrier encryption, as implemented in the setup of figure 2, was presented in [30]. That preliminary work, based on a back-to-back configuration gave promising results regarding the decoding efficiency of the encrypted pseudorandom message. Especially when the data encoding was performed by externally modulating the chaotic carrier, BER values of the recovered message were as low as three orders of magnitude, in respect to those of the encrypted one, equal to $7 \times 10^{-5}$.

In order to further improve the above decoding efficiency some modifications have been performed in this back-to-back system. The first act was to use an ultra-broadband microwave amplifier (Picosecond model 5840A) at the output of the receiver that has a very low frequency cutoff limit (80kHz). Thus, the

---

**Fig.3** Optical transmission modules: a) pre-compensation and b) post-compensation dispersion configurations. SMF: Single-mode fiber, DCF: Dispersion-compensation fiber, EDFA: Erbium-doped fiber amplifier.

<table>
<thead>
<tr>
<th>Transmission parameters</th>
<th>1st transmission module</th>
<th>2nd transmission module</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMF length</td>
<td>50649.2 m</td>
<td>49424.3 m</td>
</tr>
<tr>
<td>SMF total dispersion</td>
<td>851.2 ps</td>
<td>837.9 ps</td>
</tr>
<tr>
<td>SMF losses</td>
<td>12.5 dB</td>
<td>10.5 dB</td>
</tr>
<tr>
<td>DCF length</td>
<td>6191.8 m</td>
<td>6045.4 m</td>
</tr>
<tr>
<td>DCF total dispersion</td>
<td>-853.2 ps</td>
<td>-852.5 ps</td>
</tr>
<tr>
<td>DCF losses</td>
<td>3.8 dB</td>
<td>3.7 dB</td>
</tr>
<tr>
<td>EDFA gain</td>
<td>16.3 dB</td>
<td>14.2 dB</td>
</tr>
</tbody>
</table>
amplifier is able now to respond even at the lower spectral components of the message, especially when the bit-stream length is large ($2^{23} - 1$). The specific amplifier was also preferred for the relatively low noise figure - 5.8dB at 100 MHz - since ultra-broadband amplifiers suffer from very high noise figure, mainly at the low frequency region. The second modification was to adjust the biasing voltage of the external modulator through which the message is applied. Given that the applied message amplitude $V_{\text{mod}}$ is much smaller than $V_\pi$ of the modulator - in order to be efficiently encrypted into the chaotic carrier, one can operate among an excess of biasing voltages $V_b$ on the responsivity characteristic of the modulator. However, as depicted in fig. 3, the biasing voltage determines the operation point of the modulator and hence the quality of the modulated optical carrier. In the specific experiments that follow, we choose the biasing point of $Q_3$ (figure 4). This selection depends on the amplitude modulation, so that the “zeros” of the digital modulation signal correspond to 5V voltage, at which the modulator operates at the minimum transparency. In such a way we achieve minimal optical power of “zeros” and consequently negligible noise effect due to the system. The message “ones”, on the other hand, are obliged to an attenuated response from the modulator. By optically amplifying the modulator’s output, compensation of the optical losses induced by the modulator - due to the selected operation point - is achieved and the message gains enough optical power. Following the above guidelines, the noise effects are more prominent to the “ones”, while “zeros” exhibit a flattened noise profile, as depicted in the eye-diagram of the form of figure 4 ($Q_3$ case). Such a form can lead to lower BER values if the measurement takes place not symmetrically in respect to “one” and “zero” levels, but in a level closer to the “zero” level.

V. RESULTS AND DISCUSSION

5.1 Back-to-back system performance

By applying the conditions analyzed in section 3 in the proposed communication system and by optically injecting 0.4mW of the transmitter’s output into the receiver, the best BER curve achieved for different message bit-rates is shown in figure 5. The BER values achieved now are several orders of magnitude lower than the ones presented in [30]. The code length of the pseudorandom message is $2^{23} - 1$, however, almost the same results were obtained by using a code length of $2^{23} - 1$. The message amplitude, determined by the applied modulation voltage $V_{\text{mod}}$, is set at such levels so that the filtered encrypted message that arrives
at the receiver has a BER value of no less than $6 \times 10^{-2}$.

For each bit-rate studied, electrical filters of different bandwidth have been tested in order to ensure an optimized BER performance of the recovered message. The filter bandwidth $B_f$ selection is crucial and is not only determined by the message bandwidth $B$ but it is associated to the chaotic carrier cancellation that is achieved at the receiver. For example, at the decoding process stage, if the chaos cancellation is not significant the residual spectral components of the chaotic carriers will probably cover the largest part of the decoded message spectrum. In this case, the lowest BER value will emerge by using a filter that rejects the chaotic components, even if it rejects simultaneously part of the message itself ($B_f < B$). On the contrary, for a very good decoding performance, the lowest BER value may emerge by using a filter with $B_f > B$.

The lowest BER value measured for the recovered message was $4 \times 10^{-9}$, for message bit-rate of 0.8Gb/s. As the bit-rate is increased to a multi-Gb/s scale, the BER values are also increased monotonically. This is mainly attributed to the filtering properties of the message at the receiver. The message filtering effect has been confirmed to be larger for lower frequencies and decreases as message spectral components approach the relaxation oscillation frequency of the laser in the gigahertz regime, similar to the response of steady-state injection-locked lasers to small-signal modulation [17]. The above observation is consistent with the results of figure 5. As the message rate approaches the relaxation frequency of the receiver’s laser (~3GHz) the deteriorated message filtering leads to decrypted signal BER values higher than $10^{-4}$.

5.2 Transmission system performance

When intercepting the optical transmission path of 100km the BER values are slightly increased, when compared with the back-to-back configuration. Specifically, when no compensation of the chromatic dispersion is included in the transmission path - i.e. absence of the dispersion compensation fiber (DCF) in figures 2 - the BER values are increased over an order of magnitude (figure 6, circles). For a 0.8Gb/s message the best attained BER value is now $10^{-7}$. Such an increase is attributed to the amplified spontaneous emission noise from the amplifiers, as well as to the non-linear self-phase modulation effects induced by the 4mW transmitted signal.

When dispersion compensation is applied by including into the transmission modules the appropriate dispersion compensation fibers, the BER curves reveal a slightly better system performance in respect to the case without dispersion compensation as the message rate increases. Two different dispersion compensation configurations that are commonly used in optical communication transmission systems have been employed. The first named “symmetrical map” consists of the transmission module of figure 3a followed by the transmission module of figure 3b. The second named “pre-compensation map” consists of two transmission modules that correspond to figure 3b. The corresponding BER values of these two configurations, for the different message rates, are presented in figure 6 (up and down triangles, respectively). For message bit-rates up to 1.5Gb/s, the decryption performance is practically com-

![Fig.6 BER measurements of the encrypted, the back-to-back decoded and the decoded message after 100km of transmission, for different compensation management techniques (no compensation, symmetrical map, pre-compensation map), in respect to the message repetition rate.](image-url)
parable to the case where dispersion compensation is not included. By increasing the message rates, chromatic dispersion plays a more important role in the final decoding performance, so by including different dispersion compensation maps a slight improvement can be achieved. In fact, the pre-compensation configuration shows a very small advantage over the symmetrical map for high bit-rates, up to 2.5GHz.

VI. FIELD EXPERIMENT

6.1 The transmission infrastructure

The next step followed is to test this encryption system to the real world, by sending chaos-encrypted data in a commercially available fiber network. The transmission infrastructure is an installed optical network of single mode fiber that covers the wider metropolitan area of Athens, has a total length of 120km and is provided by Attika Telecom SA. The topology of the link is shown in the map of figure 8. The transmitter and the receiver are both in the University campus, separated by the optical fiber transmission link, which consists of three fiber rings, coupled together at specific cross-connect points. A dispersion compensation fiber (DCF) module, set at the beginning of the link (pre-compensation technique), cancels the chromatic dispersion induced by the single mode fiber transmission. Two amplification units that consist of erbium-doped fiber amplifiers and optical filters are used within the optical link for compensation of the optical losses and amplified spontaneous emission noise filtering, respectively.

6.2 Results

The system’s efficiency on the encryption and decryption performance is studied by bit-error rate (BER) analysis of the encrypted/decoded message [31]. The message amplitude is attuned so that the BER values of the filtered encrypted message do not exceed in any case the value 6×10⁻². In figure 9, spectra of the encrypted (upper trace) and the decrypted - after the transmission link - (lower trace), 2⁷⁻¹ length, 1Gb/s message are shown. The good synchronization performance of the transmitter-receiver setup leads to an efficient chaotic carrier cancellation and hence to a satisfactory decoding process. The performance of the chaotic transmission system has been studied for different message bit rates up to 2.4Gb/s and for 2 different code lengths: 2⁷⁻¹ and 2²³⁻¹ (figure 10). All BER values have been measured after filtering the electric subtraction signal, by using low-pass filters with bandwidth.
adjusted each time to the message bit rate. For sub-gigahertz bit-rates the recovered message exhibits BER values lower than $10^{-7}$, while for higher bit-rates a relatively high increase is observed. This behaviour characterizes the back-to-back and the transmission setup, with relatively small differences in the BER values, revealing only a slight degradation of the system performance due to the transmission link.

VII. CONCLUSION

The bit-error-rate performance of an all-optical communication transmission system based on chaotic encryption has been studied. Pseudorandom bit sequences that were efficiently hidden into broadband chaotic carriers were transmitted over a fiber length of 100km and successfully recovered at the receiver side, with BER values as low as $10^{-7}$. The transmission effects induced by the optical medium in optical communication systems can play an important role in the final performance. The fact that the receiver output needs to be synchronized with the signal that reaches the receiver, rather than the signal generated at the emitter, leads to this dependence on the transmission effects. Transmission non-linear effects such as self-phase modulation degrade the system from its back-to-back performance. Additionally, chromatic dispersion effects that become significant for high bit-rates can be encountered by employing different dispersion compensation maps. Our results show that information can be transmitted at high bit rates using deterministic chaos in a manner that is robust to perturbations and channel disturbances unavoidable under real-word conditions for distances in the order of 200km.

REFERENCES


[29] L. Larger, J.-P. Goedgebuer, V. Udaltsov,


BIOGRAPHY

Apostolos Argyris was born in Thessaloniki, Greece, in 1976. He received the B.S. degree in physics from the Aristotle University of Thessaloniki, Greece, in 1999, the M.Sc. degree in Microelectronics and Optoelectronics from the University of Crete, Greece, in 2001 and the Ph.D. degree in Informatics and Telecommunications from the National University of Athens, Greece, in 2006. He also received the B.Sc. degree as an Observer Meteorologist from the Hellenic National Meteorological Service in 2005. In 2000-2001 he was with the Foundation of Research and Technology Hellas, working on fiber Bragg gratings fabrication and applications. He is currently a researcher in the Optical Communications Laboratory in the Dept. of Informatics and Telecommunications of the National University of Athens, Greece and an adjunct lecturer in the Dept. of Computer Engineering, Telecommunications and Networks of the University of Thessaly, Greece. His research interests include semiconductor lasers dynamics, four-wave mixing, fiber Bragg gratings, LIDAR systems, chaotic encryption and optical communications. He serves as a reviewer for the IEEE. In 2006, he was awarded the “Ericsson Telecommunications Award 2006” for the South-East Europe region and was also named as “Top Young Innovator 2006 - TR35” from the Technology Review magazine and the Massachusetts Institute of Technology. He has authored and coauthored more than 20 articles published in international scientific journals and conferences.

Adonis Bogris was born in Athens, Greece, on June 16, 1975. He received the B.S. degree in Informatics, the M.Sc. in telecommunications, and the Ph.D. degree in all-optical processing based on fiber-based devices from the National and Kapodistrian University of Athens, Athens, Greece, in 1997 and 1999, and 2005 respectively. He is a research assistant working for the Optical Communications Laboratory of the National and Kapodistrian University of Athens participating in local and European projects. His research interests include high-speed all-optical transmission systems, non-linear effects in optical fibers, semiconductor lasers dynamics and chaotic optical cryptography. He is a reviewer for the IEEE Photonics Technology Letters.

Dimitris Syvridis received the B.Sc. degree in physics, the MSc. degree in telecommunications, and the Ph.D. degree in physics from the University of Athens, Athens, Greece, in 1982, 1984, and 1988, respectively. From 1990 to 1994, he was a Researcher with the Institute of Quantum Electronics, Swiss Federal Institute of Technology (ETH-Zurich). Since 1994, he has been with the Department of Informatics and Telecommunications, University of Athens, where he is currently an Associate Professor. He has participated in many European research projects in the field of optical communications. His research interests cover the areas of optical communications and networks, photonic devices, and subsystems, as well as photonic integration. He has authored and coauthored more than 100 articles published in international scientific journals and conferences.
PIM Interference Analysis under Multi-band Multi-signal Input in Duplex Indoor Distribution System

Pauling Huang, Wang Baohua, Senior Engineer, China Unicom
Email: Paulingzh@msn.com, wangbaohua@unicomgd.com

I. INTRODUCTION

Passive Intermodulation (PIM) is a co-existent phenomena in radio system. When two or more signals transmit through passive components (e.g. antennae, connectors, combiners, couplers, isolators, switches, diplexers and cable assembly), the nonlinear response, a decaying series of PIM products (IMPs) of all orders, are generated. If some of IMPs are at the same frequencies and strong enough with receiving signals, they interfere to them. The physical mechanism of PIM and precautions to minimize PIM level are simply described in section 2.

In modern communication system, high power (e.g. 43dBm for GSM) transmitting signal IMPs may interfere to the sensitive receiving signals (e.g. minimum -104dBm for GSM) if they are at the same frequency. This may deteriorate the communication quality, reduce capacity, make bad calls or even drop calls [1~5]. Multi-band multi-signal input will address the problem because of more complicated intermodulation, more IMPs generated and more receive bands.

Most of discussions on PIM assume the input signals in a single frequency band, so it’s reasonable they just concern about the odd order PIM (e.g. 3rd order, ab. IP3), because only odd order PIM may fall...
in the receive band. However, in current mobile networks, sites and indoor distribution systems are often shared by several mobile networks which belong to different frequency bands. In this situation, both even and odd order of PIM may fall in receive bands [1, 7, 8]. So under multi-band multi-signal input, more IMPs may be interference source except for IP3.

PIM is a rapid decaying series by its order, the lower the order is, the stronger of its magnitude. The 2nd order intermodulation product (IP2) is much stranger than IP3. If IP2 falls in a receiving channel, it can interfere to receiving signals more serious. Theoretical analysis of PIM is described and 4 lower order modulation components are deducted in section 3.

With the communication technology development and the increasingly demand of wireless service, there are more and more communication networks coming into use. In China, there are at least 5 mobile networks now. It is hard to find suitable places to locate all base transceiver stations (BTSs) for all systems separately, and even harder to get approval by landlord and residents. As a result, co-locating (for different operators) or sharing (for the same operator) sites and antenna systems are commonly appreciative. For indoor coverage, more and more indoor distribution systems have to be shared by all networks.

In crowded co-locating BTS site and multiple system shared indoor duplex antenna distribution system, there exist a large number of passive components (antennae, connectors, cable, etc.) and steel pole, tower, rusted bolts that may cause PIM. In a famous casino in Macau, a serious PIM interference in indoor distribution system made it can not make any call. Finally by adjusting some input signal channels, the system worked normally. In China, many sites are co-locating and most of the indoor distribution systems are often shared by several networks, so PIM is inevitable and worth of carefully study. For practical use, the important 4 lower orders IMPs that possibly interfere to 2G and 3G system are calculated under the presenting of 2G system signals input in section 4.

II. CAUSES OF PIM AS PHYSICAL PHENOMENA

In radio systems, passive components are usually assumed electrical linearity. In other word, signals between output and input are linear relationship. But if the input signals are strong enough, the passive components behave weak nonlinearity, producing high order harmonics and intermodulation products. In particular, when these components contain the magnetic material or are rusted, polluted by chemical material or loosely connected, the PIM can be more obvious.

As described by many studies[5,8,10], it is extremely complex about the mechanism of nonlinearity in the passive components and usually divide into three kinds:

1. Contact nonlinearity: Caused by different conductor’s contact surfaces, the possible reasons are as following: (1) The contact surface between two conductors exist microscale concave-convex, only some small bulges connect. This causes the surface electric currency nonuniformity, and the contact resistance changes; (2) The conductor surface covers a thin layer oxidation which causes “the diode effect”. When the voltage comes to a certain intensity, the tunnel effect achieves; (3) The nonuniformity rust eclipses on surface that causes the surface current density nonuniformity; (4) Soldering contamination and oxide on the connection surface etc.

2. Material nonlinearity: Ferromagnetic material, non-linear dielectric medium, conductor corrosion and environment temperature change may cause the surface resistance change. The incident electromagnetic wave response voltage, electric current change non-linearly or achieve the magnetic hysteresis, so the second radiating waves have been distorted. The commonly used non-linear material includes: Nickel, iron, cobalt and their alloy material, lanthanum material, aluminum magnesium copper alloy, ferro-
magnetic material, ferrous oxide compound material, non-linear dielectric medium material etc.

3. Surface effect: The conductor surface microscale roughness, pollution and the welding remains make the reflected electromagnetic wave non-linear.

Contact nonlinearity induces IP3 level lower than -100dBc and easily influenced by movement or vibration; Magnetic hysteretic material nonlinearity induces IP3 level from 90 to 100dBc, but it has nothing to do with external force [12]. Obviously, the material nonlinearity caused the PIM is stronger.

In GSM network with 20W (43dBm) transmitting power, the IP3 level is usually from -90 to -120dBm [8]. No data could be found yet about IP2 level. But it is reasonable that IP2 is much stronger than IP3 as IP2 is lower in order than IP3, so IP2 can more serious interfere to sensitive receiving signal if they at the same frequency.

To minimize PIM interference, some precautions should be taken into account in mobile network planning, construction and wireless channel assignment:

1. Avoid to use nickel, iron and the ferromagnetic material or their coat on passive component, the electricity silver-plating or copper-plating may reduce PIM level.

2. Use less connectors as possible, and use welding connection rather than screw connection. If screw connectors have to be used, cleanup the connect interface and screw tightly to ensure enough contact area.

3. Avoid to use braid covered flexible jumper, great angle curve, intense extrusion distortion, make the cable lay straight as possible.

4. Make enough space between transmitting and receiving antennas to avoid transmitting signal PIMs going through receiving path.

5. Change some channels to avoid lower order PIM frequencies falling into receiving channels.

III. PIM MATHEMATICAL ANALYSIS

According to nonlinear system theory, the Volterra integral commonly used to express nonlinear output [13, 14, 15], but PIM is non-time-depend weak nonlinear phenomenon, so the PIM products y(x) may simplify as Taylor series [1, 3, 10]:

\[ y(x) = \sum_{k=1}^{M} y_k = \sum_{k=1}^{M} a_k x^k, \quad a_k = \frac{\gamma^{(k)}(0)}{k!} \] (1)

The magnitude ak reduces rapidly with the increase of k, the higher of k is the much smaller ak, that is \( a_1 > a_2 > a_3 > \ldots \) y(x) is convergent. Therefore, usually only several lower terms take into consideration, the higher terms are neglected. Here below we discuss 4 lower terms.

For N sine signals input x:

\[ x = \sum_{i=1}^{N} x_i \cos(\omega_i t + \theta_i) \] (2)

\[ \omega_i = 2\pi f_i \]

In complex form:

\[ x = \frac{1}{2} \sum_{i=1}^{N} (X_i e^{j\omega_i t} + X_i^* e^{-j\omega_i t}) = \frac{1}{2} \sum_{i=0}^{N} X_i e^{j\omega_i t} \] (3)

Where \( X_i = x_i e^{j\phi_i} \), \( X_i^* = X_i \), \( \omega_i = -\omega_i \) (3) substitutes into (1):

\[ y(x) = \sum_{k=1}^{M} a_k \sum_{i=0}^{N} X_i e^{j\omega_i t} \cdots \sum_{i=0}^{N} X_i e^{j\omega_i t} \]

\[ = \sum_{k=1}^{M} a_k \frac{k!}{2^k} \prod_{i=0}^{N} \frac{1}{m_i!} \] (4)

Where, \( m_0 + \cdots + m_x + \cdots + m_y = k \), \( m_i \) is integer or zero.

From (1), (4):

\[ y_k = \frac{a_k k!}{2^k} \prod_{i=0}^{N} \frac{X_i e^{j\omega_i t}}{m_i!} \] (5)

When \( k = 1 \), from (5):

\[ y_1 = \frac{a_1}{2} \sum_{i=0}^{N} X_i e^{j\omega_i t} = a_1 \sum_{i=1}^{N} x_i \cos(\omega_i t + \theta_i) \] (6)

(6) is linear output component, proportional to input signals x.

When \( k = 2 \), from (5):
\[ y_2 = a_2 \sum_{i=1}^{N} x_i^2 + \frac{a_2}{2} \sum_{i=1}^{N} x_i^4 \cos(2 \omega_i t + 2 \theta_i) \]
\[ + a_2 \sum_{i,j,k=1, i \neq j \neq k} x_i x_j x_k \cos(\omega_i t + \omega_j t + \theta_i + \theta_j + \theta_i + \theta_j) + \cos(\omega_i t - \omega_i t + \theta_i + \theta_i + \theta_i + \theta_i) ] \]  
(7)

In (7), the first item is zero order (direct current) components, the second item is the second order harmonics, and the third item is IMPs.

When \( k = 3 \), by (5):
\[ y_3 = \frac{3a_3}{2} \sum_{i=1}^{N} x_i^3 + 12 \sum_{i=1}^{N} x_i^2 x_i \cos(\omega_i t + \theta_i) + \frac{a_3}{4} \sum_{i=1}^{N} x_i^4 \cos(3 \omega_i t + 3 \theta_i) \]
\[ + 3a_3 \sum_{i,j=1, i \neq j} x_i x_j \cos(2 \omega_i t + \omega_j t + 2 \theta_i + \theta_j) + \cos(2 \omega_i t - \omega_i t + 2 \theta_i - \theta_i) ] \]
\[ + \frac{3a_3}{2} \sum_{i,j,k=1, i \neq j \neq k} x_i x_j x_k \cos(\omega_i t + \omega_j t + \omega_k t + \theta_i + \theta_j + \theta_k) + 3 \cos(\omega_i t + \omega_j t - \omega_k t + \theta_i + \theta_j - \theta_k) ] \]  
(8)

In (8), the first item is linear components, the second item is the 3rd order harmonics, the third item is 3rd order IMPs with arbitrary two out of all input signals, the 4th item is 3rd order IMPs with arbitrary three out of all input signals.

When \( k = 4 \), from (5):
\[ y_4 = \frac{3a_4}{4} \sum_{i=1}^{N} x_i^4 + 18 a_4 \sum_{i=1}^{N} x_i^2 x_i \]
\[ + \left( \frac{a_4}{2} \sum_{i=1}^{N} x_i^4 + 3a_4 \sum_{i,j=1, i \neq j} x_i^2 x_j^2 \right) \cos(2 \omega_i t + 2 \theta_i) \]
\[ + 3a_4 \sum_{i,j=1, i \neq j} x_i x_j x_k \cos(\omega_i t + \omega_j t + \omega_k t + \theta_i + \theta_j + \theta_k) + \cos(\omega_i t - \omega_j t + \theta_i - \theta_j) ] \]
\[ + 36 a_4 \sum_{i,j,k,l=1, i \neq j \neq k \neq l} x_i x_j x_k x_l \cos(\omega_i t + \omega_j t + \omega_k t + \omega_l t + \theta_i + \theta_j + \theta_k + \theta_l) + \cos(\omega_i t - \omega_j t + \theta_i - \theta_j) ] \]
\[ + \frac{a_4}{8} \sum_{i=1}^{N} x_i^4 \cos(4 \omega_i t + 4 \theta_i) \]
\[ + \frac{a_4}{2} \sum_{i=1}^{N} x_i^4 \cos(3 \omega_i t + 3 \theta_i + \theta_i) + \cos(3 \omega_i t - \omega_i t + 3 \theta_i - \theta_i) ] \]
\[ + \frac{3a_4}{4} \sum_{i,j,k,l=1, i \neq j \neq k \neq l} x_i^2 x_j x_k x_l \cos(2 \omega_i t + 2 \omega_j t + 2 \omega_k t + 2 \omega_l t + 2 \theta_i + 2 \theta_j + 2 \theta_k + 2 \theta_l) \]
\[ + 2 \cos(2 \omega_i t + 2 \omega_j t + 2 \omega_k t + 2 \theta_i + 2 \theta_j + 2 \theta_k - \theta_i - \theta_j - \theta_k) ] \]
\[ + \cos(2 \omega_i t - \omega_j t - \omega_k t + 2 \theta_i - \theta_j - \theta_k + \theta_i - \theta_j - \theta_k) ] \]
\[ + 3a_4 \sum_{i,j,k,l=1, i \neq j \neq k \neq l} x_i x_j x_k x_l \cos(\omega_i t + \omega_j t + \omega_k t + \omega_l t + \theta_i + \theta_j + \theta_k + \theta_l) + \cos(\omega_i t - \omega_j t + \theta_i - \theta_j - \theta_k - \theta_l) ] \]
\[ + 4 \cos(\omega_i t + \omega_j t - \omega_i t + \theta_i + \theta_j + \theta_i - \theta_j + \theta_i - \theta_j - \theta_i) ] \]
\[ + 3 \cos(\omega_i t + \omega_j t - \omega_i t + \theta_i + \theta_j + \theta_i - \theta_j - \theta_i) ] \]  
(9)
In (9), the 1st item is the linear components, the 2nd item is the 2nd order harmonics, the 3rd and 4th item are 2nd order IMPs, the 5th item is the 4th order harmonics, the 6th item is the 4th order IMPs with arbitrary 1 signal and another signal’s 3rd order harmonics, the 7th item is 4th order IMPs with arbitrary 2 signal 2nd order harmonics, the 8th item is the 4th order IMPs with arbitrary 2 signals and another signal’s 2nd harmonics. The last term is the 4th order IMPs with arbitrary 4 signals.

For \( k > 4 \), \( y_k \) is more complicated but small of its magnitude. From the above discussion we can come to the conclusions:

1. The lower the order is, the much stranger of IMP magnitude, IP2 level is the strangest.
2. The multi-signals modulation can have all possible mixed products including every order IMP and harmonic.
3. The higher order modulation can have the same and lower order harmonics and IMPs.
4. The even order modulation has the even order IMPs and harmonics, the odd order modulation has the odd order IMPs and harmonics.
5. All higher order modulations have contribution to the same parity lower order IMPs and harmonics.

IV. PIM IN INDOOR DISTRIBUTION SYSTEM OF MOBILE COMMUNICATION

According to the overseas statistics, there is approximately 80% communication traffic occurred in indoor, and the percentage is increasing with the growth of 3G technology and the demand of wireless data communication. Therefore, indoor distribution system pays more and more important role to ensure high quality and heavy traffic load of indoor communication. So PIM interference in indoor distribution system must be concerned significantly.

There are usually two fundamental modes in indoor distribution system design: Duplex and Simplex [16]. In duplex mode, there is only one set of antenna and cable assembly shared by signal reception and transmission (fig. 1 A). In simplex mode, two sets of antenna and cable assembly are con-
structed for signal reception or transmission separately (fig. 1 B). Certainly, we also can design a separate one for each network, but this wastes money and difficult to construct. Actually, the indoor distribution system is often shared by several networks for money save and difficult lease.

Obviously, duplex mode is economical, but the transmitting signals and receiving signals propagate through the same antenna and cable assembly, so IMPs achieved by transmitting signals can easily interfere to the receiving signals. Simplex mode can separate transmitting signals from receiving signals, so PIM interference can be avoided effectively. But it is almost double investment for needing of double passive components (e.g. antennas, connectors, cable etc.) than simplex mode. Balancing between good quality service and investment controlling, a simplex mode system shared by all networks is practicable.

In China, the indoor distribution system covers almost all public places now. Generally, it is constructed by operator independently. But in huge public place (e.g. People’s Conference Building, Olympic Games fields and halls), only one indoor distribution system is permitted to construct for all operators. Operators have to cooperate. This was the same in Hong Kong and Macao earlier, but the situation is changing now, many indoor distribution system are constructed by all operators because of the resistance of landlord, high investment and construction difficulty.

Cooperation is beneficial, it can plan and design perfectly, avoid estate re-negotiation, reconstruction, capitals wasting, minimize the inconvenience to residents and maximize society benefits.

In China, there will be more operators and networks with the 3G system coming into use. With the competition behavior regularization and rational among operators. I’m sure that there will be more and more cooperations among operators in indoor distribution system construction in the 3G era.

It’s possible that some or all radio networks of 2G and 3G system have to share a duplex indoor distribution system. In this situation, if the IMP happens to fall into 2G or 3G system receive band, some receiving channels can not be used. So PIM interference and channel assignment should be carefully studied.

In China, the current 2G mobile systems are 3 kinds: CDMA1X, GSM900 and DCS1800 systems, networks are 5: 1 CDMA1X, 2 GSM900 and 2 DCS1800, which belong to 2 operators respectively: China Mobile owns 1 GSM900, 1 DCS1800 and China Unicom owns 1 CDMA1X, 1 GSM900, 1 DCS1800. For 3G systems, there may be 3 kinds: WCDMA, CDMA2000 and TD-SCDMA. So there may be 6 kinds of systems, 8 networks after the license provided by the government in not long future. The PIM interference will be more serious and complicated.

According to 2G systems radio frequency actual usage and 3G system radio frequency assignment published in document [2002] No.479 published by China Information Industry Ministry[17], the public mobile communication system frequency assigns as in table1:

| The Frequency Assignment for China Public Mobile Communication Systems |
|---|---|
| Actually, China Unicom CDMA system use just 3 channels (No. 283,242,201), China Mobile GSM system occupies more 5 MHz bandwidth of frequency for analog mobile system before (EGSM band). |
| For 3G system, WCDMA and CDMA2000 adopt frequency division duplex (FDD) mode, uplink and downlink use different frequency just like GSM. TD-SCDMA adopts time division duplex (TDD) mode, uplink and downlink share the same frequency by time. If the 3G system signals join into an indoor distribution system, their transmitting signals can modulate just like 2G system signals, and induce IMPs and harmonics which may interfere to wider frequency band. But fortunately, 3G system has few channels. |
| Because it isn’t clear for 3G system licensure yet, here we only discuss 2G system transmitting signal PIM interfering to 2G and 3G system receiving signals. |
Table I The Frequency Assignment for China Public Mobile Communication Systems

<table>
<thead>
<tr>
<th>System</th>
<th>Frequency</th>
<th>China Mobile</th>
<th>China Unicom</th>
<th>Unit: MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Band-Width</td>
<td>From</td>
<td>To</td>
<td>Band-Width</td>
</tr>
<tr>
<td>CDMA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uplink</td>
<td>10</td>
<td>825</td>
<td>835</td>
<td>-</td>
</tr>
<tr>
<td>Downlink</td>
<td>870</td>
<td>880</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>GSM</td>
<td>30</td>
<td>885</td>
<td>915</td>
<td>24</td>
</tr>
<tr>
<td>Uplink</td>
<td>930</td>
<td>960</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downlink</td>
<td>45</td>
<td>1,710</td>
<td>1,755</td>
<td>15</td>
</tr>
<tr>
<td>DCS</td>
<td></td>
<td>1,805</td>
<td>1,850</td>
<td></td>
</tr>
<tr>
<td>FDD</td>
<td>60</td>
<td>1,920</td>
<td>1,980</td>
<td></td>
</tr>
<tr>
<td>Uplink</td>
<td>2,110</td>
<td>2,170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downlink</td>
<td>30</td>
<td>1,755</td>
<td>1,785</td>
<td></td>
</tr>
<tr>
<td>FDD (s)</td>
<td>1,850</td>
<td>1,880</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TDD</td>
<td>1</td>
<td>40</td>
<td>1,880</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>15</td>
<td>2,010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>200</td>
<td>2,400</td>
<td></td>
</tr>
</tbody>
</table>

For N transmitting signals which frequency are \(f_1, f_2, \ldots, f_N\) respectively, from section 3, we can get PIM frequency as following: 
\[
f_{PIM} = \sum_{i=1}^{N} m_i \times f_i
\]  (10)

And \(C = \sum_{i=1}^{N} |m_i|\), where C is order of intermodulation.

Especial case: when all \(m_i\) is zero except one, then \(C = m_i\), where I is the ith signal. That is the C order of harmonic for \(f_i\).

Assume that ith transmitting signal frequency \(f_i = [(f_{0i} - \Delta_i / 2), (f_{0i} + \Delta_i / 2)]\), where fi0 is the channel center frequency, \(\Delta_i\) is channel bandwidth. Then for the existing 2G system transmission we can get following 2nd–4th order intermodulation form which may cause interference to receive bands:

1. 2nd order \(f_i + f_j\):
\[
f_{PIM} = [(f_{0i} + f_{0j} - \Delta_i / 2 - \Delta_j / 2), (f_{0i} + f_{0j} + \Delta_i / 2 + \Delta_j / 2)]
\]  (11)

2. 2nd order \(f_i - f_j\), assuming that \(f_i > f_j > 0\):
\[
f_{PIM} = [(f_{0i} - f_{0j} - \Delta_i / 2 - \Delta_j / 2), (f_{0i} - f_{0j} + \Delta_i / 2 + \Delta_j / 2)]
\]  (12)

3. 3rd order \(f_i + f_j - f_k\), assuming that \(f_i + f_j - f_k > 0\):
\[
f_{PIM} = [(f_{0i} + f_{0j} - f_{0k} - \Delta_i / 2 - \Delta_j / 2 - \Delta_k / 2), (f_{0i} + f_{0j} - f_{0k} + \Delta_i / 2 + \Delta_j / 2 + \Delta_k / 2)]
\]  (13)

4. 4th order \(f_i + f_j + f_k - f_l\), assuming that \(f_i + f_j + f_k - f_l > 0\):
\[
f_{PIM} = [(f_{0i} + f_{0j} + f_{0k} - f_{0l} - \Delta_i / 2 - \Delta_j / 2 - \Delta_k / 2 - \Delta_l / 2), (f_{0i} + f_{0j} + f_{0k} + f_{0l} + \Delta_i / 2 + \Delta_j / 2 + \Delta_k / 2 + \Delta_l / 2)]
\]  (14)

5. 4th order \(f_i + f_j - f_k - f_l\), assuming that \(f_i + f_j - f_k - f_l > 0\):
\[
f_{PIM} = [(f_{0i} + f_{0j} - f_{0k} - f_{0l} - \Delta_i / 2 - \Delta_j / 2 - \Delta_k / 2 - \Delta_l / 2), (f_{0i} + f_{0j} - f_{0k} - f_{0l} + \Delta_i / 2 + \Delta_j / 2 + \Delta_k / 2 + \Delta_l / 2)]
\]  (15)
Obviously, the PIM frequency bandwidth is widened; it is the sum of all participated transmitting signal frequency bandwidths, so it may interferes to more channels.

Because the lower order IMPs are strong. Here the important 3 lower order IMPs, which may interfere to receive band, are calculated and listed in the attached table: “  2G System 2nd- 4th Order PIM Interference to 2G And 3G Systems”.

We can look up the table by transmitting frequency to find whether the PIMs fall into receiving frequency bands, and can calculate the interfered bandwidth by (11)-(15) above. Special attention must be paid to the 2nd, 3rd order PIM products interference for their strong magnitude. If the higher order PIM products (e.g. 4th, 5th) are strong enough to compare with the sensitive receiving signal, they also should take into account.

It is difficult to locate which component causes high level PIM products in a large indoor distribution system. To avoid PIM interference, it is practical to assign transmitting signal channels to ensure all lower order PIM products out of the receiving channels.

V. CONCLUSIONS

This article discusses on PIM phenomena in duplex indoor distribution system under multi-band multi-signal input, the results are also suitable for multi-band antenna system and co-locate antenna in a crowded site. The main conclusions are as following

1. Nonlinearity of passive component causes PIM. Contact nonlinear, material nonlinear and surface effect are the physical reasons to induce PIM.

2. Under multi-band multi-signal input, both even and odd order PIM may interfere to receiving signal, and especial attention should be paid to 2nd and 3rd order PIM which may cause serious interference.

3. All channels of the joined radio networks should be carefully arranged to avoid their lower order PIM products falling into receiving channels.

4. In duplex indoor distribution system construction, it is good to use low PIM level components, such as antennas, cable jumpings, connectors, diplexer, combiners and cable, and should be careful laying and fixing to avoid distortion, loose connection or remains, chemical pollution in components.

REFERENCES


  http://www.hut.fi/Units/Radio/research /rf_applications_in_mobile_communication
[13] Tommy W. S. Chow, Hong-Zhou Tan, Yong Fang “ nonlinear system representation”
  http://www.site.uottawa.ca
[16] Specifications for indoor coverage engineering designation of wireless communication systems, YD/T 5XXX-200X, Communication Industry Standard of P. R. China
[17] Circular on the issue of frequency assignment for the 3rd generation public mobile communication system, No. [2002]479 of Dept. of Radio Management of MII of P. R. China

BIOGRAPHY

Pauling Huang is a senior engineer and member of China Institute of Communications. He received the B.S degree in radio propagation and antenna, the M.S degree in radio physics in Wuhan University and 1986 and 1991 respectively. He has been working on mobile network planning, engineering, operation over 10 year and fulfilled 4 local mobile networks construction and Macao CDMA network construction in China Unicom. His interest is in communication technology , engineering and wireless network optimization.

Wangbaohua is a senior engineer, M.S in radio electronins, he is now researching in mobile communication technology in China Unicom Guangdong branch.