Daniel Giusto
Antonio Iera
Giacomo Morabito
Luigi Atzori
Editors

The Internet of Things

20th Tyrrhenian Workshop on Digital Communications

Springer

The Internet of Things

Daniel Giusto · Antonio Iera · Giacomo Morabito
Luigi Atzori
Editors

The Internet of Things

20th Tyrrhenian Workshop on Digital Communications

Springer

Editors
Daniel Giusto
Università di Cagliari
Dipto. Ingegneria Elettrica e
Elettronica
Piazza d' Armi
09123 Cagliari
Italy
ddgiusto@diee.unica.it

Giacomo Morabito
Università Catania
Dipto. Ingegneria Elettrica
Elettronica e dei Sistemi (DIEES)
Viale Andrea Doria, 6
95125 Catania
Italy

Antonio Iera
Università Mediterranea
Fac. Ingegneria
Dipto. D.I.M.E.T
Via Graziella (Feo di Vito)
89100 Reggio Calabria
Italy

Luigi Atzori
Università di Cagliari
Dipto. Ingegneria Elettrica e
Elettronica
Plazza d' Armi
09123 Cagliari
Italy

ISBN 978-1-4419-1673-0 e-ISBN 978-1-4419-1674-7
DOI 10.1007/978-1-4419-1674-7
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010923231

Preface

In the last few years, a stimulating idea is fast emerging in the wireless scenario: the pervasive presence around us of a variety of "things" or "objects", such as RFID, sensors, actuators, mobile phones, which, through unique addressing schemes, are able to interact with each other and cooperate with their neighboring "smart" components to reach common goals.

This novel paradigm, named "The Internet of Things" (IoT) continues on the path set by the concept of smart environments and paves the way to the deployment of numerous applications with a significant impact on many fields of future every-day life. In this context, logistics, Intelligent Transportation Systems (ITS), business/process management, assisted living, and e-health are only a few examples of possible application fields in which this novel paradigm will play a leading role in the next future.

Actually, many challenging issues still need to be addressed and both technological and social nodes untied before the IoT idea being widely accepted. Central issues are making a full interoperability of such devices possible, enabling their adaptation and autonomous behavior, as well as guaranteeing trust, privacy, and security. Also, the IoT idea poses several new problems concerning the networking aspects. In fact, the "things" composing the IoT will be characterized by low resources in terms of both computation and energy capacity. Accordingly, the proposed solutions need to pay special attention to resource efficiency besides the obvious scalability problems.

The papers included in this volume present a picture of the current state of the art on the above issues; more specifically, concepts and ideas are discussed on networking, middleware, security and privacy, RFID and sensor networks, as well as electromagnetic aspects.

1 Networking Issues in the IoT

Networking issues are of great relevance in the IoT scenario, especially when sensed data and control commands need to be routed through different networks of objects or have to be delivered to servers in the Internet. Routing and addressing are two of the main issues, which need to be addressed taking into account the fact that

the network topologies (physical and logical) vary over the time so that different gateways and clusters of mobile nodes are used to transmit from one network to another. A common scenario analyzed by Bernardo Leal and Luigi Atzori in "*Objects Communication Behavior on Multi-Homed Hybrid Ad Hoc Networks*" is one of the Mobile Ad hoc NETworks (MANETs) interconnected to the fixed networks by means of different gateways. Such multi-homed MANETs can be seen as an extension to the existing infrastructure. Accordingly, MANET nodes may seamlessly communicate with those on the fixed network forwarding packets through appropriate gateways. Two of the aspects that may affect the performance are node address allocation and dynamic gateway changes. When objects in MANETs move around, they may find themselves on a different sub network from the one they have registered in and got their address from. For that reason, their IP address must be changed accordingly while ongoing connections must be maintained and the packets belonging to these connections must be delivered continuously. A crucial role in this context is that of the routing protocols, which may be either proactive, such as the OLSR (Optimized Link State Routing Protocol), or reactive, like the AODV (Ad-hoc On-Demand Distance Vector). The selection of one of these two approaches and the configuration of the adopted protocol is a big issue that needs to be evaluated depending on the type of communications and the main performance issues.

Another emerging solution proposed in the paper entitled "*Traffic Classification in the Presence of Routing Asymmetry*", authored by Manuel Crotti, Francesco Gringoli, and Luca Salgarelli, is that based on the adoption of asymmetric routing, which is a practice already common in the Internet core and is going to affect parts of the network that are closer and closer to the edges, such as the IoT networks. In order to maintain their effectiveness in this environment, appropriate traffic analysis techniques need to be made robust to the effects of asymmetric routing. Performance evaluation of routing protocols in WSNs is presented by L. Bergesio, M. Franceschinis, M. Spirito, and R. Tomasi in their paper. They specifically focus on a multi-hop routing solution, called *Hopefully Longest Jump First*, specifically proposed for WSNs arranged in a linear topology. This solution is compared with two simple approaches, named *Single Hop* and *Limited Flooder*. At the application level, a Master Node has been implemented at one side of the WSN, which progressively queries each of the N Slave Nodes. On its turn, when queried, a Slave Node replies to the Master Node. The analysis is based on performance metrics like end-to-end delay and end-to-end packet delivery success rate, also considering the impact of the distance between two adjacent nodes and thus the expected number of neighbors within a node radio coverage.

Medium access control and management continues to be a subject extensively investigated within the IoT arena, since it heavily affects the communication performance in terms of delays, losses, throughput, and reliability. When devising new strategies, it is important to focus on the self-managing, self-configuring, and self-regulating features, especially if the resulting network has to operate in emergency areas as in the case of the management of catastrophes. In this context, an important issue addressed by Emanuele Cipollone, Francesca Cuomo, and Anna Abbagnale in the paper "*A Distributed Procedure for IEEE 802.15.4 PAN Coordinator Election*

in Emergency Scenarios" is the selection of the network coordinator which has to be performed by also taking into account energy saving during data delivery, thus increasing the network lifetime in case of battery supplied devices (e.g., in case of Wireless Sensor Networks – WSNs) and delivery delay reduction. Cooperation between wireless networks has also quickly gained the interest of many researchers since several studies have shown how the performance can be improved by using principles of cooperation. In particular, a novel architecture, namely Cellular Controlled Peer-to-Peer (CCP2P), has been introduced in the last years. According to it, cellular devices can create cooperative clusters with neighboring devices in their proximity using a short range technology. Each terminal is then contributing to the cooperative cluster by sharing its cellular link. The grouped members acting in a cooperative manner can achieve better performance than a standalone device, in different scenarios. This is the context studied by Emanuele Scuderi, Rocco Parrinello, David Izal, Gian Paolo Perrucci, Frank Fitzek, Sergio Palazzo, and Antonella Molinaro in "*A Mobile Platform for Measurements in Dynamic Topology Wireless Networks*".

However, still great efforts are needed to make the cooperation effective and reliable in realistic scenarios where the peers are continuously moving and the channel is disturbed by concurrent communications in the same geographical area. Whereas cooperation can improve the performance in terms of energy consumption, broadcast techniques allow for reaching the widest area in the shortest possible time which is a feature highly demanded in case of emergency transmission. This is the case of safety-related applications in self-organizing Vehicular Ad-hoc NETwork (VANET), where vehicles share and distribute information by rebroadcasting a received information packet to their neighbors. An efficient broadcast technique can offer a high reactivity without sacrificing the communication reliability. In this context, Cluster-based Irresponsible Forwarding (CIF) is a novel broadcast multi-hop forwarding protocol introduced in "*Cluster-based Irresponsible Forwarding*" by Stefano Busanelli, Gianluigi Ferrari, and Sooksan Panichpapiboon. CIF integrates the recently proposed IF probabilistic forwarding approach with a loosely structured clustered architecture. The clusters of vehicles that naturally form in VANET are typically "ephemeral" and tend to be a source of network congestion, that penalize the performance of "pure" probabilistic approaches. CIF attempts to recognize the presence of ephemeral clusters and adapts its forwarding strategy to the underlaying instantaneous network topology. Also the paper "*Dynamic Spectrum Access Communications: Wavelet Modulation with Unequal Power Allocation*", by Marco Lixia and Maurizio Murroni, addresses network access layer issues by proposing an optimization in the power allocation when using the available spectrum in a dynamic way. An adaptive Wavelet Modulation (WM) scheme is proposed to exploit available resources of the channel while avoiding interferences with the primary users of the spectrum. The power is distributed according to both the sensitivity to channel errors and channel availability. Genetic Algorithms are used to optimize weights with the constraint of average energy per bit remaining unaffected.

Flow control is another crucial issue in heterogonous wireless networks used in the IoT scenario, especially when dealing with multimedia communications.

The paper authored by Enzo Baccarelli, Mauro Biagi, Nicola Cordeschi, Tatiana Patriarca, and Valentina Polli, "*Optimal Cross-Layer Flow-Control for Wireless Maximum-Throughput Delivery of VBR Media Contents*", focuses on the design of control policies and proposes an approach based on the maximization of the average throughput over the wireless last-hop, under constraints on the maximum connection bandwidth allowed at the Application (APP) layer, the queue-capacity available at the Data-Link (DL) layer, and the average and peak transmit energies sustained by the Physical (PHY) layer. The resulting controller is rate-based and operates in a cross-layer fashion that involves the APP, DL, and PHY layers of the underlying protocol stack.

Finally, the paper "*A Secure MPLS VPN Infrastructure for Complex Geodata Sensor Network*", by Mirko Luca Lobina and Tatiana Onali, studies the issues of interconnecting geodata sensor networks when deployed to monitor environmental factor in an extended geographical area. The use of the MPLS protocol is considered in this context, with a keen attention to issues related to the bandwidth management and security, when short messages are transmitted through the backbone network.

2 Middleware for the Internet of Things

The availability of a middleware layer hiding the details of different technologies is fundamental to exempt the programmer from details that are not directly pertinent to her/his focus, which is the development of the specific application enabled by the IoT infrastructures. The IoT may benefit a lot from the existence of such a middleware, since new services will be easily developed and objects interaction will be strongly enhanced. This is a software layer interposed between the technological and the application sub-levels. Several solutions are under study. The WhereX solution, introduced by Antonio Puliafito, Angelo Cucinotta, Antonino Longo Minnolo, and Angelo Zaia in "*Making the Internet of Things a Reality: The WhereX Solution*" has been developed responding to SOA (Service Oriented Architecture) and Multichannel Communication technologies. The use of the SOA principles provides the highest level of flexibility and scalability to the system in the organization of both the external integration processes and the exchange processes within the middleware, by favoring the addition/modification of functions and services (scalability). The use of the XMPP communication model is very important because, since this is a real-time communication protocol, it can provide a high level of interactivity and can simplify the integration with enterprise applications. The SAI (Service Application Integration) middleware solution described in "*A Scalable Grid and Service-Oriented Middleware for Distributed Heterogeneous Data and System Integration in Context-Awareness Oriented Domains*" by David Parlanti, Federica Paganelli, and Dino Giuli, is also based on the adoption of a SOA approach for easing the integration of heterogeneous resources for the development of context-aware applications in enterprise domains. The SOA approach interprets distributed systems mainly as a problem of service

specification, implementation, and composition. A "service" may be defined as a computational entity endowed with an open and addressable specification of its expected behavior. The definition of the "computation entity" is then extended to include software components encapsulating sensors/actuators functionalities. Integration of such real-world devices and business systems usually requires decoupling between service consumers and providers, thus demanding support also for one-way, notification-response, and solicit-response interaction patterns. To address such invocation requirements, SOA's implementation solutions should also be correlated with "message-oriented" approaches. Message orientation gives new insights on service provision/consumption as well as on the overall SOA architectural style. The InterDataNet architecture is another middleware solution proposed by Franco Pirri, Maria Chiara Pettenati, Samuele Innocenti, Davide Chini, and Lucia Ciofi in "*InterDataNet: A Scalable Middleware Infrastructure for Smart Data Integration*". InterDataNet is based on the SOA principles too. It is designed to enable heterogeneous objects networks to expose and integrate their smart data. At the core of the system sits the InterDataNet middleware that defines an object Information Model and the related Service Architecture operating on it in order to provide: a scalable and open service to support a consistent reuse of objects identifiers, that is a global reference and addressing mechanism for locating and retrieving objects in a Web-wide scale; a set of transparent application-services functions, namely historic data management and replica management.

The management of the Resource Identifiers of the Real World Objects (RWOs) is the main subject of the CONVERGENCE middleware framework proposed by Nicola Blefari Melazzi in "*CONVERGENCE: Extending the Media Concept to Include Representations of Real World Objects*". CONVERGENCE is aimed at enhancing current media handling with new functionality and extend the traditional concept of media to include digital representations of RWOs. The Convergence framework is based on the concept of Versatile Digital Item (VDI), a structured package of digital content and meta-information, inspired by the MPEG-21 standard, but designed to address a broader range of application domains, including the management of RWOs in the Internet of Things. The VDI is supported by an innovative middleware and by tools and applications. This framework incorporates six innovations: (1) VDIs provide uniform mechanisms to handle different classes of information, including Real World Objects; (2) VDIs are intrinsically dynamic, allowing both providers and consumers to update content; (3) VDIs support "digital forgetting" (automatic "un-publishing", automated garbage collection of VDIs after a user-defined expiry date); (4) VDIs meet security and privacy needs of both professional and non-professional consumers and providers; (5) VDIs support new modes of semantic search; (6) VDIs allow easy sharing of information across multiple, heterogeneous devices.

Particular attention has to be devoted to the middleware solutions that specifically deal with the management of emergency situations. In this scenario, addressed by Fabrizio Ronci and Marco Listanti in "*Service Oriented Middleware Solutions for Emergency Communication Networks*", a number of operators, decision-makers, and institutional and commercial service providers are usually supposed

to cooperate in order to assist involved population and environment, to overcome the crisis and to start reconstruction. As far as conventional, and possibly inadequate, communication services, basically relying on radio voice calls, yield to a wide gamut of real time, interactive and multimedia data oriented information flows, new viewpoints on network architectures arise from integrating available information, communication and media technologies. All these network resources and components, including today's mobile devices that, as for performance, are a lot more powerful than those of the early days, are able to host very sophisticated and versatile software. This fact enables the usage and exploitation of middleware solutions to integrate knowledge, but also to enhance reliability, to provide transparency, and to guarantee scalability in respect of physical, link, routing and transport technologies, and schemes.

3 Localization and Applications

As we already said, the IoT concept has the potential to change radically our life, thanks to innovative smart applications that adapt their behavior according to the specific context. One of the most important context parameters is the position of individuals or objects. Indeed, several applications can be thought that involve the localization and tracking of people, assets, or goods.

In the past, a large effort has been devoted to the problem of localization of communication nodes. However, the existing results cannot be applied *tout cour* to the IoT scenarios given that nodes have much lower processing and communication resources. Therefore, new methodologies must be devised with higher localization accuracy and higher efficiency.

To answer this research need, several approaches are possible. In this volume, there are two papers that cover the most relevant of such approaches: the use of the electromagnetic signal and the use of the acoustic signal.

More specifically, in "*Localization Issues in a ZigBee based Internet of Things Scenario*", Ugo Biader Ceipidor, Massimiliano Dibitonto, Luca D'Ascenzo, and Carlo Maria Medaglia propose a methodology that uses fixed network devices as reference points. In fact, the RSSI between mobile nodes that need to be localized and the above reference points is measured to obtain distance values and therefore, to derive localization information.

Instead, in the paper entitled "*Low-Complexity Audio Signal Processing for Localization in Indoor Scenarios*", Marco Martalò and Gianluigi Ferrari propose an innovative approach that allows to perform low complexity localization based on the audio signal. More specifically, the proposed methodology allows to localize entities that emit sound through special devices called *anchors*. Based on such an approach, both a centralized and a distributed solution have been designed. Matlab simulations show that the proposed approach allows to achieve high localization accuracy with low processing load. As a future work, authors aim to solve the localization ambiguities raising in certain special cases.

Localization is one of the objectives pursued by Stefano Tennina, Luigi Pomante, Fabio Graziosi, Marco Di Renzo, Roberto Alesii, and Fortunato Santucci in "*Integrated GPS-denied Localization, Tracking, and Personal Identification*". In that paper, a solution is presented that utilizes a biometric badge implementing positioning/tracking algorithms as well as authentication procedures based on fingerprinting matching. Such solutions can be utilized to enable or deny access to restricted areas, for example, to localize patients in hospitals. Localization and tracking are achieved by means of a novel distributed algorithm called ESD (*Enhanced Steepest Descent*). Experimental results show that ESD achieves localization accuracy comparable to the leading solutions much more rapidly.

In certain application scenarios, exact identification and localization of individuals is not necessary, it is sufficient to have information about the number of persons in a certain area and the gender distribution (i.e., how many women? How many men?). Accordingly, in "*Design and Implementation of Smartphone Applications for Speaker Count and Gender Recognition*", Alessio Agneessens, Igor Bisio, Fabio Lavagetto, and Mario Marchese propose a methodology that allows to obtain such information utilizing smart-phones. Current realization of such methodologies allows to distinguish one speaker cases from two speaker cases; however, extension to the case in which several speakers can be distinguished are possible in the future. Prototypes of the proposed methodologies have been implemented on communication devices based on Symbian OS. Experimental results assess the accuracy of the proposed scheme.

In other application scenarios, presence of individuals in a certain area can be identified by means of cameras. In the IoT scenarios, it is of paramount importance to achieve high compression gain given the limitations in the capacity of the devices. Accordingly, in the paper "*Video Coding and Motion Estimation at the Decoder*", Claudia Tonoli, Pierangelo Migliorati, and Riccardo Leonardi present a methodology that achieves high compression ratio thanks to effective and efficient motion estimation at the decoder side. The methodology is based on the Last Square Estimation (LSE) prediction and achieves good performance as demonstrated by experimental results.

Observe that the transmission of video information has strict QoS requirements. Support of such requirements is important especially in the case of disaster recovery applications. Accordingly, there is the need for appropriate QoS management solutions such as the one proposed in "*Inter-Vehicle Communication QoS Management for Disaster Recovery*" by Paolo Orefice, Luigi Paura, and Amedeo Scarpiello. In that paper, focus is on inter-vehicle communications where several problems arise given the dynamics of network topology and link characteristics. Interesting feature of the proposed solution is that it can easily manage new access technologies as they become available in the application scenario without the need of upgrades of the management procedure. Preliminary performance results show that the proposed approach works correctly in PropsimC2 emulator.

4 RFID and Sensor Networks Technologies

For what concerns RFID and sensor networks, it shall be recognized that both of them play a special role within the "Internet of Things" paradigm. According to the International Telecommunication Unit (ITU Report 2005), Internet of Things can be defined as a vision "*...to connect everyday objects and devices to large databases and networks... (using) a simple, unobtrusive and cost-effective system of item identification...*".

Therefore, according to the IoT vision, smart sensor/actuators need to be enhanced with connection capability to locally available networks to the purpose of interacting with the real world. Through the exploitation of smart distributed "objects", such as for example sensors, actuators, RFID tag, and the implementation of data fusing and mining algorithms, the end-user is allowed to identify objects, access real time data, and undertake appropriate actuation strategies, ubiquitously and via Internet.

This is the reason why the enhancements and the integration of communication potentials among RFID tags, sensors, and actuators are key issues to be addressed in any research/publication related to the IoT, together with the integration of the cited devices into hybrid wireless sensor networks.

Researches addressing the RFID and Sensor/Actuator related issues are manifold. Among them, FOSSTRAK (Free and Open Source Software for Track and Trace, an open source RFID software platform that implements the EPC Network specifications), CASAGRAS (Coordination And Support Action for Global RFID-related Activities and Standardization, a European Union-sponsored project looking at future standardization for RFID and especially RFID's role in the emerging Internet of Things), and EPoSS (European Technology Platform on Smart Systems Integration, an industry-driven policy initiative defining R&D and innovation needs as well as policy requirements related to Smart Systems Integration and integrated Micro- and Nanosystems) are examples that deserve a citation because part of their activities are preparatory to the future platforms integrating both technologies (RFID and Sensors) into single IoT platforms.

Ongoing researches testify to the fact that still the set of actions that the future objects should be able to do and the enhancements that are required to best integrate these objects into IoT frameworks is a matter of investigation. This is why in the present publication, some of the main issues relevant to both RFID and Sensor/Actuator networks will be introduced to the reader.

Aspects related to middleware-based solutions for RFID integrations are addressed within the specific session of the present publication dealing with middleware solutions for the IoT. Due to the relevance of the topic, more papers are dedicated to the RFID topic. Some of them deals with RFID technology in general, while the others more specifically address electromagnetic aspects related to the RFID technology. As a result, the reader will be able of obtaining an in-depth picture of main potentialities and limitations of the RFID technology, constraints and enabling technologies for its future evolutions, as well as its effectiveness in impacting in specific IoT application scenarios (e-health, ITS, Logistics, etc).

Since the reading of the first paper, "*Beyond the ID in RFID*" by Christian Floerkemeier, Rahul Bhattacharyya, and Sanjay Sarma, it clearly emerges that RFIDs are fast moving toward the implementation of functions going well beyond the mere identification. Battery powered wireless sensors are the most common commercial wireless sensors used today. However, limited battery life and higher costs limit their deployment in some sensing applications. For this reason, the authors analyze the advantages and shortcomings of alternative passive wireless sensing approaches by emphasizing an emerging paradigm of RFID tag antenna based sensing that offers great potential for the development of ultra low cost, long lasting wireless sensors. Authors conclude that the gains from adopting this sensing approach may outweigh the shortcomings.

A further performance study aiming at the characterization of passive ultra-high frequency (UHF) and radio frequency identification (RFID) tags is presented by Leena Ukkonen and Lauri Sydänheimo in their paper "*Performance Characterization of Passive UHF RFID Tags*". In this paper, the analyses of the effects of dipole antenna width and of the impedance matching properties of a bow-tie tag antenna are two characterization examples used to investigate the harvesting properties of the tag and the significance of the backscattered signal strength and radar cross section (RCS) of the tag.

Different RFID tag technologies and their main characteristics, design, and application are the subject of "*Chipless Tags, the Next RFID Frontier*", by S. Tedjini, E. Perret, V. Deepu, and M. Bernier, which focuses on the emerging concept of the so called "chip-less configurations". The interesting aspects of this tag family (also known as RF barcodes) lay in the fact that they do not use IC chip and the information is directly coded on the surface and/or in the volume of the structure; besides, they are very attractive in terms of cost and data security. The importance of having a clear picture of this novel technology is testified by the number of research projects worldwide dedicated to the development of efficient and versatile chipless tags and to the recent market projections showing that chipless tags will reach 60% of the RFID market before the end of the next decade.

Chip-less configuration is not the only possible enhancement to RFID. Also, the use of Ultrawide bandwidth (UWB) technology to enhance the RFID performance in specific IoT application fields, such as the accurate object localization, is a promising approach. UWB technology might allow next generation RFID systems to overcome most of the main limitations of current narrow bandwidth RFID technology, such as reduced area coverage, insufficient ranging resolution for accurate localization, low security, sensitivity to interference, and scarce multiple access capability. The paper "*Backscatter Communication using Ultrawide Bandwidth Signals for RFID Applications*", by F. Guidi, D. Dardari, C. Roblin, and A. Sibille aims at contributing to this issue, shows that it is possible to provide both identification and high-definition localization of objects by applying the UWB technology to (semi-) passive RFID based on backscatter modulation.

Passive RFIDs are the subjects of the paper "*Passive RFID Integrated Transponders for Automotive Applications*", by Alberto Toccafondi, Cristian Della Giovampaola, Paolo Braconi, and Alessio Cucini. In this paper, the authors specifically

stress the effectiveness of the RFID technology in future IoT scenarios (envisaged for Intelligent Transportation Systems), where the objective is the identification of moving vehicles within a mono-lane scenario for non-stop road-toll operation. To this aim, they propose and analyze a reference system using a HF-UHF RFID integrated transponder. Numerical simulations and experimental results conduced on a transponder prototype confirm the possibility of using this technology for the intended objective.

A different application field in which the IoT paradigm will surely play a starring role in the near future is the one in which the paper "*Sensor-Oriented Passive RFID*" by Gaetano Marrocco, Cecilia Occhiuzzi, and Francesco Amato is conceived. In fact, it is particularly focused on the monitoring of human body features, even if the proposed ideas and devices may also be applied in a variety of different scenarios. More specifically, the paper starts from the consideration that designing low-cost antennas for sensing applications is still a great challenge, especially when the human body is involved. In this view, the authors address the design of new UHF tag antennas for sensing applications able to host detectors and additional electronics but also to act as passive sensors themselves for some modification of the target.

Pharmaceutical distribution is a further sample scenario in which item-level tagging is one of the main challenges in order to improve track and trace systems. The paper "*Performance Evaluation of UHF RFID tags in the Pharmaceutical Supply Chain*", by M. De Blasi, V. Mighali, L. Patrono, and M. L. Stefanizzi, introduces a performance comparison between near field and far field UHF RFID systems. The final conclusion of this work is that the use of passive far field UHF tags could represent the de-facto solution for item-level tracing systems in the whole supply chain. Furthermore, the obtained results lead the authors to assert that the same solution can be easily extended to other sectors in which the item-level traceability is still an important aspect.

Finally, interesting contributions concerned with the business value of RFID technology in future logistics scenarios are the topic of paper "*The Benefits of RFID and EPC in the Supply Chain: Lessons from an Italian Pilot Study*", by Massimo Bertolini, Eleonora Bottani, Antonio Rizzi, Andrea Volpi and paper "*RFID Data Analytics in Apparel Retail*", by Frédéric Thiesse and Jasser Al-Kassab.

In the first one, the potential benefits of RFID technology and EPC Network on the overall fast moving consumer goods (FMCG) supply chain are quantified. The authors show that the largest part of the RFID benefits can be achieved through collaboration between multiple supply chain players. Examples of such benefits include: automation of supply chain processes, better inventory management and decrease in safety stocks, streamlining of other processes, and increase in turnover. While, in the second paper, the interesting contribution of Frédéric Thiesse and Jasser Al-Kassab deals with the business value of large amounts of data generated by RFID data collection infrastructures. The focus of the authors is on a recorded trace data derived from a department store that implemented RFID in its menswear department to seamlessly track thousands of items on their way from the distribution center to the point of sale. The output of their case study indicates that RFID poses an untapped opportunity for retail companies to improve category management,

store layout design, inventory control, and process execution, assumed that the company's individual capabilities exist to translate RFID data into value.

Besides RFID, also the "Sensors" issue is addressed in the present publication from the specific perspective of the Internet of Things. By this meaning, wireless sensors are seen like a companion technology of RFID, in the view of the deployment of common *IoT* infrastructures including RFID and Sensors. Therefore, in this publication the followed approach has been twofold: first novel models of sensor networks are addressed and studied in a couple of papers and then novel applications of Sensor and Actuator networks in very appealing scenarios are introduced as further contributions

Olivier Alphand, Andrzej Duda, Martin Heusse, Benoit Ponsard, Franck Rousseau, and Fabrice Theoleyre, contribute to the present publication with an interesting position paper, "*Towards the Future Internet of Sensors*", in which they propose a new view on the integration of wireless Sensor and Actuator Networks (SAN) in the Internet. According to their approach, the network conveys typed data chunks while applications organize communication according to the Publish/Subscribe model: data consumers subscribe to chunks advertized by producers. The main element of the proposed model is a data router device interconnecting different wireless SAN and offering a data centric view on the physical world to the rest of the Internet.

Still novel models of Sensor Networks (termed "refining" and "expanding") and related issues are the subject of the paper "*Energy and Distortion Minimization in* '*Refining*' *and* '*Expanding*' *Sensor Networks*" by Franco Davoli, Mario Marchese, and Maurizio Mongelli. The first model refers to the acquisition of measurements from a source by means of sensors deployed at different distances, and measuring random variables correlated with the source output. The acquired values are transmitted to a sink, where an estimation of the source has to be constructed, according to a given distortion criterion. The second model represents a "rich" communication infrastructure, where all sensor readings potentially bring fresh information to the sink. In the paper, the authors investigate coding strategies that obey a global power constraint and are decentralized.

As already addressed, three more papers, are focused on very novel and exciting scenarios in which Internet of Things instances based on sensor networks are going to play a starring role.

In the first, "*An IEEE 802.15.4 Wireless Sensor Network for Energy Efficient Buildings*", Chiara Buratti, Alberto Ferri, and Roberto Verdone find in the realization of energy-efficient buildings a very innovative and challenging field of application for wireless sensor networks (WSNs). What they aim at is minimizing the building energy consumption and optimizing the energy use. The eDIANA project, funded by FP7 of the European Commission through the ARTEMISIA framework, is focused on this target. Different network topologies are studied and compared and preliminary outputs of simulation studies are illustrated.

In the second, "*A Real Implementation and Deployment for Wine Production Management Based on Wireless Sensor Network technology*", Luca Bencini, Giovanni Collodi, Davide Di Palma, Antonio Manes, and Gianfranco Manes

describes a successful application of the Internet of Things concept in a challenging environmental monitoring context, concerning the remote management of a vineyard. The shown Wireless Sensor Network System is entirely set into the IoT vision, since it is a valid solution to monitor common parameters using simple, unobtrusive, commercial and cheap sensors, forwarding their measurements by the means of an heterogeneous infrastructure, consisting of wireless sensor network technology, GPRS communications, and ordinary Internet data transfer (TCP-IP protocol).

Last, in the paper "*Performance Evaluation of an IEEE802.15.4 Standard Based Wireless Sensor Network in Mars Exploration Scenario*", Renato Pucci, Demis Boschetti, Enrico Del Re, and Luca Simone Ronga, consider the opportunity to use an IEEE 802.15.4 standard based network in Martian planetary exploration context. By considering a network formed by a mobile rover and 40 sensors, they demonstrate through simulation that an IEEE802.15.4 based WSN can be used in planetary exploration context. Such WSN works pretty well, also in case of transmission within terrains with high density of rocks. The results shown in this paper, demonstrate that WSN should be used in future mission of planetary exploration, following a test campaign finalized to validate simulated and predicted data.

As a last contribution to the issue of "smart objects" (whose category RFID, Sensors, and Actuators belong to) handling, access, and interconnection within a unique IoT platform, in the present publication some reports from the most relevant projects funded by the European Commission dealing with embedded systems and enabling technologies are included. More specifically, the reader will find the papers: "*The PECES Project: Ubiquitous Transport Information Systems*", by Antonio Marqués, Manuel Serrano, "*Probabilistic Information Dissemination for MANETs: the IPAC Approach*", by Odysseas Sekkas, Damien Piguet, Christos Anagnostopoulos, Dimitrios Kotsakos, George Alyfantis, Corinne Kassapoglou-Faist, and Stathes, and "*HYDRA: A development platform for integrating wireless devices and sensors into Ambient Intelligence systems*", by Markus Eisenhauer, Peter Rosengren, and Pablo Antolin.

5 Security and Privacy Issues

Security and privacy issues are a central problem in all ICT scenarios. As such, security has received a lot of attention in the past and is today part of the fundamental know how of any engineer working in all ICT fields. In recent years, attention on the privacy issues raising in several communication scenarios is increasing. Indeed, privacy is recognized among the fundamental human rights and, as the pervasiveness of communication technology increases, new efforts are focusing on privacy problem from both a legislative and technical point of view.

In the IoT case, security and privacy become even more critical as their support becomes more difficult. The reasons of such difficulties lay in both the amount and sensitivity of data that will be generated and will flow through the network, and the

limitations of the computing and communication devices which will be included in the IoT and that are, therefore, much more vulnerable to all kind of security and privacy attacks.

In fact, note that in the IoT concept, objects communicate between themselves and with the information and communication infrastructure as they interact with human beings in everyday life. Therefore, it is evident that in this process they will handle information that could be fused to gain sensible insights into the habits and actions of humans. Furthermore, observe that personal information might be required to deploy added-value context-aware services. In this case, it is important to guarantee that such information is utilized only for the purposes of the service and the only information strictly needed is disclosed to the service provider. Control on the possession and the flow of such information is a must to guarantee an acceptable level of privacy. To this purpose, we observe that the lack of support of a sufficient level of privacy may jeopardize the acceptance of the IoT technologies from a societal point of view.

As we already said, WSNs and RIFD will play an important role in the IoT. Such technologies are characterized by extremely strict resource limitations. In fact, their processing capabilities will be much lower than other computing and communication technologies. Furthermore, their batteries will have low capacity and will be difficult to be replaced or, in case of RFID, they may not have autonomous energy supply at all. Such limitations must be taken into account as a constraint in the design of solutions supporting privacy and security. This basically requires that the proposed solutions must be simple and robust to system failures.

In the paper entitled "*An Overview of Privacy and Security Issues in the Internet of Things*", by Carlo Maria Medaglia and Alexandru Serbanati, the authors provide a survey on the above problems. Both wireless sensor networks and RFID technologies are considered and analysis encompasses both the short term and the long term technologies. Where the short term regards technologies that are already available or will be ready soon; whereas the long term regards technologies that today can only be envisioned. In the context of wireless sensor networks, the problems taken into account are authorization of nodes to join the network so that denial of service attacks are prevented, authentication, and data confidentiality and integrity. In the RFID case, the above problems are also taken into account considering that RFID tags always respond to readers (the user cannot decide when and to whom RFIDs should respond). The analysis suggests that all possible solutions should be set in the standardization effort, which will be the key enabling factor of the IoT technology as a whole.

A more detailed analysis of the privacy issues when RF tags are taken into account is presented in the paper "*Privacy Challenges in RFID Systems*" by Yong Ki Lee, Lejla Batina, and Ingrid Verbauwhede. In fact, RFID systems require a specific study as the solutions that have been recently proposed to guarantee privacy in systems with scarce resources cannot be applied effectively in RFID systems. In fact such solutions do not have the scalability properties, the robustness to cloning, replay, tracking, and DoS attacks, that are required in RFID systems. Accordingly, in the above paper, an analysis of the privacy threats is provided along with the

solutions proposed up to date. Solutions of such threats lay in the design of efficient and effective schemes for data protection (which involve secure algorithms for the exchange of the keys) and authentication. In such an analysis, besides standard solutions, the schemes recently proposed by the research community have been considered as well.

A possible set of solutions to the privacy problems in RFID systems is provided in the paper "*Security and Privacy Protection in Contactless Devices*" by Olivier Savry and François Vacherand. The solution proposed in this paper aims at two different objectives. The first one is to build a mechanism to offer to the user the capabilities to control the information that is released by the RFIDs. More specifically, such a scheme should allow the user to decide and control who can read any of her/his RFID tags as well as when and where this task can be accomplished. Such a management information is handled by a specific device called "Contactless privacy manager". Such device jams the transmissions of an RFID tag when it should not be read. The second objective of the proposed solution is to protect the information transmitted by an RFID. This is achieved by modulating the signal emitted by the RFID with a pseudo-noise. In this way, the RFID can be read and intercepted only by readers that can generate the same pseudo-noise. Accordingly, the seed utilized to generate the above noise can be used as a cryptographic key. The proposed solution has been implemented.

Finally, the paper "*Private Location-Based Information Retrieval via k-Anonymous Clustering*" by David Rebollo-Monedero, Jordi Forné, and Miguel Soriano deals with the problem of location privacy when user localization is a parameter utilized for the provision of context aware service. Indeed, most of the context aware services that will be deployed in the IoT need knowledge of the position of the user. However, the precision of such an information may change depending on the specific application. Accordingly, appropriate solutions should be found able to tradeoff the precision required by the application with the location privacy desired by the user. In this paper, authors propose to use the k-anonymity concept. An appropriate network element called "trusted third party" (TTP) receives location information of all the users. Then, location privacy is achieved by letting the TTP disclose location information with a precision such that the position of more than k users cannot be distinguished from each others. In this way, for example, higher level of privacy can be guaranteed in densely populated areas. The proposed solution uses a modification of the Lloyd algorithm to introduce the right amount of distortion in the location information, while the Levenberg-Marquardt algorithm is applied to adjust the quantization cell size in order to satisfy the privacy constraint.

Cagliari, Italy

Luigi Atzori
Daniele Giusto
Antonio Iera
Giacomo Morabito

Contents

Contributors

Anna Abbagnale INFOCOM Departments, University of Rome "Sapienza", via Eudossiana, 18, 00184 Rome, Italy

Alessio Agneessens Department of Communications, Computer and System Science, University of Genoa, Genoa, Italy

Roberto Alesii Department of Electrical and Information Engineering and Center of Excellence in Research DEWS, University of L'Aquila, 67040 Poggio di Roio, L'Aquila (AQ), Italy

Jasser Al-Kassab SAP Research CEC St. Gallen, SAP (Switzerland) Inc. & Institute of Technology Management, University of St. Gallen, St. Gallen, Switzerland

Olivier Alphand Grenoble Informatics Laboratory, University of Grenoble, Grenoble, France

George Alyfantis Pervasive Computing Research Group, Department of Informatics and Telecommunications, University of Athens, Panepistimioupolis, Illissia, 15784 Athens, Greece

Francesco Amato Dipartimento di Informatica Sistemi e Produzione, University of Roma Tor Vergata, Via del Politecnico, 1, 00133 Rome, Italy

Christos Anagnostopoulos Pervasive Computing Research Group, Department of Informatics and Telecommunications, University of Athens, Panepistimioupolis, Illissia, 15784 Athens, Greece

Pablo Antolin Telefonica I+D, Zaragoza Area, Spain

Luigi Atzori Department of Electrical and Electronic Engineering, University of Cagliari – Italy, 09123 Cagliari, Italy

Enzo Baccarelli INFO-COM Department, "Sapienza" University of Rome, via Eudossiana, 18, 00184 Rome, Italy

Lejla Batina ESAT/SCD-COSIC, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Belgium

Luca Bencini Department of Electronics and Telecommunications, University of Florence, via di Santa Marta, 3, 50139 Firenze, Italy

L. Bergesio Pervasive Radio Technologies Lab, Istituto Superiore Mario Boella (ISMB), Torino, Italy
and
Dipartimento di Automatica e Informatica (DAUIN), Politecnico di Torino, Italy

M. Bernier Grenoble-inp/LCIS, ESISAR, F 26902 Valence, France

Massimo Bertolini Department of Industrial Engineering, University of Parma, viale G.P. Usberti, 181/A, 43100 Parma, Italy

Rahul Bhattacharyya Auto ID Labs, Massachusetts Institute of Technology, Boston, MA, USA

Mauro Biagi INFO-COM Department, "Sapienza" University of Rome, via Eudossiana, 18, 00184 Rome, Italy

Igor Bisio Department of Communications, Computer and System Science, University of Genoa, Genoa, Italy

D. Boschetti Thales Alenia Space Italia, University of Florence, Florence, Italy

Eleonora Bottani Department of Industrial Engineering, University of Parma, viale G.P. Usberti, 181/A, 43100 Parma, Italy

Paolo Braconi Wavecomm S.r.L., Loc. Belvedere, 53034 Colle Val d'Elsa, Siena 53100, Italy

Chiara Buratti WiLAB, DEIS, University of Bologna, Bologna, Italy

Stefano Busanelli Department of Information Engineering, University of Parma, Parma, Italy

Ugo Biader Ceipidor Centro per le Applicazioni della Televisione e delle Tecniche di Istruzione a Distanza (CATTID), University "Sapienza", Rome, Italy

D. Chini Electronics and Telecommunications Department, University of Florence, Florence, Italy

L. Ciofi Electronics and Telecommunications Department, University of Florence, Florence, Italy

Emanuele Cipollone INFOCOM Department, University of Rome "Sapienza", via Eudossiana, 18, 00184 Rome, Italy

Giovanni Collodi Department of Electronics and Telecommunications, University of Florence, via di Santa Marta, 3, 50139 Firenze, Italy

Nicola Cordeschi INFO-COM Department, "Sapienza" University of Rome, via Eudossiana, 18, 00184 Rome, Italy

M. Crotti Università degli Studi di Brescia, Brescia, Italy

Alessio Cucini Wavecomm S.r.L., Loc. Belvedere, 53034 Colle Val d'Elsa, 53100 Siena, Italy

A. Cucinotta Engineering Faculty, University of Messina, Contrada Papardo, S. Sperone, 98166 Messina, Italy

Francesca Cuomo INFOCOM Department, University of Rome "Sapienza", via Eudossiana, 18, 00184 Rome, Italy

D. Dardari WiLAB, DEIS, University of Bologna at Cesena, via Venezia 52, 47023 Cesena (FC), Italy

Luca D'Ascenzo Centro per le Applicazioni della Televisione e delle Tecniche di Istruzione a Distanza (CATTID), University "Sapienza", Rome, Italy

Franco Davoli DIST-University of Genoa, Via Opera Pia 13, 16145 Genoa, Italy

M. De Blasi Department of Innovation Engineering, University of Salento, Lecce, Italy

V. Deepu Grenoble-inp/LCIS, ESISAR, F 26902 Valence, France

E. Del Re Thales Alenia Space Italia, University of Florence, Florence, Italy

Massimiliano Dibitonto Centro per le Applicazioni della Televisione e delle Tecniche di Istruzione a Distanza (CATTID), University "Sapienza", Rome, Italy
and
Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

Davide Di Palma Department of Electronics and Telecommunications, University of Florence, via di Santa Marta, 3, 50139 Firenze, Italy

Marco Di Renzo French National Center for Scientific Research (CNRS) Laboratory of Signals and Systems (LSS) Ecole Supérieure d'Electricité (SUPELEC) 3 rue Joliot-Curie, 91192 Gif-sur-Yvette (Paris), France

Andrzej Duda Grenoble Informatics Laboratory, University of Grenoble, Grenoble, France

Markus Eisenhauer Fraunhofer Institute for Applied Information Technology FIT, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Gianluigi Ferrari Department of Information Engineering, CNIT Research UNIT, University of Parma, Parma, Italy

Alberto Ferri WiLAB, DEIS, University of Bologna, Bologna, Italy

F.H.P. Fitzek Department of Electronic Systems, Aalborg University, Aalborg, Denmark

Christian Floerkemeier Auto ID Labs, Massachusetts Institute of Technology, Boston, MA, USA

Jordi Forné Information Security Group, Department of Telematics Engineering, Technical University of Catalonia (UPC), E-08034 Barcelona, Spain

M. Franceschinis Pervasive Radio Technologies Lab, Istituto Superiore Mario Boella (ISMB), Torino, Italy

Cristian Della Giovampaola Department of Information Engineering, University of Siena, Via Roma 56, 53100 Siena, Italy

Dino Giuli Department of Electronics and Telecommunications, University of Florence, Florence, Italy

Fabio Graziosi Department of Electrical and Information Engineering and Center of Excellence in Research DEWS, University of L'Aquila, 67040 Poggio di Roio, L'Aquila (AQ), Italy

F. Gringoli Università degli Studi di Brescia, Brescia, Italy

F. Guidi WiLAB, DEIS, University of Bologna at Cesena, via Venezia 52, 47023 Cesena (FC), Italy

Stathes Hadjiethymiades Pervasive Computing Research Group, Department of Informatics and Telecommunications, University of Athens Panepistimioupolis, Illissia, 15784 Athens, Greece

Martin Heusse Grenoble Informatics Laboratory, University of Grenoble, Grenoble, France

S. Innocenti Electronics and Telecommunications Department, University of Florence, Florence, Italy

D. Izal Department of Electronic Systems, Aalborg University, Aalborg, Denmark

Corinne Kassapoglou-Faist CSEM, Centre Suisse d'Electronique et de Microtechnique S.A. Jaquet-Droz 1, CH-2002 Neuchâtel, Switzerland

Yong Ki Lee University of California, Los Angeles, 56-125B Engineering IV Building 420 Westwood Plaza, Los Angeles, CA 90095-1594, USA

Dimitrios Kotsakos Pervasive Computing Research Group, Department of Informatics and Telecommunications, University of Athens, Panepistimioupolis, Illissia, 15784 Athens, Greece

Fabio Lavagetto Department of Communications, Computer and System Science, University of Genoa, Genoa, Italy

Bernardo Leal Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

Riccardo Leonardi Department of Electronics for Automation, Signals and Communication Laboratory, University of Brescia, Brescia, Italy

Marco Listanti INFOCOM Department, University of Rome "La Sapienza", Via Eudossiana, 18, 00184 Rome, Italy

Marco Lixia Department of Electrical and Electronic Engineering, University of Cagliari, 09123 Cagliari, Italy

M.L. Lobina Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

Agostino Longo SELEX Sistemi Integrati SpA, Rome, Italy

Antonio Manes Department of Electronics and Telecommunications, University of Florence, via di Santa Marta, 3, 50139 Firenze, Italy

Gianfranco Manes Department of Electronics and Telecommunications, University of Florence, via di Santa Marta, 3, 50139 Firenze, Italy

Mario Marchese Department of Communications, Computer and System Science, University of Genoa, Genoa, Italy
and
DIST-University of Genoa, Via Opera Pia 13, 16145 Genoa, Italy

Antonio Marqués ETRA I+D, Tres Forques 147, 46014 Valencia, Spain

Gaetano Marroccoa Dipartimento di Informatica Sistemi e Produzione, University of Roma Tor Vergata, Via del Politecnico, 1, 00133 Rome, Italy

Marco Martalò WASN Lab, Department of Information Engineering, University of Parma, Parma, Italy

Carlo Maria Medaglia Centro per le Applicazioni della Televisione e delle Tecniche di Istruzione a Distanza (CATTID), University "Sapienza", Rome, Italy

Nicola Blefari Melazzi University of Rome, Tor Vergata, Rome, Italy

V. Mighali Department of Innovation Engineering, University of Salento, Lecce, Italy

Pierangelo Miglioratia Department of Electronics for Automation, Signals and Communication Laboratory, University of Brescia, Brescia, Italy

A. Longo Minnolo Engineering Faculty, University of Messina, Contrada Papardo, S. Sperone, 98166 Messina, Italy

A. Molinaro Università "Mediterranea" di Reggio Calabria, Reggio Calabria, Italy

Maurizio Mongelli DIST-University of Genoa, Via Opera Pia 13, 16145 Genoa, Italy

Maurizio Murroni Department of Electrical and Electronic Engineering, University of Cagliari, 09123 Cagliari, Italy

Cecilia Occhiuzzi Dipartimento di Informatica Sistemi e Produzione, University of Roma Tor Vergata, Via del Politecnico, 1, 00133 Rome, Italy

T. Onali Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

P. Orefice Laboratorio Nazionale di Comunicazioni Multimediali, CNIT, via Cinthia, Monte S. Angelo, 80126 Napoli, Italy

Federica Paganelli National Interuniversity Consortium for Telecommunications (CNIT), Firenze, Italy

S. Palazzo Università di Catania, Catania, Italy

Sooksan Panichpapiboon Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

David Parlanti National Interuniversity Consortium for Telecommunications (CNIT), Firenze, Italy

R.E. Parrinello Department of Electronic Systems, Aalborg University, Aalborg, Denmark

Tatiana Patriarca INFO-COM Department, "Sapienza" University of Rome, via Eudossiana, 18, 00184 Rome, Italy

L. Patrono Department of Innovation Engineering, University of Salento, Lecce, Italy

L. Paura Laboratorio Nazionale di Comunicazioni Multimediali, CNIT, via Cinthia, Monte S. Angelo, 80126 Napoli, Italy
and
Dipartimento di Ingegneria Biomedica, Elettronica e delle Telecomunicazioni, Università di Napoli Federico II, via Claudio 1, 80125 Napoli, Italy

E. Perret Grenoble-inp/LCIS, ESISAR, F 26902 Valence, France

G.P. Perrucci Department of Electronic Systems, Aalborg University, Aalborg, Denmark

M.C. Pettenati Electronics and Telecommunications Department, University of Florence, Florence, Italy

Damien Piguet CSEM, Centre Suisse d'Electronique et de Microtechnique S.A. Jaquet-Droz 1, CH-2002 Neuchâtel, Switzerland

F. Pirri Electronics and Telecommunications Department, University of Florence, Florence, Italy

Valentina Polli INFO-COM Department, "Sapienza" University of Rome, via Eudossiana, 18, 00184 Rome, Italy

Luigi Pomante Department of Electrical and Information Engineering and Center of Excellence in Research DEWS, University of L'Aquila, 67040 Poggio di Roio, L'Aquila (AQ), Italy

Benoît Ponsard Grenoble Informatics Laboratory, University of Grenoble, Grenoble, France

R. Pucci Thales Alenia Space Italia, University of Florence, Florence, Italy

A. Puliafito Engineering Faculty, University of Messina, Contrada Papardo, S. Sperone, 98166 Messina, Italy

David Rebollo-Monedero Information Security Group, Department of Telematics Engineering, Universitat Politècnica de Catalunya (UPC), E-08034 Barcelona, Spain

Antonio Rizzi Department of Industrial Engineering, University of Parma, viale G.P. Usberti, 181/A, 43100 Parma, Italy

C. Roblin ENSTA-ParisTech, 32 Boulevard Victor, 75739 Paris Cedex 15, France

Fabrizio Ronci INFOCOM Department, University of Rome "La Sapienza", Via Eudossiana, 18, 00184 Rome, Italy

L. Ronga Thales Alenia Space Italia, University of Florence, Florence, Italy

Peter Rosengren CNet Svenska AB, Svärdvägen 3B, 182 33 Dandaryd, Sweden

Franck Rousseau Grenoble Informatics Laboratory, University of Grenoble, Grenoble, France

L. Salgarelli Università degli Studi di Brescia, Brescia, Italy

Fortunato Santucci Department of Electrical and Information Engineering and Center of Excellence in Research, DEWS, University of L'Aquila, 67040 Poggio di Roio, L'Aquila (AQ), Italy

Sanjay Sarma Auto ID Labs, Massachusetts Institute of Technology, Boston, MA, USA

O. Savry Commissariat à l'énergie atomique, LETI, 38054 Grenoble, France

A. Scarpiello Laboratorio Nazionale di Comunicazioni Multimediali, CNIT, via Cinthia, Monte S. Angelo, 80126 Napoli, Italy

E. Scuderi Department of Electronic Systems, Aalborg University, Aalborg, Denmark

Odysseas Sekkas Pervasive Computing Research Group, Department of Informatics and Telecommunications, University of Athens, Panepistimioupolis, Illissia, 15784 Athens, Greece

Alexandru Serbanati Centro per le Applicazioni della Televisione e delle Tecniche di Istruzione a Distanza (CATTID), University "Sapienza", Rome, Italy

Manuel Serrano ETRA I+D, Tres Forques 147, 46014 Valencia, Spain

A. Sibille ENSTA-ParisTech, 32 Boulevard Victor, 75739 Paris Cedex 15, France

Miguel Soriano Information Security Group, Department of Telematics Engineering, Universitat Politècnica de Catalunya (UPC), E-08034 Barcelona, Spain
and
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Parc Mediterrani de la Tecnologia (PMT), 08860 Barcelona, Spain

M. Spirito Pervasive Radio Technologies Lab, Istituto Superiore Mario Boella (ISMB), Torino, Italy

M.L. Stefanizzi Department of Innovation Engineering, University of Salento, Lecce, Italy

Lauri Sydänheimo Rauma Research Unit, Department of Electronics, Tampere University of Technology, Kalliokatu 2, 26100 Rauma, Finland

S. Tedjini Grenoble-inp/LCIS, ESISAR, F 26902 Valence, France

Stefano Tennina Department of Electrical and Information Engineering and Center of Excellence in Research DEWS, University of L'Aquila, 67040 Poggio di Roio, L'Aquila (AQ), Italy

Fabrice Theoleyre Grenoble Informatics Laboratory, University of Grenoble, Grenoble, France

Frédéric Thiesse Institute of Technology Management, University of St. Gallen, St. Gallen, Switzerland

Alberto Toccafondi Department of Information Engineering, University of Siena, Via Roma 56, 53100 Siena, Italy

R. Tomasi Pervasive Radio Technologies Lab, Istituto Superiore Mario Boella (ISMB), Torino, Italy

Claudia Tonoli Department of Electronics for Automation, Signals and Communication Laboratory, University of Brescia, Brescia, Italy

Leena Ukkonen Rauma Research Unit, Department of Electronics, Tampere University of Technology, Kalliokatu 2, 26100 Rauma, Finland

F. Vacherand Commissariat à l'énergie atomique, LETI, 38054 Grenoble, France

Ingrid Verbauwhede University of California, Los Angeles, 56-125B Engineering IV Building 420 Westwood Plaza, Los Angeles, CA 90095-1594, USA
and
ESAT/SCD-COSIC, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, Belgium

Roberto Verdone WiLAB, DEIS, University of Bologna, Bologna, Italy

Andrea Volpi Department of Industrial Engineering, University of Parma, viale G.P. Usberti, 181/A, 43100 Parma, Italy

A. Zaia Inquadro s.r.l., Via Nino Bixio, 77, 98123 Messina, Italy

Part I
Networking Issues

Objects Communication Behavior on Multihomed Hybrid Ad Hoc Networks

Bernardo Leal and Luigi Atzori

1 Introduction

According to the "The Internet of Thing" paradigm, physical objects connect to Internet for sharing information about themselves and their surroundings [1]. When the considered objects move around, it is necessary to use wireless means to connect them to Internet. But, when the paths followed by these objects are unpredictable and/or when the objects move away from networks structures, MANET (Mobile Ad Hoc Networks) may be the only way to maintain connection. MANET consists of a number of self-organized mobile nodes or objects with routing capabilities, which may be implemented isolated or connected to structured networks by means gateways [2,3]. The integration of MANETs with fixed infrastructures, as Internet, must be carefully studied to evaluate its capabilities. In such integrated scenarios, commonly known as hybrid ad hoc networks, a MANET can be seen as an extension to the existing infrastructure, whose mobile nodes may seamlessly communicate with those on the fixed network forwarding packets through the gateways found on the edge which joins both types of network.

Much of the MANET research has primarily focused on its isolated performance without considering how it behaves when connected to a fixed network. Performance of hybrid ad hoc networks is strongly impacted by node mobility on the MANET. Two of the aspects that may affect this performance are MANET node address allocation and the dynamic gateway changes. When objects on MANETs move around, they may find themselves on a different MANET subnetwork from where they registered and got their address from, and for that reason, their IP address must be changed accordingly while maintaining ongoing connections and delivering the packets belonging to these connections continuously. After changing address, mobile nodes will be required to use a different gateway to continue forwarding and receiving packets that flow between the MANET and the fixed network. Address

B. Leal (✉) and L. Atzori
Department of Electrical and Electronic Engineering, University of Cagliari – Italy
e-mail: bernardo.leal@diee.unica.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_1,

and gateways changes may cause packet delivery interruption, packet losses, and even connection losses that may probably affect communication between moving objects and fixed nodes.

In this chapter, a comparative evaluation is made between the performance of two scenarios of Hybrid ad Hoc Networks: one in which a proactive protocol is used in the MANET side of the network, like Optimized Link State Routing Protocol (OLSR), and one in which a reactive one is used, like Ad-hoc On-Demand Distance Vector (AODV). In both scenarios, we consider that the interconnection between the MANET and the fixed network will be by means of two or more gateways placed away from each other allowing the formation of different subnetworks; one for each gateway. A mobile object with MANET communication capabilities will be allowed to move from the vicinity of one gateway to the vicinity of the others while engaged in a communication with a host placed on the wired network. Then, estimations of packet losses, delay and jitter are evaluated.

The rest of this work is organized as follows: in Sect. 2, the scenarios to be evaluated are presented, describing how addresses are allocated on MANET objects, how gateways are chosen, and how MANET routing protocols work in this scenarios. In Sect. 3, a detailed description of the scenarios is presented and a conceptual analysis of the events that occur when an object engaged in a communication with a host in the fixed network moves and have to change gateway to continue forwarding packets vs. the fixed network. Finally, on Sect. 4, results and conclusions are presented.

2 The Multihomed Scenarios

Hybrid ad hoc networks, as it is shown on Fig. 1, are composed of three different parts: (1) the fixed network, where hosts remain always in the same subnetwork without changing their address prefixes, and a traditional Internal Gateway Protocol

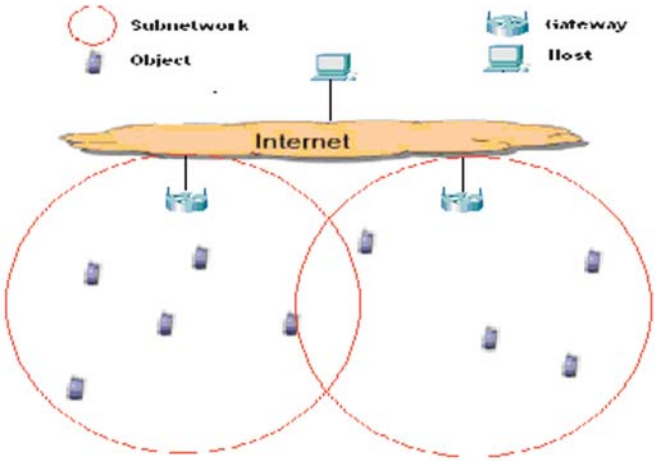


Fig. 1 Hybrid ad hoc network

(IGP) is used to find usable routes. (2) The MANET, where mobile objects may move and change their subnetwork associations and addresses, besides running a MANET routing protocol to find usable routes. (3) The gateways, which are special routers that interconnect the MANET to the fixed network, allowing not only that data packets traverse from one network to the other, but that the routing protocols from each of the networks may share their known routes. It means that gateways must have at least one interface belonging to the fixed network and one interface belonging to the MANET. When two or more gateways connect the MANET to the fixed network, it is referred as Multihomed Hybrid Ad Hoc Networks.

On "The Internet of Things," it is expected that mobile objects participate into information networks without position communication limitations, and if MANET networks are used as a part of its supporting structure, then we should consider the effects that over ongoing communications appears when these objects move from one subnetwork to another, especially those related to address reallocation, dynamic gateway changes, and routing protocol convergence.

2.1 Address Allocation

Address allocation on MANET objects that communicate to the Internet is preferably done by using a stateless autoconfiguration mechanism based on network prefixes advertised by one or more gateways nodes. This solution is adopted because it deals better with network partitions on MANET [4]. With stateless autoconfiguration, mobile nodes set its IP address according to the network prefix announced by the closest gateway. In this way, it is possible the formation of subnets of nodes sharing a common network prefix, A host realizes that it is in a zone belonging to a different subnetwork when it recognizes that its distance to another gateway, measured in route hop counts, is less than that from which it got its current address from. Address reallocation is done dynamically according to object mobility, and thus, routing tables on MANET nodes and gateways will have to adjust their routes and summaries, which may probably cause, connection and packet losses and packet forwarding delay.

2.2 Gateways

The paths used to forward packets between mobile and fixed networks may also affect communication performance. Before setting its address, MANET nodes must select a gateway for traffic forwarding to and from the fixed network. Gateways discovery is associated to the MANET protocol used, thus, it may be done using one of two mechanisms: a reactive one and a proactive one [5,6]. In the reactive version, when an object requires global connectivity, it issues a request message which is flooded throughout the MANET. When this request is received by a gateway, it

sends a message which creates reverse routes to the gateway on its way back to the originator. The proactive approach is based on the periodic flooding of gateway advertisement messages, allowing mobile objects to create routes to the Internet in an unsolicited manner. If objects receive routes to more than one gateway, they choose the closest one, but only on the proactive approach objects may be sure that the selected gateway will remain the closest, since on the reactive approach, gateway updates only occur when its routes are lost. Changing forwarding gateways during ongoing connections will bring time gaps during which packets are not forwarded or are lost. Even more, the connection to the distant host may be lost.

2.3 MANET Protocols

The MANET routing protocol used on hybrid ad hoc network also affects its performance significantly when object move between different subnetworks. Standard MANET protocols may be grouped in two types: Reactive MANET Protocols and Proactive MANET protocols [7]. Reactive protocols discover routing paths only when traffic demands it, and as a result, when there are route changes, trading off longer packet delays in the interest of lower protocol overhead. AODV is an example of reactive protocols. Proactive protocols maintain and regularly update full sets of routing information, trading off greater protocol overhead and higher convergence time in the interest of smaller packet delays. OLSR is an example of a proactive protocol. Paradoxically, reactive protocols tend to take less time than proactive protocols to recover from route losses, especially as a consequence of object mobility. This is because it uses a smaller time to declare a lost route and only cares about recovering specific routes. Each MANET protocol type will react differently when objects move between different MANET subnetworks and find routes to new gateways to keep their ongoing communications active. The important parameter to observe is the time taken for each protocol to reach convergence. To better understand their behavior, a brief description of one protocol of each type follows.

2.3.1 AODV

AODV [8] focuses only on learning about those neighbors that are useful in order to transmit data to a particular destination. To learn about a new destination, a Route Request (RREQ) is broadcast within a specified area, initially set at 1 hop. With each failed Route Request, the broadcast area is increased. When the RREQ reaches an object that has information to the required destination, it responds with a Route Reply message. If an active route fails, a Route Error is sent from the object that has noted the failed link and a new RREQ is initiated. Active routes in AODV are maintained via periodic Hello messages. According to RFC 3,561, Hello messages are transmitted with a frequency of 1 s. If a Hello from an active object is not received

Table 1 Main parameters of the MANET protocols

MANET protocol	Route/neighbor discovery	Identification of route change
AODV	Route request	No Hello within 2 s
	Route reply	
	Hello for active nodes (1 s)	
OLSR	Hello (2 s)	No Hello within 6 s
	Topology control (5 s)	

within 2 s, the route is considered unreachable, a Route Error message is broadcast to all nodes, and another series of Route Requests are broadcast.

2.3.2 OLSR

OLSR [7] is a proactive protocol in which periodic HELLO messages are used to establish neighbor links and to distribute MultiPoint Relays (MPRs), determined by a particular algorithm. Hello messages track link connectivity. Topology Control (TC) messages, distributed by MPRs, propagate link state information throughout the network, and are broadcast periodically as well as when there is a change to the topology. Control traffic consists of periodic hellos and TC messages. Overhead is controlled by MPR broadcast and redistribution of TC messages throughout the network, rather than broadcasts of link state from each router.

The time that each type of protocol takes to help objects discover new gateways, set its addresses, and find adequate routes to given destinations on the fixed network in the presence of object mobility heavily impact hybrid ad hoc networks performance. Table 1 shows main timing values for AODV and OLSR protocols. It can be seen that AODV only keeps routes to requested destinations, reducing thus congestion and routing table size, but most important, AODV takes less time than OLSR to react on the event of lost routes. Even more, AODV is only interested in recuperate that specific lost route and not every possible route.

3 Multihomed Hybrid Ad Hoc Networks Analysis

As shown in Fig. 1, the scenarios analyzed consider the interconnection of a MANET and a fixed network by means of two gateways placed away from each other, providing each one a different network prefix, allowing the formation of two different subnetworks. A mobile object will be allowed to move from one subnetwork to the other following a straight path, while keeping a communication connection with a host placed on the wired network. Packet losses, delay and jitter are evaluated during this transition. The MANET routes announced by the gateways to the fixed network, if necessary, may be summarized in order to reduce frequently update exposure coming from the MANET routing.

The mobile object will set its IP address in correspondence to the public prefix announced by the closest gateway. Alternatively, node addresses may be manually fixed or dynamically autoassigned using private address, which can later be translated to public address by means of NAT servers loaded on gateways. In either case, when an object moves closer to a different gateway from which it got its original address, it must set a new one corresponding to the new subnetwork prefix, and use it to forward packets toward the fixed network throughout the new gateway. On their way back, packets coming toward objects on MANET should enter using the same gateway used by the packets exiting MANET. This is not always true, especially when, to reduce frequently routing update exposure coming from the MANET, route summarization is implemented on gateways, hence reducing granularity on MANET routes. In order to avoid its loosing when return packets try to enter MANET using the wrong gateway, physical links between gateways should be implemented, which will permit packets to find its way vs. the originating object [9].

The objective of this paper is to compare traffic performance for the two types of MANET routing protocols, when a moving MANET object maintains a communication connection whit a node placed on the fixed network, which is connected to the MANET by means of two or more gateways. The considered metrics to evaluate MANET protocol performance are:

- Packet Delivery Ratio (PDR): The ratio of the number of data packets received to the number of data packets transmitted
- End-to-End Delay: The time needed to deliver a packet from the data source to the data destination
- Jitter: Variability of End-to-End Delay

3.1 Scenario 1

AODV. When an object on MANET needs to forward packets vs. the Internet, but does not have a valid route to its destination, it broadcasts a request. This request is forwarded by neighbor objects until a route is found. For destinations outside MANET, gateways, if present, will respond with a route. Among those that respond to, the originating object chooses the closest gateway, from which it also gets its address prefix, which will use to forward its packets. The gateway will forward all packets received from the mobile object toward its destination on the fixed network. Return packets will use the same gateway in its way back to the originating object. When objects move and routes to the gateway get lost, they use new requests to find new ones. New routes may or may not use the same gateway for destinations outside the MANET. In any case, until new routes are found, there will be a time lapse when packets will not be forwarded or will be lost. This time is not always the same, and will depend on the links that are set or lost between mobile objects, but will always be superior to 2 s, which is the time needed before declaring a lost route in AODV. If the new destination route goes throughout a different gateway, then the object will have to change its address before continuing to forward packets. It is also important

to note that since AODV is a reactive protocol, as long as they have a route, mobile objects will not notice if they are closer to a different gateway from which they got their address prefix, thus, they will continue forwarding packets throughout the same gateway, even if they take a longer path, until gateways routes are lost. Being AODV a reactive protocol, it will not generate as much routing traffic as proactive ones, thus it won't be required to summarize MANET routes to reduce exposure over the IGP on the fixed network, hence helping it on finding better return routes.

3.2 Scenario 2

OLSR. Without needing to forward any packet, objects on MANET discover routes to any possible destination by establishing neighborhood relations to some nearby nodes. Besides its known routes, gateways on MANET announce routes to the fixed network as a default route. Mobile objects choose, between those routes going outside MANET, the one going throughout the closest gateway, from which it also gets its address prefix that uses to forward its packets. The gateway will forward all packets received from the mobile object toward its destination on the fixed network. Return packets will use the same gateway in its way back to the originating node. When any mobile object moves, and link connections are added or lost, routes must be recalculated on the whole MANET. New routes going outside the MANET may or may not use the same gateway. In any case, there is a hold time before declaring a route to be lost, in which, packets forwarded using lost routes will also be lost. The time to discover new routes is not always the same, and will depend on the links that are set or lost between mobile objects, but will always be superior to 6 s, which is the time needed before declaring a lost route. Since OLSR is a proactive protocol, any route recalculation on the MANET will make all objects notice if they are closer to a different gateway from which they got its address prefix, so they will have to change it according to the new prefix before continuing to forward packets throughout the new gateway. Finally, OLSR generates so much routing traffic, that MANET route summarization will be required on the fixed network in order to reduce routing exposure over the fixed network IGP, thus decreasing granularity on MANET routes.

4 Results and Conclusions

After evaluation of the two types of MANET routing protocols and how they react when MANET objects move between different MANET subnetworks, the most important characteristics are presented on Table 2. It may be observed that AODV reacts better to object mobility. Although it would not have a route to a given destination right away, as it will have OLSR, when routes are lost, it will recuperate them faster. This is not only because it uses less time to declare lost routes but also

Table 2 Expected behavior for each routing protocol

MANET protocol	Protocol characteristic	Object mobility impact
AODV		Do not require route summarization
		Do not require gateway interlinks
	2 s to declare lost routes	PDR will be smaller
	Only rediscover lost routes	End-to-end delay is bigger
	Minor routing congestion	Jitter will be smaller
OLS		Require route summarization
	6 s to declare lost routes	Require gateway interlinks
	Rediscover every routes	PDR will be bigger
	Major routing congestion	End-to-end delay is slower
		Jitter will be bigger

it recovers only those routes that are needed. Additionally, AODV uses less routing packets to get and maintain its routes, thus creating less congestion.

As a consequence, PDR will be higher on AODV than on OLSR. Packets from an object to a node in the fixed network won't be delivered from the moment that a route to the current gateway is lost until it is rediscovered. In AODV, this time includes 2 s for declaring that route as lost, and some additional time required to find a new route. How many packets are lost also depends on its generation rate and on the object buffer size. On the other hand, OLSR uses 6 s to declare a route as lost, and will take a longer time to find new routes, because all objects must reach the convergence. Additionally, OLSR, besides generating bigger congestion, it will stop forwarding any packets when any route is lost, and not only those aimed to nodes outside the MANET.

End-to-End Delay will be usually longer on AODV because it will not be recognized unconditionally if there is a closer gateway, and may then use longer paths to forward its packets toward the fixed network. However, because AODV does not require the use of summarization, return packets may find shorter routes on the fixed network trajectory, and thus reducing the delay of returning packets, although this may not compensate the delay found on the MANET part path.

Since AODV reacts only when a required route is lost, there will not be as many routing table changes as when OLSR is used. In others words, routes on AODV will last longer, and thus there will be less delay variation. For this reason, Jitter will also be lower in AODV than in OLSR.

We may then finally conclude that it may result more convenient to use a reactive protocol than a proactive one on a MANET whose objects, being part of "The Internet of Thing," are engaged in communication connections and are moving. This is not the case when objects are static, for which proactive protocols where proposed to deliver a better performance. Future work on this area is aimed to evaluate the effects over the fixed network produced by object mobility on multihomed hybrid ad hoc network and to propose alternative solutions.

References

1. Commission of The European Communities (2008) Future networks and the Internet. Early challenges regarding the "Internet of Things". COM (2008) 594. SEC (2008) 2507
2. Engelstad P, Tønnesen A, Hafslund A, Egeland G (2004) Internet connectivity for multi-homed proactive ad hoc networks. First IEEE international conference on sensor and ad hoc communications and networks, 4–7 Oct 2004
3. Ratanchandani P, Kravets R (2003) A hybrid approach to Internet connectivity for mobile ad hoc networks. In: Wireless communications and networking (WCNC 2003), New Orleans, USA, Mar 2003, pp 1522–1527
4. Ros F, Ruiz P, Gomez-Skarmeta A (2006) Performance evaluation of interconnection mechanisms for ad hoc networks across mobility models. J Netw 1(2), 9–17
5. Andreadis G (2002) Providing Internet access to mobile ad hoc networks. In: Proceedings London Communication Symposium, 9–10 September 2002, London
6. Bayer N, Sivchenko D, Xu B, Hischke S, Rakocevic V, Habermann J (2005) Integration of heterogeneous ad hoc networks with the Internet. In: Proceedings of international workshop on wireless ad-hoc networks, 23–26 May 2005, London
7. Kiwior D, Lam L (2007) Routing protocol performance over intermittent links. Military communications conference, In: Proceedings MILCOM Conference, 29–31 October, Orlando, Florida USA
8. Perkins C, Das S (2003) Ad hoc On-Demand Distance Vector (AODV) Routing. IETF RFC 3561, July 2003
9. Spagnolo P, Henderson T (2007) Connecting OSPF MANET to larger networks. Military communications conference, In: Proceedings MILCOM Conference, 29–31 October, Orlando, Florida USA

Classification of Emerging Protocols in the Presence of Asymmetric Routing

Manuel Crotti, Francesco Gringoli, and Luca Salgarelli

1 Introduction

A widely accepted notion of the Internet of Things (IoT) is referred to the possibility of equipping everyday objects with adequate technology to allow them to communicate using TCP/IP with other objects, identify themselves, or even participate to distributed computing [1]. A rapid technological evolution is taking the IoT phenomenon an essential step further: by embedding short-range transceivers into a wide array of everyday items, new forms of communication are growing not only between people and things but also between things themselves. These new devices are expected to produce significant amount of traffic not only locally but also globally, for example when swarms of sensors are probed and accessed remotely through the Internet by their automated controllers.

In order to maintain and improve the way QoS delivery and security are dealt with in the IoT, traffic analysis mechanisms need to be able to cope with the emergence of the new protocols used by "things," and with how such protocols impact the current Internet. At the same time, both economical and technological reasons are behind another change that the Internet is experiencing, and that will be even more accelerated by the advent of the IoT. Asymmetric routing [2, 3], a practice already common in the Internet core, is going to affect parts of the network that are closer to the edges in a few years [4]. In order to maintain their effectiveness in this environment, Internet traffic analysis techniques need to be made robust to the effects of asymmetric routing.

In this chapter, we focus on the two problems described above and present the analysis of a commonly used traffic classification technique with respect to its capabilities of recognizing "unknown" protocols and its ability to operate when only one of the two counter-propagating half-flows is available.

M. Crotti (✉), F. Gringoli, and L. Salgarelli
Università degli Studi di Brescia, Brescia, Italy
e-mail: manuel.crotti@ing.unibs.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_2,

1.1 Research Contributions

In this chapter, we offer two main research contributions. First, we introduce an effective and modular statistical classification technique that optimizes the precision of a statistical classifier in recognizing the emerging (i.e., "unknown") protocols in an asymmetrical routing environment. Then, we investigate on how the technique can be expanded so as to make it work effectively on bidirectional flows.

Second, we analyze on three data sets the differences in classification accuracy when half-flows are considered as opposed to bidirectional ones. The traces we use are quite different with respect to the environment in which they were captured, the type of protocols they contain, and the time at which the captures were made. This should ensure the generality of the results of our analysis.

The rest of the chapter is organized as follows: Sect. 2 describes an optimization technique conceived for the chosen unidirectional classifier. We then introduce in Sect. 3 two bidirectional classifiers that can be obtained combining the results of the unidirectional verdicts. In Sect. 4, we describe a couple of performance parameters that are useful in determining the improvement introduced by the bidirectional classifiers. Section 5 details the datasets we used in our experiments. In Sect. 6, we report numerical results and we analyze them in Sect. 7. Finally, Sect. 8 concludes the paper.

2 Tuning the Unidirectional Classifiers

The classification algorithm that we developed in [5] and [6] has been extended in this work to maximize the capability of the classifier to recognize emerging protocols. To do so, we introduced two free parameters: the acceptance threshold T^i_{acc} for each of the ω_i protocols under observation and the number n_{p} of packets used to compute the *anomaly score*. We choose to fix those parameters by means of an optimization algorithm based on the maximization of the classifier's overall *recall*.

The recall measurement R is a performance parameter widely adopted in the machine learning (ML) literature [7] and, in our context, it is defined as follows:

$$R_i = \frac{T_{\text{p}_i}}{T_{\text{p}_i} + F_{n_i}},$$

where T_{p_i} denotes the true positives that are the half-flows correctly labeled as belonging to protocol ω_i and F_{n_i} denotes the false negatives, i.e., the half-flows incorrectly labeled as not belonging to ω_i.

The value of R_i for each protocol $\omega_i \in \Omega$ (i.e., the set of "known" protocols) is obtained by running the classification algorithm on a set G_Ω composed by flows generated by application protocolsinside Ω that didn't enter the training

phase. Similarly, R_{n+1} is obtained by classifying a set $G_{\omega_{n+1}}$ composed by flows that have been generated by other protocols than those in Ω. Those two sets compose the *optimization set* G:

$$G = G_{\Omega} \cup G_{\omega_{n+1}}.$$

We finally determine T^i_{acc} and n_{p} that maximize the classifier's overall recall by imposing the following cost function:

$$\max_{n_{\text{p}}} \left(\sum_{i=1}^{n} R_i(T^i_{\text{acc}}, n_{\text{p}}) + R_{n+1}(n_{\text{p}}) \right). \tag{1}$$

3 Combining Unidirectional Classifiers

In this section, we propose two methods to combine the outputs of the unidirectional classifier applied to F_{I} and F_{R} (i.e., the two half-flows propagating respectively from Initiator to Responder and vice-versa) to obtain a bidirectional classifier. The first technique aims at the maximization of the classifier's recall, and is based on a *maximum a posteriori* probability (MAP) approach. The second technique aims at the maximization of the classifier's precision P that is defined as follows:

$$P = \frac{T_{\text{p}}}{T_{\text{p}} + F_{\text{p}}}.$$

The precision parameter is also widely adopted in the ML literature [7] and, differently from the recall parameter, increases when the false positives F_{p} decrease and indicates the reliability of the classifier's verdict. In our context, the precision maximization is based on the evaluation of *coincident fault* probability of unidirectional classifiers (i.e., when both classifiers emit the same wrong verdict).

3.1 First Technique: Recall Maximization

Here, we describe how the MAP approach can be applied to the output of the unidirectional classifiers described in Sect. 2.

Starting from the elements that belong to the optimization set G we gather the unidirectional classification results and we build, for each of the $n + 1$ classes, a classification matrix:

$$A^{(\omega_i)} = \begin{bmatrix} a^{(\omega_i)}_{1,1} & \dots & \dots & \dots & a^{(\omega_i)}_{1,n+1} \\ \dots & \dots & a^{(\omega_i)}_{i,i} & \dots & \dots \\ a^{(\omega_i)}_{n+1,1} & \dots & \dots & \dots & a^{(\omega_i)}_{n+1,n+1} \end{bmatrix}.$$

Here, $a_{j,k}^{(\omega_i)}$ represents the percentage of flows belonging to ω_i class marked by the F_I classifier as belonging to ω_j class and marked by the F_R classifier as belonging to ω_k class.

From a statistical point of view, $A^{(\omega_i)}$ can be seen as a joint probability matrix and each $a_{j,k}^{(\omega_i)}$ element represents a conditional probability:

$$a_{j,k}^{(\omega_i)} = P(\hat{\omega}_\mathrm{I} = \omega_j, \hat{\omega}_\mathrm{R} = \omega_k | \omega_i). \tag{2}$$

If the Maximum A-posteriori Probability condition is imposed to the output of the classifier that combines F_I and F_R verdicts, under the assumption that the flows are uniformly distributed across the ω_i classes, the estimated $\hat{\omega}$ class for each i, j would be:

$$\hat{\omega}(j,k) = \arg\max_{\omega_i} \left(a_{j,k}^{(\omega_i)} \right). \tag{3}$$

The behavior of the statistical classifier based on a MAP estimation can be summarized with the following classification matrix $\hat{\Omega}$:

$$\hat{\Omega} = \begin{bmatrix} \hat{\omega}(1,1) & \dots & \dots & \dots & \hat{\omega}(1,n+1) \\ \dots & \dots & \hat{\omega}(i,i) & \dots & \dots \\ \hat{\omega}(n+1,1) & \dots & \dots & \dots & \hat{\omega}(n+1,n+1) \end{bmatrix}.$$

In practice, the input to the above matrix are the two outputs of the unidirectional classifier applied to F_I and F_R. Such outputs serve as indexes to the element of the matrix that indicates the bidirectional classifier's outcome.

3.2 Second Technique: Precision Maximization

The proposed unidirectional classification algorithm, as explained in [5, 6], can assign a half-flow to any of a known class of protocols Ω or mark the flow as *unknown* (i.e., as belonging to the ill-defined class ω_{n+1}). Since we aim at the maximization of precision, the way of combining unidirectional verdicts is straightforward: we assign a flow to a class $\omega_i \in \Omega$ if the classifiers' verdicts on the two halves of the flow agree, otherwise the flow is marked as unknown (i.e., assigned to class ω_{n+1}).

Applying the described Precision Maximization criteria (MaxP), the elements $\hat{\omega}(i, j)$ of the classification matrix $\hat{\Omega}$ would be set up as follows:

$$\hat{\omega}(i,j) = \begin{cases} \omega_i & \text{if } i = j \\ \omega_{n+1} & \text{otherwise} \end{cases}$$

4 Comparing the Unidirectional and Bidirectional Classifiers

In the previous sections, we proposed two different methods for combining the information coming from F_{I} and F_{R} into a bidirectional classifier. Since in this paper we are interested in comparing the results of the bidirectional approaches with the unidirectional ones, here, starting from the literature, we develop a couple of performance indicators that we will use in Sect. 7 to help us with the analysis.

4.1 MAP: Recall Comparison

Basically we need a measure to determine if and how an MAP approach impacts on the Recall parameter with respect to a unidirectional classifier. Since the optimum recall results R_{I} and R_{R} of the unidirectional classifiers are known, they can be used as a lower bound for the recall of combined classification as follows:

$$R_{\min} = \max\left(R_{\mathrm{I}}, R_{\mathrm{R}}\right).$$

The estimation of the overall recall for the MAP approach is simply obtained with the following:

$$R = \frac{1}{n+1} \sum_{i=1}^{n+1} f_{\hat{\Omega}}(\omega_i), \tag{4}$$

where $f_{\hat{\Omega}}(\omega_i)$ represents the recall of the MAP classifier for each class ω_i of the evaluation set and it is derived as follows:

$$f_{\hat{\Omega}}(\omega_i) = \sum_{\substack{j,k=1 \\ \hat{\omega}(j,k)=\omega_i}}^{n+1} a_{j,k}^{(\omega_i)}.$$

Obviously, if $R > R_{\min}$ the bidirectional classifier achieves better results than each of the unidirectional classifiers. In order to give an indication of the improvement of classification accuracy due to the MAP approach we define the parameter Δ_{R}:

$$\Delta_{\mathrm{R}} = \frac{R - R_{\min}}{1 - R_{\min}},$$

normalized in order to give a non-negative value when the MAP approach leads to better results with respect to the unidirectional classification and a maximum value of 1 in case of perfect classification.

4.2 MaxP: Independence of Wrong Verdicts

As previously depicted, the chosen approach for Precision Maximization imposes that a classification result is accepted only if both unidirectional classifiers emit the same verdict. Therefore, a desirable property for the maximization of precision would be a high rate of coincident hits (i.e., the two classifiers emit the correct verdict) and a low rate of coincident fault. This property holds when the classifiers' faults are independent or, at least weakly dependent, therefore a measure of independence for the faulty classification is needed.

Several works of literature such as [8, 9] adopt the so-called *diversity measures* in order to determine the degree of dependence of classifiers. One of them is the *double fault measure* Df (defined by Giacinto et al. [10]) that allows to determine the rate of double fault of two-class classifiers.

Since in this paper we are working with a couple of multiple-class classifiers that can emit two types of verdict (i.e., the pattern belongs to $\omega_i \in \Omega$ or it belongs to ω_{n+1}) we need to refine the definition of double fault measure given in. [10] taking into account the fact that two classifiers can fail emitting two distinct wrong verdicts. We are interested in the analysis of the frequency of coincident faulty verdicts, therefore we state that a *coincident fault* happens when both unidirectional classifiers emit the same wrong verdict. Hence, the frequency of coincident fault $\mathrm{Df}(i)$ for the ith class is:

$$\mathrm{Df}(i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{j,j}^{(\omega_i)},$$

and then the overall coincident fault frequency is simply derived by the following:

$$\overline{\mathrm{Df}} = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathrm{Df}(i).$$

The degree of dependency of the classifiers is obtained by comparing $\overline{\mathrm{Df}}$ with the coincident fault probability of independent classifiers $\overline{\mathrm{Df}_0}$ and the coincident fault probability of positively correlated classifiers $\overline{\mathrm{Df}_1}$. In our case, the behavior of F_{I} and F_{R} classifiers are known and the probability that both classifiers assign a pattern to a class are:

$$P_0(i,j) = P_0(\hat{\omega}_{\mathrm{c}} = \hat{\omega}_{\mathrm{s}} = \omega_j | \omega_i) = P(\hat{\omega}_{\mathrm{c}} = \omega_j | \omega_i) \cdot P(\hat{\omega}_{\mathrm{s}} = \omega_j | \omega_i),$$

if the two classifiers are independent and:

$$P_1(i,j) = \min\left(P(\hat{\omega}_{\mathrm{c}} = \omega_j | \omega_i), P(\hat{\omega}_{\mathrm{s}} = \omega_j | \omega_i)\right),$$

in case the classifiers are correlated. Therefore, the probability of coincident fault for the ith class can be estimated as follows:

$$Df_l(i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} P_l(i, j). \tag{5}$$

Here, $l = 0$ gives the double fault measure for independent classifiers and $l = 1$ is related to the completely dependent classifiers. The estimation of overall coincident fault probability is therefore:

$$\overline{Df_l} = \frac{1}{n+1} \sum_{i=1}^{n+1} Df_l(i),$$

where l assumes the same meaning of (5).

Finally, the parameter Δ_P can be obtained as follows:

$$\Delta_P = 1 - \frac{\overline{\mathrm{Df}} - \overline{\mathrm{Df}_0}}{\overline{\mathrm{Df}_1} - \overline{\mathrm{Df}_0}}.$$

Here, Δ_P assumes nonnegative values. If $0 < \Delta_P < 1$ the classifiers are positively correlated in emitting the wrong verdict, otherwise (i.e., $\Delta_P > 1$) they are negatively correlated.

5 Datasets

We base our experimental analysis on the results obtained applying the classifiers described in this paper to three different datasets, as described below.

5.1 UNIBS Dataset (2006)

The packet traces from this set were collected in early 2006 at the border router of our Faculty network. Having full monitor access to this router, we can apply pattern-matching mechanisms to assess the actual application that has generated each TCP flow, in some cases with the help of manual inspection. Because of this, we consider the traffic information derived from UNIBS relatively reliable with respect to the preclassification, i.e., with respect to knowing, independently from our classifier, which application generated each flow ("ground truth").

Our network infrastructure comprises several 1,000 Base-TX segments routed through a Linux-based dual-processor box and includes about a thousand

workstations with different operating systems. All the traffic traces were gathered on the 100 Mb/s link connecting the edge router to the Internet for a period of 3 weeks: a total of 50 GB of traffic was collected by running Tcpdump [11] for 15 every hour.

According to the optimization procedure, three different and independent datasets have been collected during the course of several weeks:

- A training set composed of six protocols. For each protocol, we fixed a maximum amount of 20,000 flows. We chose to train the classifier with the protocols a network administrator would be interested to monitor, such as mail, web, chat, and file transfer.
- An optimization set composed of the six protocols mentioned above and another set of flows, $G_{\omega_{n+1}}$ that have not been generated by the applications that appear in the training set. To obtain $G_{\omega_{n+1}}$, we simply collected the network traffic at the edge router and then we extracted the "certified" G_{Ω} flows. The remaining flows compose $G_{\omega_{n+1}}$. A payload inspection of flows composing $G_{\omega_{n+1}}$ for the UNIBS dataset revealed a wide variety of protocols, such as Kazaa, Gnutella, SSH, Netbios, ...
- An evaluation set $E = E_{\Omega} \cup E_{\omega_{n+1}}$ composed by the same protocols of the optimization set and used to verify the classifier's ability to recognize protocols different than those used during the training phase.

5.2 LBNL Dataset (2004–2005)

The LBNL traffic traces we used were collected at the Lawrence Berkeley National Laboratory under the Enterprise Tracing Project [12]. The packet traces were obtained at the two central routers of the LBNL network and they contain more than one hundred hours of traffic generated from several thousand internal hosts. The traffic traces are public, but they are completely anonymized, so ascertaining the "ground truth" on the application behind each recorded flow is not possible. Therefore, for this set, we built protocol sets according to the TCP destination port number of each flow, an accepted practice in these cases [13, 14].

We used the traffic traces captured on December 15 and 16, 2004 to obtain the training and the optimization sets and those captured on January 6 and 7, 2005 to build the evaluation set. Once again we performed the training by using the most frequently used port numbers in the dataset.

5.3 CAIDA Dataset (2002)

We built this data set starting from 3 hour long traces obtained by the Cooperative Association for Internet Data Analysis (CAIDA) [15], and collected at the AMES

Internet Exchange (AIX) along an OC48 link on August 14, 2002. We used flows extracted from the first hour (corresponding to the interval 16.15–17.00 UTC) to build the training set the optimization set and from the third hour (18.00–18.10 UTC) to build the evaluation set. As for the previous set, these traces are also anonymized, so port numbers are used as indicators of each protocol. The selection of flows composing the training, optimization and evaluation sets has been made following the same approach adopted for the LBNL and UNIBS datasets.

6 Results

6.1 Unidirectional Classification

Here, we report the classification results obtained by tuning the unidirectional classification algorithm with the optimum parameters. For each of the three datasets, we show the overall recall, the misclassification rate for untrained protocols (i.e., $\omega_{n+1} \rightarrow \Omega$), and the misclassification rate for trained protocols (i.e., $\omega_i \rightarrow \omega_j$).

As it can be seen in Table 1, the classifier takes the right decision in almost 90% of cases for the UNIBS and LBNL dataset after the analysis of the first three F_{I} or F_{R} data packets. A small portion of traffic belonging to trained protocols is misclassified, but a noteworthy portion of traffic belonging to the ω_{n+1} class is incorrectly marked as belonging to the Ω set. The CAIDA dataset exhibits the worst classification results since the classifier reaches, in the best case, only the 83% of recall and a large portion of ω_{n+1} traffic is labeled as belonging to Ω set.

6.2 Bidirectional Classification

Here we show classification results obtained by applying the two bidirectional classification techniques described in Sect. 3.

As can be seen in Table 2 where the bold numbers evidence a classification verdict that surpasses the best results obtained using a unidirectional classifier, the effects of combining classification verdicts a positive impact on the Recall parameter when an MAP approach is adopted (it is always greater than 90% for each of the

Table 1 Unidirectional classification: optimal results obtained after the analysis of the first three data packets

	CAIDA		UNIBS		LBNL	
	F_{I}	F_{R}	F_{I}	F_{R}	F_{I}	F_{R}
Recall	83.4	81.4	89.0	87.0	90.8	89.7
$\omega_{n+1} \rightarrow \Omega$	60.9	95.8	15.2	30.5	35.5	24.1
$\omega_i \rightarrow \omega_j$	3.6	5.5	3.6	2.8	2.7	5.2

Table 2 Classification results obtained with the two bidirectional classifiers

	CAIDA		UNIBS		LBNL	
	MaxP	MAP	MaxP	MAP	MaxP	MAP
Recall	87.8	91.4	84.1	93.3	88.4	92.4
$\omega_{n+1} \rightarrow \Omega$	5.1	10.1	5.1	17.6	13.5	36.8
$\omega_i \rightarrow \omega_j$	0.8	1.7	0.1	3.6	0.2	2.0
$\Delta_P \vert \Delta_R$	0.719	0.478	0.845	0.394	0.388	0.172

three network environments) and bring to substantive reduction of misclassification rates choosing a MaxP approach that, in the worst case, decreases more than 40% and in the best case of over 90% (CAIDA dataset, $\omega_{n+1} \rightarrow \Omega$). Finally, the last line of the table shows the values of the comparison parameters Δ_P and Δ_R that will be taken into account in the next section, where the unidirectional and bidirectional classification results will be compared.

7 Discussion

Looking at the results obtained for unidirectional and bidirectional classification, the impact of asymmetric routing on the correctness of classifier's verdict is clearly visible: we are losing both in classification accuracy and precision when we choose to (or we are forced to) deploy a unidirectional classifier.

In the following sections, first we analyze the overall results evidencing the differences between unidirectional and bidirectional classifiers on the different network environments, then we deepen the analysis on each network environment showing how the classification accuracy, for different protocols, often depends on the direction of half-flows.

7.1 Overall Results

The results obtained by combining unidirectional classifiers are, for the CAIDA dataset, definitely better than those obtained with unidirectional classification. As shown in Table 2, with a MaxP approach, the misclassification of flows belonging to untrained protocols falls from 60% to 5%. This behavior, due to the substantial incorrelation between F_I and F_R classifiers in emitting wrong verdicts is well captured by the comparison parameter Δ_P that measures the tendency of the unidirectional classifiers in emitting the same wrong verdict. This fact reflects also on the classification results obtained applying an MAP approach: the misclassification rates substantially decrease and the overall recall that increases from 83% to 91% is evidenced by Δ_R parameter that, as explained in Sect. 4, measures the recall improvement between unidirectional and bidirectional classification results.

Similar values of the comparison parameters Δ_P and Δ_R hold for the UNIBS dataset, and as a consequence, a similar behavior is observed in terms of improvement of overall recall and decrease of overall misclassification rate. Obviously, since the accuracy of the unidirectional classifiers of the UNIBS dataset is greater than the one of the CAIDA dataset (e.g., the misclassification rate of ω_{n+1} traffic settles around 30% in the worst case), the accuracy improvement is not as remarkable as the one obtained for the CAIDA network environment.

On the other hand, the low value of comparison parameter Δ_P computed for the LBNL dataset evidences a substantial correlation between the two classifiers' verdicts. This reflects on the classification results of the bidirectional classifiers: applying a MaxP approach, the misclassification rate of ω_{n+1} traffic is not dramatically different from the one obtained by the "best" unidirectional classifier (it shifts from 24.1% to 13.5%) and the MAP approach slightly improves the classification results in comparison to the ones of the F_I unidirectional classifier.

7.2 On the Impact of Contrasting Classifiers' Verdicts

In the previous section, we stated that the correlation between F_I and F_R classifiers' verdicts can impact on the results of bidirectional classification. Here, we deepen the analysis showing what could happen if classifiers' verdicts differ.

In Table 3 for each of the three analyzed dataset, we show a couple of protocols that reach high accuracy in classification (in terms of precision and recall) just for one traffic direction.

It can be seen that the pairing techniques show classification results that, at least, are comparable with the best results obtained for unidirectional classification (e.g., the results shown for the LBNL dataset), but if the classifiers are not strictly correlated in emitting the classification verdicts, as in the case of ω_{n+1} traffic for the CAIDA dataset or HTTP traffic for the UNIBS dataset, the accuracy of the bidirectional classification surpasses the best unidirectional classifier results.

Once more we are demonstrating that when unidirectional classifiers exhibit uncorrelated behaviors, an improvement in classification results could be expected if

Table 3 Example of different accuracy in unidirectional classification

CAIDA						
	F_I		F_R		Joint	
Port	R	P	R	P	MAP	MaxP
21	0.91	0.92	0.99	0.86	0.98	0.97
ω_{n+1}	0.39	0.46	0.04	0.10	0.90	0.68

UNIBS						
	F_I		F_R		Joint	
Port	R	P	R	P	MAP	MaxP
Torrent	0.89	0.98	0.69	0.87	0.90	1.00
http	0.80	0.93	0.93	0.85	0.97	0.95

LBNL						
	F_I		F_R		Joint	
Port	R	P	R	P	MAP	MaxP
110	1.00	0.97	0.92	1.00	0.99	1.00
993	0.89	0.87	0.94	0.92	0.93	0.93

classifiers' verdicts are combined. We are still not able to determine which network conditions make the unidirectional classifiers emit uncorrelated verdicts, but we are able to determine the degree of correlation between those verdicts and then, we can qualitatively estimate the performance improvement that could be expected combining the classification verdicts.

8 Conclusions

In this chapter, we have proposed a supervised statistical traffic classifier based on the analysis of a small set of features of the network traffic. Our aim is the development of an effective classification system that can deal with the continuous growth of emerging communication protocols. Such growth, that is typical of IoT scenarios, requires a classification system that should, at least, be able to recognize if a communication belongs to a set of services or it belongs to a new protocol in order to implement on the (un)classified flows QoS policies or security policies.

In order to deal with another emerging practice, i.e., asymmetric routing, that is spreading even on the last legs of Internet connectivity, we have also analyzed how the precision of the statistical traffic classifier can be affected by the availability of both halves of flows during the decision process. Experimental results support a hypothesis that is quite easy to understand: the classifier that can work on flows in their entirety achieves better results than one that is forced to operate on half-flows only. However, reducing our analysis to a mere empirical proof of this assumption, would be a mistake.

First of all, the numerical comparison between the unidirectional and bidirectional classifiers points out that the loss of accuracy when only half-flows are available is much more substantial in terms of increased false positives as opposed to decreased true positives.

Second, the accuracy loss of unidirectional classifiers varies through the different network environments: the more heterogeneous a network environment is, the less the unidirectional classification is precise.

Finally, the analysis points out that some protocols seem to exhibit more statistical correlation between the two half-flows composing each session than others. We think that this fact needs to be taken into consideration when designing mechanisms that measure traffic flows, including traffic classifiers: certain protocols are more amenable than others at reducing their observation to only one direction of traffic.

References

1. Rellermeyer JS, Duller M, Gilmer K, Maragkos D, Papageorgiou D, Alonso G (2008) The software fabric for the internet of things. In: IOT, pp 87–104.
2. Mao ZM, Qiu L, Wang J, Zhang Y (2005) On AS-level path inference. In: SIGMETRICS '05: Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, ACM, New York, pp 339–349.

3. He Y, Faloutsos M, Krishnamurthy S (2004) Quantifying routing asymmetry in the internet at the as level. In: Proceedings of the GLOBECOM 2004 Conference, IEEE Computer Society Press, Dallas, TX.
4. The Cooperative Association for Internet Data Analysis (CAIDA) – Observing routing asymmetry in Internet traffic. http://www.caida.org/research/traffic-analysis/asymmetry/.
5. Crotti M, Dusi M, Gringoli F, Salgarelli L (2007) Traffic classification through simple statistical fingerprinting. ACM SIGCOMM Comput Commun Rev 37(1):5–16.
6. Dusi M, Gringoli F, Salgarelli L (2008) IP traffic classification for QoS guarantees: the independence of packets. In: Proceedings of The 1st IEEE International Workshop on IP Multimedia Communications (IPMC 2008), St. Thomas, U.S. Virgin Islands, August 2008.
7. Witten IH, Frank E (1999) *Data mining: practical machine learning tools and techniques with java implementations*. Morgan Kaufmann, San Francisco, CA.
8. Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 51(2):181–207.
9. Chen D, Hua D, Sirlantzis K, Ma X (2005) On the relation between dependence and diversity in multiple classifier systems. In: ITCC '05: proceedings of the international conference on information technology: coding and computing (ITCC'05), Vol I. IEEE Computer Society, Washington, DC, pp 134–139.
10. Giacinto G, Roli F, Fumera G (2000) Design of effective multiple classifier systems by clustering of classifiers. In: Proceedings of ICPR2000, 15th International Conference on Pattern Recognition, vol 2. pp 3–8.
11. Tcpdump/Libpcap. http://www.tcpdump.org.
12. LBNL/ICSI Enterprise Tracing Project. http://www.icir.org/enterprise-tracing.
13. Karagiannis T, Broido A, Faloutsos M, Claffy KC (2004) Transport layer identification of P2P traffic. In: IMC'04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, ACM Press, New York, pp 121–134.
14. Karagiannis T, Broido A, Brownlee N, Claffy KC, Faloutsos M (2004) Is P2P dying or just hiding? In: Proceedings of the GLOBECOM 2004 Conference, IEEE Computer Society Press, Dallas, TX.
15. The Cooperative Association for Internet Data Analysis (CAIDA). http://www.caida.org.

Performance Evaluation of Routing Protocols in WSNs Arranged in Linear Topologies

Luca Bergesio, Mirko Franceschinis, Maurizio Spirito, and Riccardo Tomasi

1 Introduction

The application of Wireless Sensor Network (WSN) technology to many heterogeneous fields, such as environmental monitoring, control and automation, logistics, assisted living, and e-health, has widened very much recently. Intelligent Transportation Systems (ITS) is another area where the use of low-consumption wireless devices has shown a promising interest due to cable replacement opportunity and reduced system maintenance costs.

Within the ITS context, monitoring vehicular traffic and detecting road accidents along a road section through WSN systems is one of the most attractive applications. The basic idea is to deploy a number of sensor nodes along the roadside, at regular distances from the predecessor and from the successor, thus forming a linear chain. In order to detect the transit of vehicles and car crash events, each node should be equipped with suitable sensors: e.g., accelerometers, pyrometers, and magnetic sensors. Traffic information locally achieved by a sensor node should be delivered to a sink node, positioned at one edge of the chain, which possibly performs data fusion operations on the pieces of traffic information coming from the many sensor nodes [1].

Each sensor node and the sink node are equipped with a radio transceiver making them able to communicate with each other. However, while low power consumption is a desired feature characterizing WSN technology, it also implies constrained node transmission power and consequently reduced radio coverage when compared with the typical length of the road section under monitoring. Definitively, multihop communication is necessary to let the sink node receive data packets from distributed sensor nodes.

The investigation of networks arranged in linear topologies, the just mentioned WSN as well as the metropolitan network standard DQDB based on a two-level grid, is very interesting from both practical and theoretical viewpoints. On the one hand,

L. Bergesio, M. Franceschinis, M. Spirito (✉), and R. Tomasi
Pervasive Radio Technologies Lab, Istituto Superiore Mario Boella (ISMB), Torino, Italy
e-mail: guglielmo@ismb.it

L. Bergesio
Dipartimento di Automatica e Informatica (DAUIN), Politecnico di Torino, Italy

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_3,

WSN applications inducing linear topologies are common in the real world, and the vehicular traffic monitoring mentioned so far is only an example. Other application fields could be the structural monitoring of infrastructures like railway bridges [2] and the industrial process monitoring of a production line. On the other hand, linear topologies are characterized by symmetries making possible the derivation of closed-form results through mathematical modeling analysis. Linear topologies play a central role in several works in the literature. The objective in [3] is to determine the best deployment of a WSN whose monitored area is known, the reference performance metric being the network lifetime. The goal of [4] is the simulation analysis and comparison of some routing protocols. [5] presents an analytical model to investigate the performance limit of a WSN as a function of the number of nodes, focusing on MAC protocols. The authors of [6] deal with the combination of directional antennas and suitable MAC protocols in linear WSNs and explicitly mention roadside or highway scenarios as reference applications. A cross-layer methodological study based on convex programming is presented in [7]. Finally, a linear topology is the scenario considered in [8], where an improved version of S-MAC protocol is introduced.

Inspired by the relevance of linear topologies and motivated by vehicular traffic monitoring through roadside WSN deployment as reference application (in particular, see [1] and the EU-funded SAFESPOT Integrated Project [9]), in this paper we propose three routing algorithms designed for WSN systems arranged in linear topologies and discuss their performance in terms of information delivery success rate and delay. The three algorithms, called *Single Hop (SH)*, *Limited Flooder (LF),* and *Hopefully Longest Jump First (HLJF)*, can be categorized as geographic routing: they are all based on packet forwarding rules exploiting the strict correlation between node address and relative position in the linear topology. The first two algorithms, *SH* and *LF*, are simplistic and are used as benchmarks. The performance evaluation is carried out experimentally. The experimental test-bed is based on an indoor WSN consisting of 11 Telos motes, on which the three routing strategies have been implemented.

The chapter is organized as follows. Section 2 introduces the system in all its aspects, including WSN deployment, test application, and performance metrics. A detailed description of the proposed routing protocols is provided in Sects. 3, 4, and 5, respectively. In Sect. 6, experimental results are presented and protocols performance commented as well. Finally, Sect. 7 briefly sketches future work and concludes the paper.

2 System Description

2.1 WSN Deployment

We consider a WSN composed of $N+1$ nodes: N sensor nodes, addressed from 1 to N, and 1 sink node whose id is 0. In light of the network application illustrated in Sect. 2.2, hereafter sensor nodes and the sink node are referred to respectively as Slave Nodes (SNi, where $i = 1, \ldots, N$) and the Master Node (MN). We assume

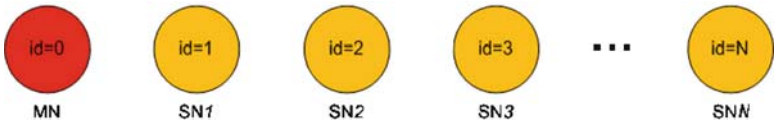


Fig. 1 Schematic representation of a WSN composed of $N+1$ nodes, 1 MN and N SNs, arranged in a linear topology

that the monitored area can be considered monodimensional, i.e., nodes are deployed along a line, in increasing order of address, as shown in Fig. 1.

2.2 Test Application

The performance of routing protocols is experimentally evaluated based on a test application implemented on WSN nodes. In order to estimate the end-to-end communication effectiveness between the MN and the many SNs, the following simple querying paradigm is utilized. The MN interrogates SNi by transmitting a query packet destined to the final recipient SNi. The query packet is forwarded by intermediate nodes until, hopefully, it is received by SNi. Depending on the specific multihop algorithm implemented on nodes, the query packet follows one or multiple paths joining MN to SNi. If SNi receives the query packet, it reacts by sending back a reply packet to the MN. Since packet losses can be experienced along paths, it is not guaranteed that a packet round trip concludes successfully. The MN progressively queries sensor nodes in sequence, from SN1 to SNN. Nonetheless, query cycles come one after another in order to collect large sets of data samples. The test application is run until a statistically sufficient number of samples is collected and the end-to-end packet delivery success rate can be estimated as the fraction of round trip successfully completed. A constant time interval, whose duration is longer than the maximum round trip time experienced, elapses between any two consecutive queries.

2.3 System Cares and Assumptions

A number of system cares and assumptions, discussed in the following, are taken and made in order to simplify the analysis of experimental results.

- The distance between any two adjacent WSN nodes, MN and SN1 or SNi and SN$i+1$, for all $i = 1, \ldots, N-1$, is constant in the deployment of our WSN system inside ISMB Lab.
- Communication channel features are supposed not to vary along time and space, or, at most, the time dependence could be very slow. This appears reasonable in the indoor environment where we deployed our WSN, although long duration

test sessions inside ISMB Lab have exhibited appreciably different channel conditions when comparing day vs. night and working vs. holiday days.
- All nodes have equivalent hardware performance, in particular referring to transmission power, receiver sensitivity, and on-board antennas. Even if this sounds as an obvious assumption, it is worth noting that nominally identical hardware components could exhibit significantly different performance due to typical tolerances and functioning ranges.
- The Free Space Path Loss (FSPL) model can be adopted to characterize node-to-node signal propagation. This model neglects the impact of the specific environment and represents a very simplified model, particularly in indoor scenarios. On the other hand, the ISMB Lab in which tests have been conducted is overall space homogeneous.
- When running experiments in an indoor space-constrained scenario, Telos transmission power is set to the minimum value, i.e., −25 dBm, in order to severely limit the radio coverage of a node and to balance the typically short distance among nodes.

2.4 Performance Evaluation Metrics

Experimental analysis is based on the joint evaluation of the two following performance metrics:

- $P_{e2e}(n)$: The end-to-end information delivery success rate is defined as the probability that, when querying SNn, the MN receives back a reply packet from that sensor node. This metric exists for all $n = 1, \ldots, N$. It is empirically calculated as the ratio between the number of reply packets the MN receives from SNn and the total number of queries destined to SNn, provided that the test session duration is long enough.
- $D_{e2e}(n)$: The end-to-end delay is defined as the time elapsed from the instant when the MN sends a query packet to the final recipient SNn and the instant when the MN receives a correspondent reply packet from SNn. Even in this case the definition must be extended to all $n = 1, \ldots, N$. We take many different samples of $D_{e2e}(n)$ by time-stamping via software, at MN side, the two events in correspondence of each successful end-to-end communication. Thus, we are able to estimate the distribution of the random variable $D_{e2e}(n)$.

3 Single Hop Algorithm

Single Hop (SH) is a simplified reference routing scheme and works as follows. When the MN originates a query packet specifying SNn as the final recipient, it also explicitly indicates SN1 as next-hop destination. If receiving the packet, only SN1 forwards it, by keeping SNn as the final destination and expressing SN2 as

next-hop recipient. On the contrary, any other SN receiving the packet simply discards it. Generalizing the procedure, SNi forwards the packet to SN$i+1$, the only node authorized to iterate the process, while ..., SN$i-1$, SN$i+2$, ..., if receiving, discard the packet. The functioning is symmetrical for the reply packet originated by SNn and addressed to the MN: when forwarding the packet, SNi indicates SN$i-1$ as next-hop destination while other possibly receiving nodes would discard it. No packet retransmission procedure is supported to recover from packet losses.

4 Limited Flooder Algorithm

The *Limited Flooder* (LF) algorithm is a simplified one-dimensional version of the Directed Flood-Routing approach [10], based on the simple idea of flooding the network with multiple copies of the same packet in order to increase the probability that it could successfully reach the final recipient. However, it is well known that the basic packet flooding approach cannot be sustained in a large network being responsible for exponential growth of the number of packets overall transmitted.

The LF algorithm partially reduces the actual number of packet copies in two ways. First of all, when receiving a packet whose final destination is different from itself, a node forwards it in broadcast mode only if it is the first time that packet is received. This is made possible by reserving a field of the header packet for reporting a packet sequence number.

In addition, the correlation between node address and relative physical node location in the network is exploited by letting a node forward the packet only if its position is closer to the final destination than the last relay node. To this aim, for any p and q, the distance between SNp and SNq is calculated as $|p-q|$, the absolute value of the difference of nodes addresses. Note that backward jumps may occur: for instance, at least theoretically, the path followed by a query packet addressed to SN7 could result as ... –4–9–7 since SN9 is closer to SN7 than SN4 is. In fact, backward jumps have been regularly observed during real test sessions, even if limited to one-hop neighbors of the final destination. These events randomly happen and they depend on nodes hardware efficiency.

Definitively, when receiving from SNi a packet originated by the MN, addressed to SNn and identified by sequence number k, SNj forwards it in broadcast mode only if $|n-j| > |n-i|$ and SNj has never received the same packet before. In order to avoid the possibility of backward jumps, an alternative forwarding rule could be proposed as follows: SNj forwards a packet received by SNi and destined to SNn if $i < j < n$.

Nonetheless, the number of packet copies generated (and possibly received, due to the broadcast mode) can still remain very large and, moreover, it is likely that several nodes try to access the shared medium almost at the same time to transmit a packet. This means that collisions could occur more frequently and that every node, endowed with poor resources, could be overloaded.

To conclude, it is worth noting that, even if no packet retransmission procedure is foreseen, the final recipient could receive multiple copies of the same packet and that any duplicate is anyway discarded.

5 Hopefully Longest Jump First Algorithm

The *Hopefully Longest Jump First* (HLJF) is a novel algorithm that tries to take the most desirable features of SH and LF algorithms, while at the same time overcoming their expected weaknesses. In particular, it pursues two objectives. On the one hand, it aims to achieve more reliable communication and to support network scalability in terms of packet delivery success making use of acknowledgments on a link-basis. Clearly, this is obtained to the detriment of longer average end-to-end delays. On the other hand, HLJF algorithm aspires to minimize the number of intermediate forwarders along the source-destination path by forwarding a packet as close to the destination as possible. This forethought could have beneficial effect for reducing delays.

According to HLJF algorithm, when SNi receives a query packet addressed to SNn, $n > i$, SNi forwards the packet to its current *Farthest Reliable Neighbor Node* (FRNN). The FRNN can be any SN whose address is larger than the current forwarder (i.e., concerning SNi, its FRNN could be SN$i+1$, SN$i+2$, ...). Note that if the FRNN address exceeds n, the packet is directly transmitted to SNn.

The core of the protocol resides in the procedure for updating the node FRNN. Let SN$i+k$ be the current FRNN of SNi for some $k > 0$. Then, SNi forwards the query packet to SN$i + k$ which, in its turn, confirms the correct packet reception by sending back an acknowledgment (ack, from now on) to SNi. If the ack is received, no FRNN update is made and the forwarding procedure is shifted on SN$i + k$. On the contrary, if either the query packet is not received by SN$i + k$ or SNi does not receive back the ack, after a Time-Out SNi sets SN$i + k - 1$ as its current FRNN and retransmits (only once) the packet. This couple of actions, FRNN reduction and packet retransmission, is potentially protracted until SN$i + 1$ becomes the current FRNN of SNi. In case not even SN$i + 1$ turns out to be reliable, the path is interrupted and the packet is definitively lost.

So far, FRNN refreshments have only concerned downgrading operations, which always proceed with unitary decrements. Obviously, increments, which are not necessarily unitary, are concerned too. The rule is quite simple and exploits the shared nature of the radio channel. Each time SNi receives a packet (indifferently query packets, reply packets and acks), regardless of SNi being or not the real packet destination, it checks whether the packet sender (not the original source) address is larger than the current FRNN and, if that's the case, the FRNN is consequently updated.

Actually, SNi keeps the addresses of two distinct FRNNs: beside the one described above, a similar one is referred when forwarding reply packets. Since reply packets flow in the opposite direction, from a certain SN to the MN, the address of the related FRNN is always smaller than the current forwarder. Nonetheless, note

that reply packets and related acks influence the updating of both the FRNNs, the one referred for reply packets forwarding as well the one concerning query packets. In the following paragraphs, we focus our attention on the FRNN regarding query packets, however extensions are immediate when considering the twin FRNN.

The behavior in two particular circumstances needs to be specified in order to accomplish the protocol description. The former situation concerns the initialization phase, when SNi completely ignores how reliable the direct communication with its neighbor nodes is. The latter happens when query packet forwarding to SNi + 1, the current FRNN of SNi, fails. When indifferently one of these circumstances occurs, SNi conventionally sets itself as FRNN. Finally, broadcast mode is selected as temporary forwarding strategy when SNi downgraded its FRNN to itself. In other words, from the relay node perspective, HLJF and LF algorithms coincide in such circumstances.

Setting the Time-Out is tricky: it should be large enough to allow the reception of acks avoiding useless retransmissions; however, too large values would result in high end-to-end latencies and overall protocol inefficiency. We experimentally measured the typical distribution of a single-link RTT and observed that the average RTT was about 20 ms, but at least 40 ms was necessary to include around 95% of samples.

To conclude, note that, differently from SH algorithm, multiple paths joining source and destination could be built at the same time. In fact, this could happen because HLJF trivially reduces to LF under certain particular conditions.

6 Experimental Results

The three protocols detailed in Sects. 3, 4, 5, along with the test application described in Sect.2.2 have been implemented on Telos motes using TinyOS development tool. In order to evaluate the effectiveness of the three routing schemes, a linear network has been deployed inside our Lab. The WSN consists of 10 SNs and 1 MN, connected to a pc where data is stored.

The distance between each couple of adjacent nodes in the test-bed is homogeneous and in the order of about 30 cm. This value was chosen after performing several preliminary tests in order to empirically characterize the typical node radio coverage area when the node transmission power was set to the minimum available value, -25 dBm. The goal was to induce with good approximation a scenario where, as schematically depicted in Fig. 2, SNi is:

- (Almost) always able to directly communicate with SNi − 1 and SNi + 1
- Often able to directly communicate with SNi − 2 and SNi + 2
- Rarely able to directly communicate with SNi − 3 and SNi + 3
- Never able to directly communicate with farther nodes than SNi − 3 and SNi + 3

For each proposed routing algorithm, test sessions have run along multiple consecutive days, including both working and weekend days. In the case of HLJF algorithm,

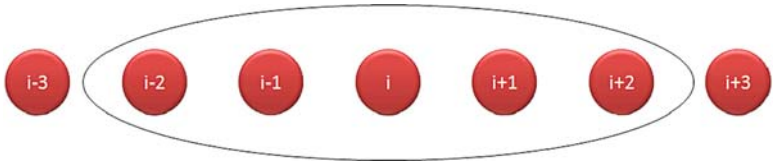


Fig. 2 Expected node radio coverage for the indoor WSN deployment

the Time-Out has been set equal to 50 ms. In the following section, we report and discuss a selected set of trials, concerning a "night" and a "day" scenarios, which have taken place during an ordinary working day. More precisely, the "day scenario" covers the 10-h span of time from 9 in the morning to 7 in the afternoon, i.e., that interval when most of the people are working at their desks and commercial radio devices as wireless networks are likely switched on. The "night scenario" is complementary to the previous one, lasts from 19 pm to 9 am of the day after thus having a duration of 14 h.

6.1 Packet Delivery Success Rate

Figure 3 shows the end-to-end information delivery success rate as a function of the target node address SNi, for the "night" and the "day" scenarios respectively.

The first observation regards a substantial difference of results obtained in the "night scenario" with respect to the "day scenario." This is likely due to a noisier channel available when people move around the test field and, above all, when other wireless technologies operating in the same 2.4 GHz ISM frequency band, such as IEEE 802.11 networks and Bluetooth, are active at the same time. In fact, other test sessions, not reported here, performed during weekend days in the absence of humans and with WiFi APs disabled, have not at all exhibited the remarkable performance heterogeneity from day to night.

Nonetheless, the performance degradation experienced by HLJF algorithm in the two scenarios is much more limited than in the case of both SH and LF schemes. This is achieved thanks to the (though single) packet retransmission procedure supported by the protocol in case of Time-Out expiration: it allows the success rate for the worse node, SN8, to be kept around 80%. The performance degradation suffered by SH is still more evident because, differently from LF, only one end-to-end path is admissible and, in addition, it always involves a larger number of node-to-node communications. The behavior of the SH protocol, mainly the deterministic end-to-end path, makes the interpretation of its results simpler and theoretically more easily predictable. In particular, an exponential decaying of delivery success rate as a function of the distance between MN and target SN is expected. Approximately, this is what really appears in Fig. 3. The LF algorithm exhibits a more irregular trend than

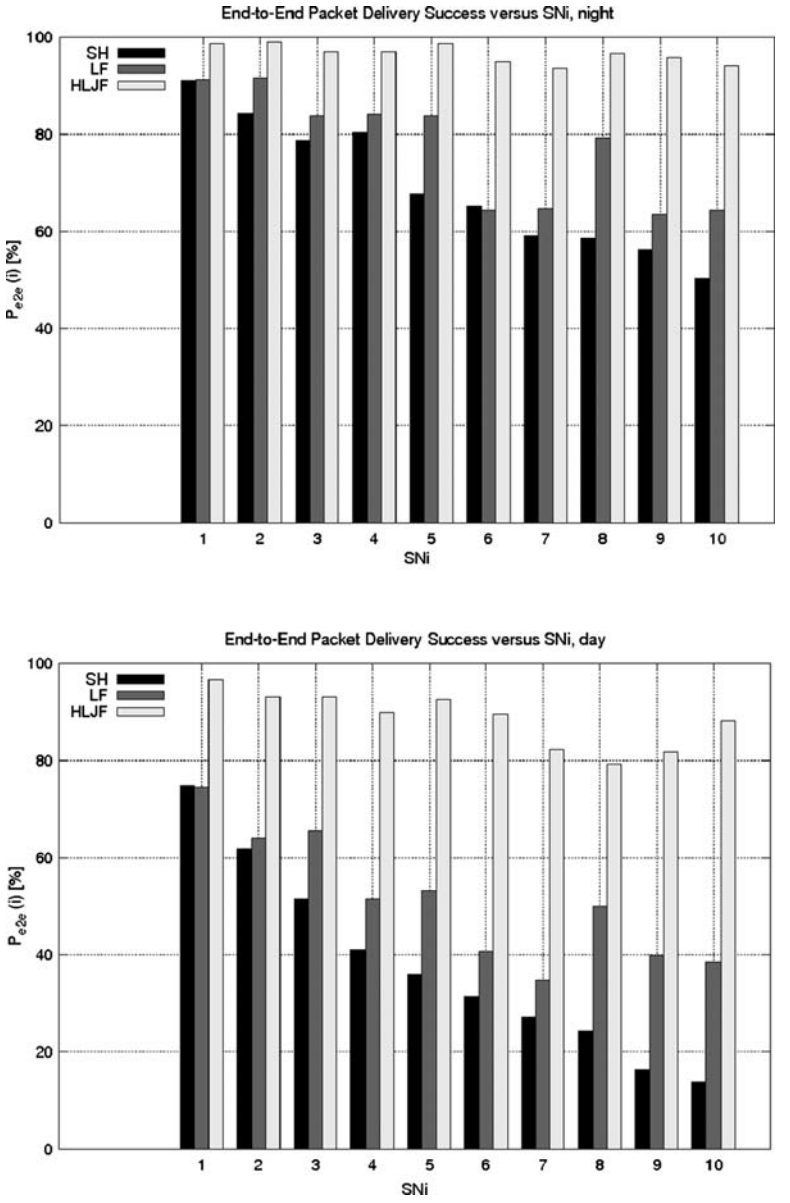


Fig. 3 Performance comparison of routing schemes from the viewpoint of the end-to-end communication success rate, according to "night scenario" (on the top) and "day scenario" (on the bottom)

the other two routing schemes in both the scenarios, and this is coherent with its functioning that leaves open several possible parallel paths and, as a consequence, many different destinies to an end-to-end communication packet. Definitively, the HLJF algorithm outperforms SH and LF in terms of communication reliability.

6.2 Delays

Average end-to-end delays are plotted in Fig. 4, which refers to the "day scenario." The analogous picture for the "night scenario" is not reported since they are qualitatively very similar. The only significant difference concerns the HLJF algorithm that shows a bit larger delays in the less favorable "day scenario." To provide further details about delays, Tables 1 and 2 respectively report minimum and maximum delays in the two scenarios.

The curve of SH average delays is perfectly linear as could be imagined considering how this protocol works. The slope of the curve is about 20 ms/node; thus,

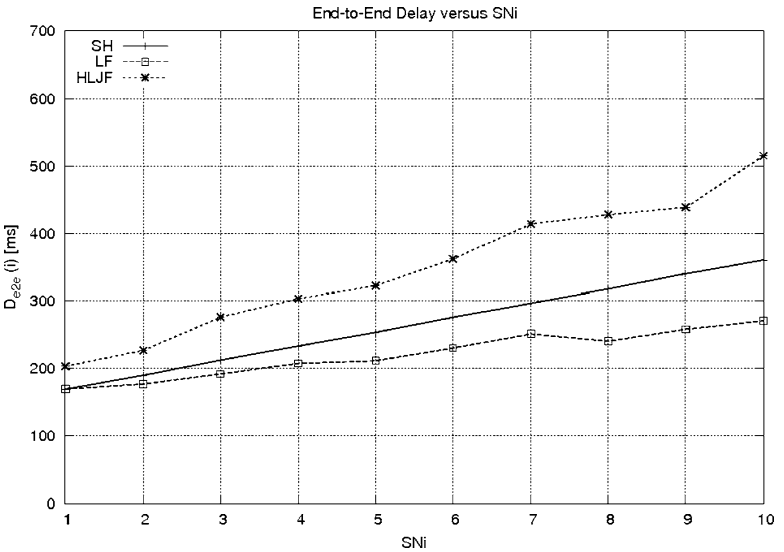


Fig. 4 Average end-to-end delays for "day scenario." Qualitatively similar results hold for the "night scenario"

Table 1 Minimum end-to-end delays, in ms

	SH		LF		HLJF	
	Day	Night	Day	Night	Day	Night
SN1	164.28	164.73	164.46	164.55	186.10	185.70
SN2	182.37	183.11	164.61	164.70	186.07	185.82
SN3	203.31	199.89	176.61	176.64	188.05	189.61
SN4	220.37	218.72	186.83	182.98	189.48	212.59
SN5	237.98	242.31	186.61	183.41	227.51	224.79
SN6	261.17	259.31	208.53	205.29	228.82	237.21
SN7	280.73	279.75	224.33	199.83	269.90	273.71
SN8	303.47	301.51	205.38	205.38	273.93	275.48
SN9	322.97	319.70	229.52	224.79	269.62	293.70
SN10	345.58	342.53	241.00	230.41	314.73	344.51

Table 2 Maximum end-to-end delays, in ms

	SH		LF		HLJF	
	Day	Night	Day	Night	Day	Night
SN1	186.98	178.62	188.11	188.57	338.68	289.92
SN2	202.67	205.51	201.93	195.98	411.35	333.01
SN3	229.86	228.67	240.05	219.06	422.06	452.24
SN4	262.63	252.11	231.69	226.87	669.34	525.48
SN5	287.02	271.42	245.42	238.83	590.85	463.62
SN6	297.67	295.75	272.86	262.05	653.23	551.97
SN7	320.37	316.86	298.43	284.12	686.86	577.39
SN8	347.93	345.83	292.82	275.09	786.41	555.15
SN9	359.34	369.75	311.61	307.46	702.39	701.63
SN10	376.07	409.52	306.82	312.81	970.46	751.46

the cost of each hop is in the order of 10 ms (recall that querying the node SNi according to SH algorithm requires $2 \cdot i$ hops). The conclusion is the same even when considering minimum or maximum delay samples.

Overall, the LF protocol is the one with the best performance in terms of delays: this can be explained by observing that this approach joins the twofold advantage of not managing acks and packet retransmissions while, at the same time often selecting the path with a number of hops as small as possible.

The performance of HLJF routing scheme is the worst on the average since the support for packet retransmissions after Time-Out expirations causes a large variance in experienced delays. Looking at Fig. 4, we can see that its trend in mean delays growth is higher when compared to SH, while the one of LF is the smallest. Coherently, the increase of HLJF maximum delays is the most emphasized too.

On the other hand, HLJF shows minimum delays that are smaller than the ones obtained with SH. This is not surprising because in case of no retransmissions (the actual source of delays), HLJF is able to reduce delays thanks to long jumps as LF.

Definitively, the LF algorithm is preferable in terms of end-to-end delays, but, depending on the application requirements, even a modified version of HLJF could be competitive and adapt to application constraints.

7 Conclusions

In this paper, we have proposed three different routing protocols designed for linear WSNs. Their performance in terms of end-to-end communication reliability and delays have been studied based on results coming from an indoor WSN composed on 11 nodes.

The performance evaluation of a larger experimental network, with up to 30 nodes, is the next step in order to investigate protocol scalability from an experimental perspective. In parallel, the development of some mathematical model would allow to reciprocally validate experimental and theoretical results.

Acknowledgments The research leading to the results presented in this paper has been carried out within the SAFESPOT Project [9], funded by the European Community's Sixth Framework Programme FP6.

References

1. Franceschinis M, Gioanola L, Messere M, Tomasi R, Spirito MA, Civera P (2009) Wireless sensor networks for intelligent transportation systems. 2009 IEEE 69th vehicular technology conference: VTC2009-Spring, Barcelona, Spain, 26–29 Apr 2009
2. Haridas H (2006) BriMon: design and implementation of railway bridge monitoring application. Master Thesis, Indian Institute of Technology, Kanpur
3. Barboni L, Valle M (2008) Wireless sensor networks power aware deployment. Sensor technologies and applications, 2008. SENSORCOMM '08. Second international conference on, 25–31 Aug 2008, pp 252–257
4. Hellman K, Colagrosso M (2006) Investigating a wireless sensor network optimal lifetime solution for linear topologies. J Interconn Netw 7:91–99
5. Gibson J, Xie GG, Xiao Y. (2007) Performance limits of fair-access in sensor networks with linear and selected grid topologies. Global telecommunications conference, 2007. IEEE GLOBECOM '07, 26–30 Nov 2007, pp 688–693
6. Karveli T, Voulgaris K, Ghavami M, Aghvami AH (2008) A collision-free scheduling scheme for sensor networks arranged in linear topologies and using directional antennas. Sensor technologies and applications, 2008. SENSORCOMM '08. Second international conference on, 25–31 Aug 2008, pp 18–22
7. Wang H, Yang Y, Ma M, Wu D (2007) Network lifetime optimization by duality approach for single-source and single-sink topology in wireless sensor networks. Wireless and optical communications networks, 2007. WOCN '07. IFIP international conference on, 2–4 July 2007, pp 1–7
8. Koutsakis P, Papadakis H (2006) Efficient medium access control for wireless sensor networks. Wireless pervasive computing, 2006 1st international symposium on, 16–18 Jan 2006
9. SAFESPOT Integrated Project. http://www.safespot-eu.org/
10. Maroti M (2004) Directed flood-routing framework for wireless sensor networks. Proceedings of the 5th ACM/IFIP/USENIX international conference on middleware, Toronto, Canada, 2004, pp 99–114

A Distributed Procedure for IEEE 802.15.4 PAN Coordinator Election in Emergency Scenarios

Emanuele Cipollone, Francesca Cuomo, and Anna Abbagnale

1 Introduction

Communication networks play a fundamental role in the response to massive catastrophes, such as earthquakes, floods, fires, and so on. When one of these events happens in an urban area, public authorities are expected to undertake all actions that are necessary to control and to limit damages for people and for buildings. A central authority, usually, takes the role of coordinating all operations in the emergency scenario and, to this aim, it needs to know in real-time the exact situation in the place where the disaster has happened and to communicate with all the teams deployed within the catastrophe area.

In such a scenario, the presence of a reliable communication infrastructure is fundamental, in order to allow communications, especially among emergency agents (like policemen, firemen, doctors, etc.) and to send and to receive information to/from a remote center, responsible for the emergency management. This communication infrastructure will be composed by a set of different network technologies, devoted on one side to the transmission, also at long distances, of information (e.g., third generation cellular networks or WiMax networks) and, on the other side, to the collection of data within the disaster area and to the communication, in the same area, among emergency agents [e.g., wireless personal area networks (WPANs)].

In this chapter, we focus on network technologies that allow, in an emergency scenario, the communication among people and the collection of specific data of interest (like temperature and humidity degree in the event of fire). Such technologies should be able to self-configure quickly and to guarantee a lifetime sufficient for an efficient emergency management. An IEEE 802.15.4 WPAN is a good candidate to take this role, thanks to its pervasive nature [1]. In this network, a specific node (called *PAN coordinator*) takes the control of the network and its position has a significant impact on the performance [2–4].

E. Cipollone (✉), F. Cuomo, and A. Abbagnale
Department of INFOCOM, University of Rome "Sapienza," via Eudossiana 18, 00184 Rome, Italy
e-mail: cipollone@infocom.uniroma1.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_4,

We present a study on the self-configuration of IEEE 802.15.4 WPANs in emergency scenarios, with specific attention to procedures of election of the PAN coordinator. As stated in [5], the development of self-managing, self-configuring and self-regulating networks, and communication infrastructures is an area of considerable research and industrial interest. In an application context, as the one relevant to the management of catastrophes, the use of autonomic and self-configuring techniques for controlling the selection of the PAN coordinator opens new prospectives.

We propose a distributed procedure, that works in an autonomic way, to elect the best node to perform the PAN coordination. This heuristic procedure aims at moving, whatever node starts the network formation in accordance to the standard IEEE 802.15.4, the PAN coordinator role to a target position which guarantees the minimum *network depth* (its meaning is explained in Sect. 2). Specific consequences of selecting the PAN coordinator in accordance to our procedure are energy saving during data delivery, thus increasing the network lifetime in case of battery supplied devices [e.g., in case of wireless sensor networks (WSNs)] and delivery delay reduction. On the contrary, if any other node is selected as PAN coordinator, network performance worsen. In self-configuring networks, our distributed procedure can be used after the network has just formed to reorganize the topology selecting in a suitable way a new PAN coordinator.

The chapter is structured as follows: Sect. 2 recalls the main characteristics of the IEEE 802.15.4 topology formation. In Sect. 3, we describe the reference architecture scenario. In Sect. 4, we present our distributed procedure for PAN coordinator election and we evaluate it in Sect. 5. Finally, the overall conclusions of the paper are provided in Sect. 6.

2 Self-Configuration of an IEEE 802.15.4 WPAN

An IEEE 802.15.4 WPAN is composed of one PAN coordinator and a set of nodes [6]. A typical network topology defined in the standard is the so-called cluster tree, where nodes associated to a single PAN coordinator are arranged in a tree with parent–child relationships. In an IEEE 802.15.4 network, it is possible to have Full Function Devices (FFDs) that allow the association of other nodes to the network, and Reduced Function Devices (RFDs) that do not permit the association of other nodes. The PAN coordinator is always a FFD, intermediate nodes allowing data relay (router) are FFDs too, whereas RFDs are leaves of the tree.

The standard defines a set of procedures implemented by the PAN coordinator to initiate a new WPAN and by other nodes to join an existing WPAN. The PAN coordinator starts by selecting a suitable channel. This selection is performed by the Energy Detection (ED) scan which measures the interference (i.e., the *peak energy*) of each available channel (16 channels in the 2.4 GHz ISM band). The procedure adopted by nodes to join a WPAN is named *association procedure* and it establishes relationships between nodes within a WPAN. The operations performed by a node

to join a WPAN are (1) the node searches for the available WPANs, (2) it selects a coordinator[1] belonging to the available WPANs and (3) it starts a message exchange with the selected coordinator to associate with it.

The discovery of available WPANs is performed by scanning the *beacon frames* broadcasted by the coordinators.

The level l of a node in the tree is intended as the distance (in terms of number of hops) of the node from the PAN coordinator. We define *network depth* (or *tree depth*) L as the maximum distance of a node from the PAN coordinator within the tree, i.e., the maximum value of l. We indicate with L_i the tree depth when the node i is the PAN coordinator. The variable g is instead the mean level of a node in the tree (g_i is the mean level of a node in the tree when node i is the PAN coordinator). These values are affected by the position of the PAN coordinator.

3 Considered Network Architecture for Emergency Scenarios

As explained in Sect. 1, the management of an emergency event requires the use of a communication infrastructure in the disaster area. In this context, thanks to the pervasive nature of their devices, the use of IEEE 802.15.4 WPANs seems particularly appropriate for the following reasons:

1. Emergency agents (like policemen, firemen, doctors, etc.) require a communication infrastructure able to self-configure in a fast manner and to guarantee a lifetime sufficient for an efficient emergency management.
2. In the emergency area, the collection of specific data of interest (like temperature, humidity, movements, etc.) could be very important for the emergency management.

In Fig. 1, an example of the use of IEEE 802.15.4 WPANs in an emergency area is reported. IEEE 802.15.4 FFDs and RFDs can be interconnected to form several WSNs. Each of them has a limited coverage area (e.g., some dozen of meters) and it is responsible for data collection in the area where it is deployed. Other IEEE 802.15.4 FFDs, instead, are interconnected to form one or more WPANs. These networks have a coverage area greater than WSNs, and they allow communications among emergency agents. Moreover, they are responsible for the collection of data revealed by WSNs and, for this reason, some nodes of WPANs have to be connected with RFDs. Each WPAN has its own PAN coordinator, instead a WSN can have or not its own PAN coordinator. In the latter case, the role of PAN coordinator of the WSN is taken by an FFD of the WPAN directly connected to the same WSN.

With reference to Fig. 1, the procedure for PAN coordinator election that we propose in this paper has an impact on both WPANs and WSNs (in terms of network

[1] Coordinators are all nodes that can act as relay nodes.

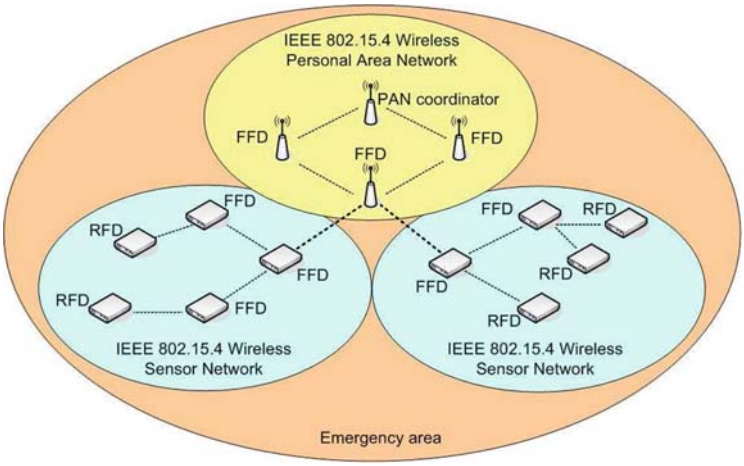


Fig. 1 A possible (pervasive) use of IEEE 802.15.4 WPANs in an emergency area

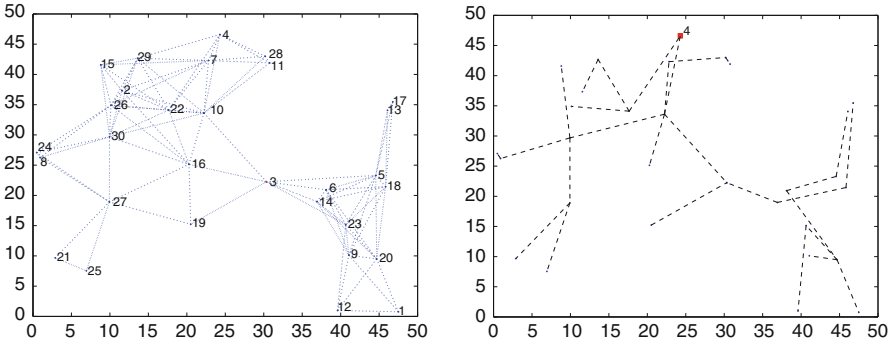
topology and performance). The selection of the node having the role of PAN coordinator affects the following topological characteristics:

- The structure of the parent–child relationships
- The number of nodes at different levels of the tree
- The tree depth

The optimization performed in the cluster-tree topology, reflects to the routing delay and energy consumption when a hierarchical routing protocol (see paper [7]) is applied. In this case, data generated by nodes and directed to the PAN coordinator are routed upward to the root tree along the parent–child relationships, i.e., every node relays data to its parent. Therefore, the energy consumption and the delivery delay due to data transmission are proportional to the number of hops of the path from the source node to the destination node (i.e., PAN coordinator).

4 A Distributed Procedure for PAN Coordinator Election

We propose a distributed procedure (named PANEL – PAN coordinator ELection) aiming at moving the PAN coordinator role to a node that has a specific position in the formed network. We consider N nodes, having transmission range R, randomly deployed in a given area (as in the example of Fig. 2a). A generic node starts the WPAN in accordance to the IEEE 802.15.4 standard and form a network. The starting point is then a network topology formed by means of the IEEE 802.15.4 association procedure. In Fig. 2b, the network resulting from a formation started by node 4 is shown. This network topology is a tree having as root node a PAN coordinator p^{IN}, that initially is in general nonoptimal, and presenting a tree depth L_{IN}. PANEL is iterative algorithm. The idea behind PANEL is the following: in order to



(a) Nodes and nodes' visibilities.

(b) Network formed by running the IEEE 802.15.4, node 4 is the PAN coordinator.

Fig. 2 An example of network including 30 nodes. The numbers represent nodes' identifiers

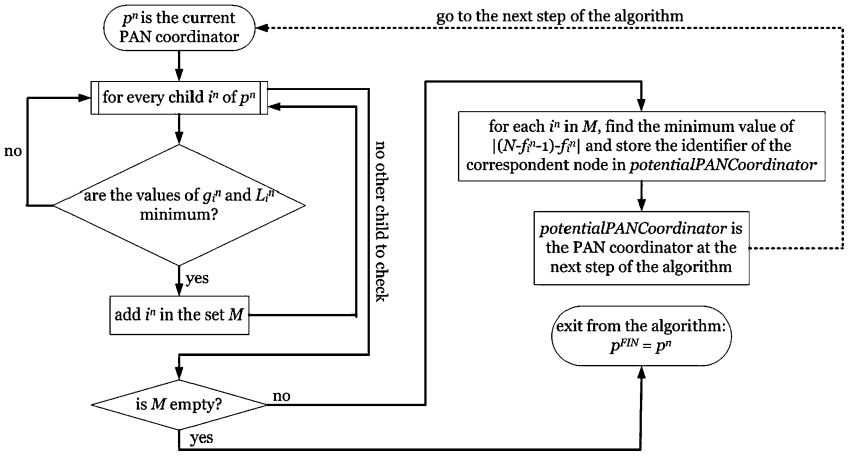


Fig. 3 Simplified flow chart describing PANEL

improve performance of the topology resulting from the IEEE 802.15.4 (in terms of energy consumption and data delivery delay), the node that should be elected as PAN coordinator is the one that allows:

1. To decrease the mean level of nodes within the tree
2. To decrease the tree depth
3. To guarantee the best *network balancing*, among all nodes that satisfy the previous two conditions

We point out that PANEL results in the best choice of the PAN coordinator given a specific starting topology and, thus, the output of the procedure depends on the topology given as input.

The n-th iteration of PANEL implements the steps of the flow chart in Fig. 3. Let p^n be the PAN coordinator at the beginning of iteration n. Every child i^n of

p^n ($i^n = k_{1^n}, k_{2^n}, ..., k_{m^n}$, where m^n is the number of children of p^n) sends to it three data: the mean level of a node in the tree g_{i^n} and the tree depth L_{i^n} if i^n was elected PAN coordinator, and the number of its descendants f_{i^n}, that is the number of nodes of the subtree having it as root. Nodes get this information from data structures (i.e., *vectors*) built during the association procedure. In fact, by exploiting messages exchanged in this phase, every node i^n computes a vector V_{i^n} whose jth element $V_{i^n}[j]$ indicates the number of nodes that are j hops away from the node i^n in the subtree having i^n as root. Let V_{p^n} the vector of p^n (the PAN coordinator at the beginning of iteration n) and V_{i^n} the vector of the generic child i^n of p^n; p^n sends V_{p^n} to all its children, so the generic child i^n computes the operations described in Pseudo-code 1, in order to calculate g_{i^n}, L_{i^n} and f_{i^n}.

Pseudo-code 1 COMPUTATION OF g_{i^n}, L_{i^n} AND f_{i^n}

1: $V^*_{i^n} = V_{i^n}$;
2: add in queue to $V^*_{i^n}$ a number of *zero* elements equal to: *size* (V_{p^n}) - *size* (V_{i^n}) + *1*;
3: $V^{**}_{i^n} = V^*_{i^n}$;
4: shift right one position the elements of $V^{**}_{i^n}$;
5: add in queue to V_{p^n} one element equal to 0;
6: $V^{**}_{i^n} = V_{p^n} - V^{**}_{i^n}$;
7: $V^{**}_{i^n}[1] = V^{**}_{i^n}[1] - 1$;
8: shift right one position the elements of $V^{**}_{i^n}$;
9: $V^{**}_{i^n}[1] = V^{**}_{i^n}[1] + 1$;
10: $V^{**}_{i^n} = V^{**}_{i^n} + V^*_{i^n}$;
11: $g_{i^n} = 0$;
12: $sumElements = 0$;
13: **for** $j = 1, 2, ..., size(V^{**}_{i^n})$ **do**
14: $g_{i^n} = g_{i^n} + j \cdot V^{**}_{i^n}[j]$;
15: $sumElements = sumElements + V^{**}_{i^n}[j]$;
16: **end for**
17: $g_{i^n} = g_{i^n}/sumElements$;
18: remove from $V^{**}_{i^n}$ potential zero element in queue;
19: $L_{i^n} = size(V^{**}_{i^n})$;
20: $f_{i^n} = 0$;
21: **for** $k = 1, 2, ..., size(V_{i^n})$ **do**
22: $f_{i^n} = f_{i^n} + V_{i^n}[k]$;
23: **end for**

The vector $V^{**}_{i^n}$ as obtained at line 10 of Pseudo-code 1 indicates the number of nodes at different levels of the tree if i^n was elected PAN coordinator. In Fig. 4, it is shown as an example of computation of this vector made by the generic child i^n of a PAN coordinator p^n. After this computation, the node i^n calculates g_{i^n}, L_{i^n} and f_{i^n}, in accordance with the steps from line 11 to line 23 of Pseudo-code 1. In case of Fig. 4, at the end of Pseudo-code 1, i^n will obtain: $g_{i^n} = 1.778$, $L_{i^n} = 3$ and $f_{i^n} = 5$. We point out that the computational and storage resources requested to nodes are very low and thus in step with the constrained capacities of IEEE 802.15.4 nodes.

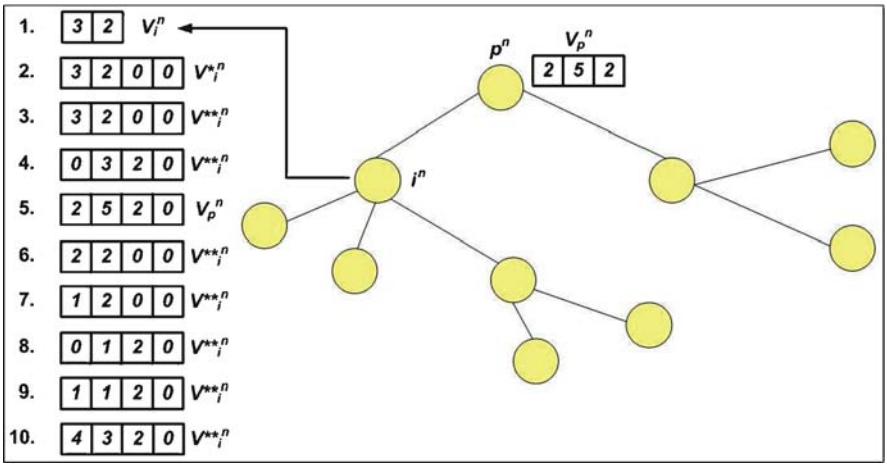


Fig. 4 Operations computed by a generic child i^n of the PAN coordinator p^n at the beginning of iteration n of the procedure PANEL. The numbers near the vectors indicate the corresponding steps of Pseudo-code 1

After receiving these data from all its children, p^n computes the best (i.e., the minimum) values of g and L (indicated with *bestMeanLevel* and *bestTreeDepth*, respectively), taking into account also its own values. Then, it creates a set M, where it puts, among all its children, the generic node i^n iff:

1. g_{i^n} is equal to *bestMeanLevel*
2. L_{i^n} is equal to *bestTreeDepth*

These conditions guarantee that if a new potential PAN coordinator is effectively chosen among the children of p^n, it will be characterized by the optimal (i.e., the minimum) values of g and L. After performing this phase, p^n checks if the set M is empty or not. If M is empty, this means that there are no nodes, among the children of p^n, able to improve network performance if elected as PAN coordinator, so the procedure stops and p^n is effectively elected PAN coordinator of the network ($p^{\text{FIN}} = p^n$). In the other case, instead, among all nodes in M, p^n selects as new potential PAN coordinator (indicated as *potentialPANCoordinator*) the node that guarantees the best network balancing; we define the best network balancing as the minimum difference between the number of the descendants (f_{i^n}) of a node i^n and the number of the other nodes ($N - f_{i^n} - 1$) in the tree. After this choice *potentialPANCoordinator* becomes the PAN coordinator at the beginning of the iteration $n + 1$ and the procedure goes on.

The conditions present in PANEL are justified by the fact that our goal is to move (downward the parent–child relationships) the PAN coordinator toward a part of the network characterized by a greater number of nodes and a high mean level of the same nodes. The movement will reduce the mean level of nodes and the tree depth, once the new PAN coordinator is elected, and the new topology will result in

a tree having all branches more balanced, in terms of number of nodes and depth. As example, in the case of Fig. 2b, PANEL moves the PAN coordinator from node 4 to node 3.

5 Performance Analysis

A performance analysis of the proposed procedure has been carried out by testing it on different network topologies. Our aim is to understand how nodes distribute at different levels of the tree when PANEL is applied on already formed IEEE 802.15.4 networks with $N = 50$ and 100. For each value of N, we simulate N times the formation of the network by using NS-2 and by varying each time the position of nodes in the considered area. For each topology, we randomly choose the PAN coordinator and collect the distribution of nodes at different levels. Then, we apply PANEL on formed networks and we collect the resulting distribution of nodes at different levels. Figure 5a, b show the obtained results for $N = 50$ and 100,

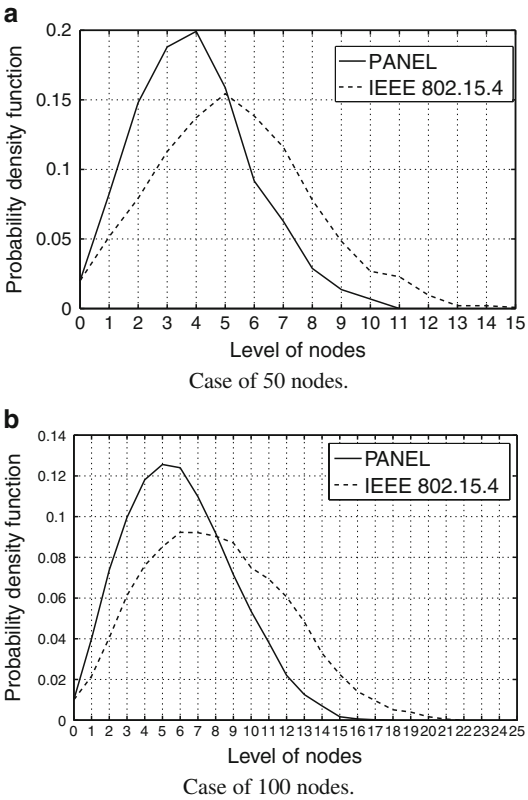


Fig. 5 Probability density function of nodes' level

respectively. In both cases, PANEL obtains improvements in the topology. In fact the curve that represents the probability density function of nodes' level related to topologies obtained with PANEL becomes higher and stamps out if compared with the one related to topologies obtained with the IEEE 802.15.4 association procedure. This means that by applying PANEL, the probability to find nodes at levels close to the PAN coordinator becomes higher when compared with the case of IEEE 802.15.4 networks. Therefore, there is an improvement of network performance, since in average shorter paths arise from sensor nodes to the PAN coordinator compared with the case of IEEE 802.15.4 networks: this results in energy saving and low data delivery delay within the network.

The improvement of network performance is confirmed from Figs. 6 and 7. They show, respectively, the average gain in terms of tree depth and mean level of nodes as function of the number of nodes in the topology, for the same networks previously analyzed. The gain of Fig. 6 is computed as the difference between the initial tree depth of topologies formed by using NS-2 (i.e., formed in accordance with the IEEE 802.15.4 procedure), and the tree depth achieved by PANEL. In the same way, Fig. 7 shows the gain in the mean level of nodes. Basically, the average value of this gain, in both cases, is always positive, thus confirming the goodness of PANEL and it

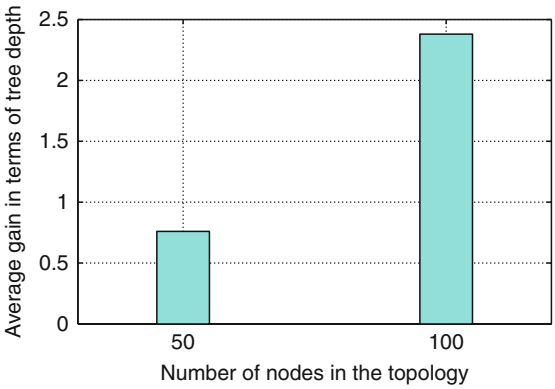


Fig. 6 Average gain in terms of tree depth as function of the number of nodes in the topology

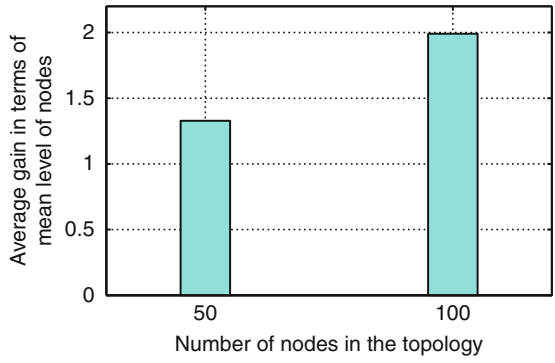


Fig. 7 Average gain in terms of mean level of nodes as function of the number of nodes in the topology

increases when the number of nodes increases. Therefore, PANEL reconfigures the network resulting in lower tree depth and mean level of nodes compared with IEEE 802.15.4, with a consequent improvement of network performance.

6 Conclusions

In this chapter, we presented a distributed solution, called PANEL, for PAN coordinator election in IEEE 802.15.4 WPANs. It can be efficiently applied in emergency scenarios, when it is required that this kind of network is able to self-configure and to guarantee a high lifetime and low data transfer delays. We showed that if a specific node assumes the role of PAN coordinator, network performance has a significant improvement.

PANEL works well in whatever network configuration, requires the exchange of simple control information and is compliant with the standard IEEE 802.15.4. For these reasons, it is suitable to be applied in emergency scenarios.

Acknowledgments This work has been partially supported by the IT-funded FIRB/PNR INSYEME (protocol number: RBIP063BPH).

A special thanks goes to Matteo Antonetti, for the support in the simulations.

References

1. Zheng J and Lee M (2004) Low rate wireless personal area networks for public security. In: IEEE 60th vehicular technology conference. VTC2004-Fall, vol 6. September 2004, pp 4568–4572
2. Abbagnale A, Cipollone E, Cuomo F (2008) Constraining the network topology in IEEE 802.15.4. In: Annual Mediterranean ad hoc networking workshop, MED-HOC-NET '08, 23–27 June 2008
3. Liang Q (2003) Designing power aware self-reconfiguring topology for mobile wireless personal area networks using fuzzy logic. IEEE Trans Syst Man Cybernet C: Appl Rev 33(3):390–394
4. Jung S, Chang A, Gerla M (2007) Comparisons of zigbee personal area network (pan) interconnection methods. In: 4th international symposium on wireless communication systems, ISWCS 2007, October 2007, pp 337–341
5. Dobson S, Denazis S, Fernández A, Gaïti D, Gelenbe E, Massacci F, Nixon P, Saffre F, Schmidt N, Zambonelli F (2006) A survey of autonomic communications. ACM Trans Auton Adapt Syst 1(2):223–259
6. Part 15.4: Wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (WPANs), IEEE Std. 802.15.4, 2006
7. Cuomo F, Luna SD, Monaco U, Melodia T (2007) Routing in ZigBee: benefits from exploiting the IEEE 802.15.4 association tree. In: IEEE international conference on communications 2007, IEEE ICC '07, June 2007, pp 3271–3276

A Mobile Platform for Measurements in Dynamic Topology Wireless Networks

Emanuele Scuderi, Rocco Emilio Parrinello, David Izal, Gian Paolo Perrucci, Frank H. P. Fitzek, Sergio Palazzo, and Antonella Molinaro

1 Introduction and Motivation

In the last years, cooperative wireless networks have quickly gained the interest of many researchers in the scientific community. Several studies have shown how performance of a wireless network can be improved by using principles of cooperation [2, 7]. A novel architecture, namely cellular controlled peer-to-peer (CCP2P), has been introduced in [3]. According to it, cellular devices can create cooperative clusters with neighboring devices in their proximity using a short range technology. Each terminal is then contributing to the cooperative cluster by sharing its cellular link. The grouped members acting in a cooperative manner can achieve better performance than a stand-alone device, in different scenarios. Performance can be improved in terms of data rate, as described in [9], where authors show the improvements achieved for mobile web browsing applications. In other cases, performance can be improved in terms of energy consumption, as shown in [6, 8]. In these papers, authors show that by using the CCP2P architecture for streaming and file downloading, the energy consumption can be reduced and the quality of the service can even be increased. For other examples of performance improvements, we refer the interested reader to [4].

Authors in [10] have carried out some energy and throughput measurements on state-of-the-art mobile phones, which cooperate over IEEE 802.11b/g short-range links. During the measurement phase, the mobile phones were in a fixed position. This is in contrast with what usually happens in a real world scenario, where the topology of the network is rapidly changing due to the user mobility. Therefore, with such kind of measurements, it is not possible to quantify the effect that the changes in the topology would cause. To overcome this problem, some software can be used

E. Scuderi (✉), R.E. Parrinello, D. Izal, G.P. Perrucci, and F.H.P. Fitzek
Department of Electronic Systems, Aalborg University, Denmark
e-mail: escuderi@es.aau.dk

S. Palazzo
Università di Catania, Catania, Italy

A. Molinaro
Università "Mediterranea" di Reggio Calabria, Reggio Calabria, Italy

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_5,

to simulate cooperative wireless networks, namely Netlogo [1]. However, simulations usually give only an approximation of real world scenarios. Another approach could be to make experimental measurements in a real environment. However, this is not trivial to implement, especially if limited resources are available.

To trade off the need of a realistic performance evaluation and the costs of on-the-field measurements in this chapter, we design and test a simple and cheap platform, which can be used to perform measurements in mobility. Recently, the term *Internet of Things* has gained popularity in the scientific community. It refers to a number of technologies and research disciplines that enable the Internet to reach physical objects. This is another research area where it could be important to have a tool like the one we designed, which helps to make experimental measurements of a dynamic environment.

This chapter is organized as follows. In Sect. 2, a detailed overview of the designed platform is given. Sections 3 and 4 introduce the possible communication architectures for the platform and some examples of applications, respectively. Finally, Sect. 5 presents our conclusions.

2 The Proposed Mobile Platform

In this chapter, we introduce a mobile platform designed to optimize measurements for mobile wireless networks with dynamic topology. The platform, in the following referred as the *robot*, is shown in Fig. 1. The robot consists of two main blocks, namely the *mechanical block* (used for the motion) and the *controller block* (used for controlling the engines for motion).

The *mechanical block*, shown in Fig. 2, is composed of two engines, two wheels, a battery pack, and a metal platform to host a mobile phone.

The core of the *controller block* is the *Opensensor* (see Fig. 3), a circuitry board developed by Mobile Devices Group at Aalborg University in Denmark [5]. It is a highly integrated device that supports different kinds of data communication interfaces such as Bluetooth, RS232 (or USB), and an nRF905 antenna. By using these different interfaces, the *Opensensor* can receive commands for the motion and transmit them to the engines. This makes it possible for the robot to be remotely controlled by:

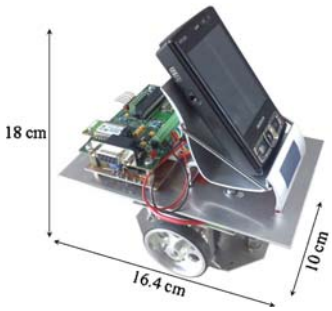
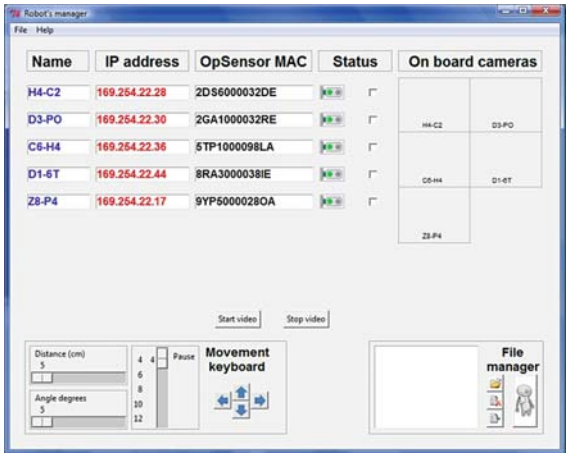


Fig. 1 Picture of the *robot*, the mobile platform designed for carrying out measurements for mobile wireless networks with dynamic topology

Fig. 2 Mechanical block: it is composed out of two engines, two wheels, a battery pack and a metal platform to host the mobile phone

Fig. 3 Opensensor, the core of the controller block

Fig. 4 Graphical user interface of the application developed for controlling the robot using a PC



- PCs (using either Bluetooth or a cable)
- Bluetooth-enabled devices (including mobile phones or Opensensor)
- Opensensors (using the nRF905 antenna)

To easily control multiple robots during measurements, an application for PCs has been developed. A daemon program is running in the background and waiting for robots to connect. Upon a successful connection, the information regarding the connected robot pops up on the screen. By using the Graphical User Interface, shown in Fig. 4, it is possible to select one or more robots to control. Each of them

can receive a single movement command or a script with a list of movement commands for automatizing measurements.

All the information for building, programming, and using the robot are available for the scientific community on the project webpage [11].

3 Communication Architectures

As mentioned in the previous section, the high flexibility of the platform allows the robots to be remotely controlled by using several technologies. In this chapter, we give a short description of four possible communication architectures for controlling the robot. Each of them is characterized by different communication links and devices involved, as shown in Fig. 5.

Architecture 1. In this architecture, a Bluetooth connection is established between the Opensensor board and the phone sitting on the robot. Another phone can be used to remotely control the robot by establishing a Wi-Fi ad-hoc connection with the phone (that will forward the commands back to the Opensensor). Finally, the Opensensor sends the right pulses to the engines for the motion. If many robots need to be controlled, the messages are broadcasted to all the phones sitting on the robots.

Architecture 2. This architecture is very similar to the previous one, but a PC is used as a controller instead of a phone. The Graphical User Interface on the PC makes it very easy to control one or more robots at the same time.

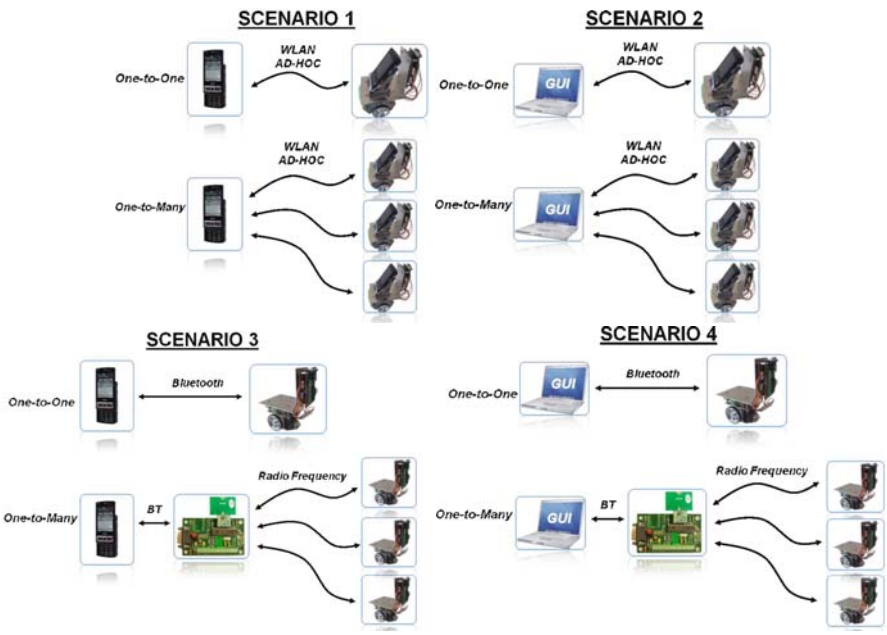


Fig. 5 Architectures overview

Architecture 3. For this architecture, in the one-to-one topology, the mobile phone controlling the robot is connected directly to the Opensensor using Bluetooth. The main limitations of this architecture are two. Firstly, the number of devices in a piconet is limited to 8 by the Bluetooth standard. Secondly, the communication range is limited to 15–20 m. For these reasons, to implement the one-to-many topology, an extra Opensensor is connected to the phone and the commands are sent to the robot using the nRF905 antenna, as shown in Fig. 5.

Architecture 4. This architecture is very similar to the previous one, with the difference that a PC is used instead of a mobile phone to act as the controller.

4 Examples of Applications

In this section, we present some usage examples of the robot. There are two main categories of usage depending on the scenario:

Passive transport. In this scenario, the robot carries a phone around as a passenger, meaning that no communication link is established between the motion part of the robot and the phone. The phone is free to take measurements in the environment with its internal sensors. The robot is controlled using a second phone, as in the communication architectures 3 and 4 described in Sect. 3. An example for this scenario is to use the mobile phone carried on the robot for making energy or throughput measurements over the IEEE802.11b/g connection with other mobile phones. In this case, it is very important that the phone does not have any other open connection to avoid interference with the measurements. The movement commands can be preset in a script file sent to the robot before starting the measurements.

Active transport. In this scenario, the phone sitting on the robot has a connection open with it. This allows the phone to control directly the robot and therefore being in charge of the mobility. This can be done using the communication architectures 1 and 2 described in Sect. 3. This could be useful in scenarios where the robot has to track and to chase moving objects or if the phone collects data using its internal sensors (GPS, light sensor, microphone, camera), and it has to move accordingly to some occurring events.

4.1 Passive Transport Scenario

In order to test the robot in a *passive transport scenario*, we have made some throughput measurements of the IEEE802.11b/g connection between the controlling phone and the transported one. It is important to note that the main goal of this chapter is to present this new tool for measurements in wireless networks, rather than give accurate measurements results.

4.1.1 Example 1: Direct Transmission

The location for this measurement is a corridor, approximately 30-m long. To carry out this measurement campaign two Nokia N95 8GB mobile phones have been used, one as the sender and the other one as the receiver. The sender is standing at one end of the corridor, whereas the receiver is placed on the robot. The robot is moving in a straight line from the sender to the opposite end of the corridor with low speed (see Fig. 6). The sender is continuously sending UDP packets of 1 KB to the receiver over a Wi-FI connection. The receiver sitting on the robot measures the throughput. In Fig. 7, the throughput of the transmission is shown and, as expected, the throughput decreases along with the distance between sender and receiver.

4.1.2 Example 2: Transmission Relay

In the second example, the sender and the receiver are not in the transmission range. They are located one at the beginning and one at the end of a 30 m L-shaped corridor. The setup of the experiment is shown in Fig. 8. In this setup, the direct transmission is not possible because the sender and the receiver are not in the range of connection. Therefore, a relay becomes strictly necessary to assure a transmission from the sender to the receiver. We use two phones as sender and receiver, and a third one, sitting on a robot, as relay. After setting up a Wi-Fi ad-hoc network among the three phones, the sender starts the transmission. The relay receives the packets, computes the throughput, and forwards them to the receiver, which computes the

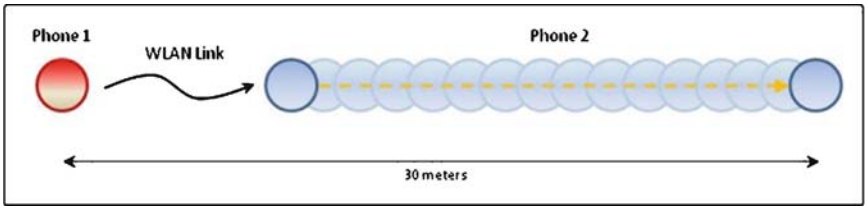


Fig. 6 Sketch of the measurements setup for the direct transmission

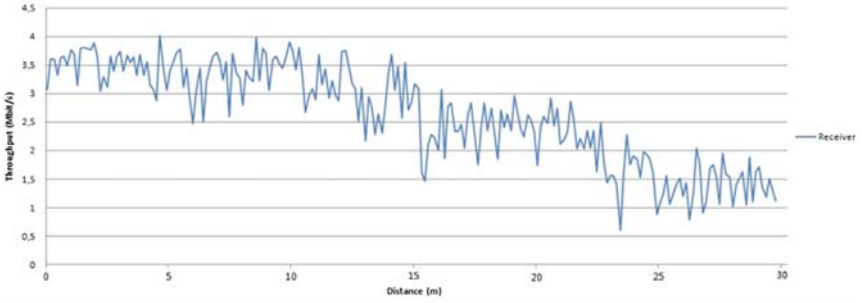


Fig. 7 Throughput measured on the receiver side

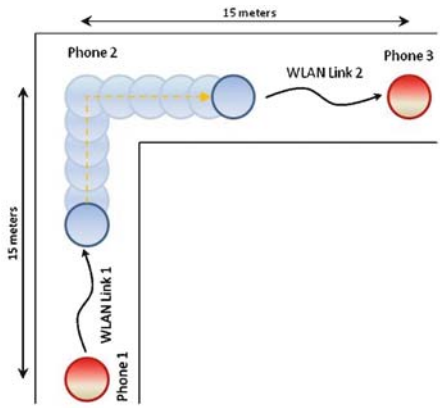


Fig. 8 Setup for the measurements in the transmission relay scenario

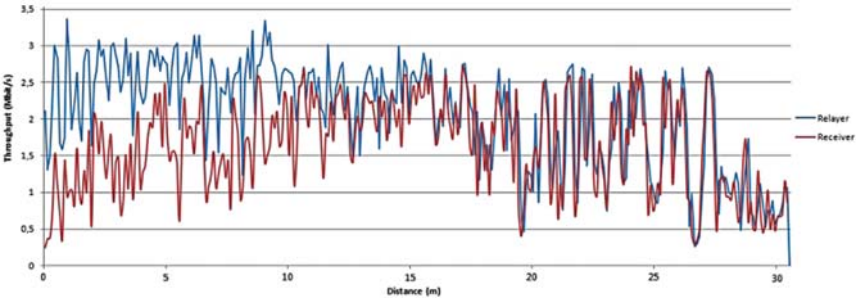


Fig. 9 Throughput at the relay and the receiver side for the relay transmission setup

throughput as well. While the transmission is taking place, the robot is moving with low speed from the sender toward the receiver. As we can notice from the plot in Fig. 9, the throughput of the receiver has a parabolic trajectory. The maximum value of the throughput is around 2.6 Mbit/s and is given at 15 m, when the relay reaches the corner of the corridor.

4.2 Active Transport Scenario

In the following, we present an application developed to give a simple example for the *active transport scenario*. The goal of this application is to remotely control a robot used as a video stream server. The phone placed on the robot is streaming the video from its camera and broadcasting it by using IEEE802.11b/g.

4.2.1 Remote Camera Application

The purpose of this application is to show how the robot can be be used according to the architecture 1 described in Sect. 3. In this example, the phone sitting on the

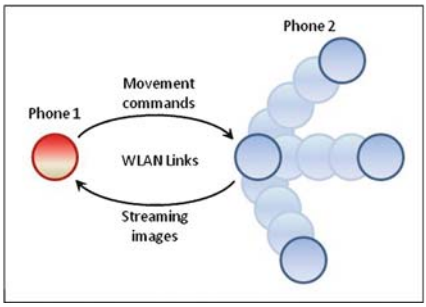


Fig. 10 Setup description for the remote camera application

robot uses the camera to collect images and streams them to another phone using a Wi-Fi connection. As shown in Fig. 10, the *Phone 2* is connected to the robot and streaming the images, whereas the *Phone 1* is receiving the images and it is able to send movement commands to the *Phone 2* that will forward them to the robot. A possible applicative scenario where this application can be used is an advanced video surveillance. For example, if a user wants to monitor some areas, he/she can control the robot over the Internet (with the help of an Access Point) by sending movement commands to the *Phone 2*, which will stream back the images.

5 Conclusions

In this chapter, we have presented a mobile platform designed to optimize measurements for mobile wireless networks, especially those with dynamic topology. This platform has wheels to allow movements and can be controlled by several kinds of devices (e.g., mobile phones, PC, Opensensors) using different wireless technologies (Bluetooth, Wi-Fi, RF antennas). The main purpose of this work is to offer to the scientific community a new interesting, very flexible and freely available framework for performance evaluation of wireless networks and devices. Therefore, all the information for the hardware and the software can be found on the webpage of the project [11].

References

1. Albiero F (2006) Wireless-coop-mobile, netlogo community models. http://ccl.northwestern.edu/netlogo/models/community/
2. Fitzek FHP and Katz M (eds) (2006) Cooperation in wireless networks: principles and applications – real egoistic behavior is to cooperate! ISBN 1-4020-4710-X. Springer, Heidelberg
3. Fitzek FHP, Katz M, Zhang Q (2006) Cellular controlled short-range communication for cooperative p2p networking. In: *Wireless World Research Forum (WWRF)*
4. Fitzek F, Perrucci G, Petersen M (2008) *Heterogeneous wireless access networks: architectures and protocols*. Springer, Heidelberg

5. Grauballe A, Perrucci GP, Fitzek FHP (2008) Opensensor – an open wireless sensor platform. In: 4th international mobile multimedia communications conference(MobiMedia 2008), Oulu, Finland, July 2008. ICTS/ACM
6. Militano L, Fitzek FHP, Iera A, Molinaro A (2007) On the beneficial effects of cooperative wireless peer to peer networking. In: *Tyrrhenian international workshop on digital communications 2007 (TIWDC 2007)*, Ischia Island, Naples, Italy, September 2007
7. Militano L, Iera A, Molinaro A, Fitzek FHP (2008) Wireless peer-to-peer cooperation: when is it worth adopting this paradigm? In: International symposium on wireless personal multimedia communications(WPMC), September 2008
8. Perrucci GP, Fitzek FHP, Petersen MV (2008) Chapter in heterogeneous wireless access networks: architectures and protocols – energy saving aspects for mobile device exploiting heterogeneous wireless networks. Springer, Heidelberg
9. Perrucci GP, Fitzek FHP, Zhang Q, Katz M (2009) Cooperative mobile web browsing. EURASIP J Wirel Commun Networking. doi:10.1155/2009/543054
10. Petersen MV, Perrucci GP, Fitzek FHP (2008) Energy and link measurements for mobile phones using ieee802.11b/g. In: The 4th international workshop on wireless network measurements (WiNMEE 2008) – in conjunction with WiOpt 2008, Berlin, Germany
11. http://kom.aau.dk/$\sim$daizal/

Cluster-Based Irresponsible Forwarding

Stefano Busanelli, Gianluigi Ferrari, and Sooksan Panichpapiboon

1 Introduction

In the last decades, Inter Vehicular Communication (IVC) systems have attracted a significant attention from universities, public administrations, and automotive companies. Despite huge efforts, these applications have not yet found the way to the market, but the intensity of the research activity still remains high. As for other ICT technologies, the success or the defeat of IVC systems depends on the appearance of killer applications. Today, the most promising areas seem to be related to accident prevention and vehicular traffic optimization. Due to the their high dynamical nature and the lack of fixed infrastructure (also for economic reasons), typically IVC systems are exploited by the so-called Vehicular Ad-Hoc Networks (VANETs).

In order to satisfy the requirements of the aforementioned applications, several authors have proposed broadcast transmission techniques, but the design of an efficient and reliable broadcasting forwarding protocol is not an easy challenge [1]. Among the various approaches, we focus on two categories of forwarding protocols, namely probabilistic and cluster-based, and we try to merge them. From pioneering works, such as [2], cluster-based networks have found a fertile application ground in the field of wireless sensor networking. In fact, in these applications, cluster-based approaches are beneficial from several points of view: they allow to reduce network congestion, to increase the spectral efficiency, and to simplify routing issues, data aggregation and dissemination. Despite their evolution and their potential advantages, cluster-based networks have not been able to obtain the same success in VANETs. The high dynamism of these networks is one of the main obstacles

S. Busanelli (✉) and G. Ferrari
Department of Information Engineering, CNIT Research UNIT, University of Parma, Parma, Italy
e-mail: busanelli@tlc.unipr.it; gianluigi.ferrari@unipr.it

S. Panichpapiboon
Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
e-mail: sooksan@alumni.cmu.edu

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_6,

against the implementation of cluster-based networking protocols. In fact, the high dynamism leads to a very short lifetime of the clusters, thus yielding a high overhead for cluster construction and maintenance, partially vanishing their potential benefits. Besides these problems, there are nonetheless good reasons for employing cluster-based approaches. One of the strongest motivations is provided by [3], where the authors show that, according to realistic mobility models, VANETs naturally evolve to clustered configurations. Among the more recent cluster-based protocol proposed in VANETs, some interesting approaches can be found in [4, 5]. In the latter work, communications are typically broadcast but, when possible, short-lived clusters are created to constitute a backbone. It is then possible to employ unicast communications among the nodes of the backbone, leading to a higher reliability without sacrificing network performance.

In [6], the authors propose an innovative probabilistic forwarding technique, named irresponsible forwarding (IF), in which every node properly computes its own transmission probability in a per-packet manner, taking in account the vehicle spatial density and the distance from the source. From the simulation analysis of the IF protocol, in IEEE 802.11 networks [7], performed in [8], it emerges that IF is a quite promising approach to broadcasting. Being a probabilistic protocol, however, its reliability is not perfect. Hence, in this work, we apply the concept of IF to derive a new broadcast technique, denoted as cluster-based irresponsible forwarding (CIF), that integrates the probabilistic approach of the original IF protocol with a cluster-based structure, to improve its performance. The key characteristic of CIF is that a clustered structure is not imposed. Rather, CIF opportunistically exploits the "ephemeral" clusters that appear in VANET.

After a short description of the IF protocol in Sect. 2, we will describe the CIF protocol in Sect. 3. In Sect. 4, we will define the simulation setup. Finally, in Sect. 5, we will present and discuss some simulation results, which will highlight the improvement brought by intelligent exploitation of "ephemeral clusters." Sect. 6 will then conclude the chapter.

2 Irresponsible Forwarding

In order to understand the basic operational principle of the IF protocol, we sketch its behavior in a one-dimensional network with a single source placed on the leftmost margin (this is the case, for instance, of a highway lane). After the initial packet transmission from the source, denoted as the 0th hop transmission, the packet is then received by a subset of the source neighbors, that are the potential rebroadcasting nodes. Their union constitutes the so-called 1st transmission domain (while the source itself identifies the 0th *transmission domain*). Every node of the 1st transmission domain extracts a value p uniformly distributed in the interval [0, 1], and then it rebroadcasts it only if $\Delta p \triangleq p_{\text{th}} - p > 0$, where p_{th} is given by the following probability assignment function, originally presented in [6]:

$$p_{\text{th}} = e^{-\frac{\rho_S(z-d)}{c}} \tag{1}$$

where d is the distance between the sender and the receiver (dimension: [m]), ρ_s is the one-dimensional vehicle spatial density (dimension: [veh/m]), z is the node transmission range (dimension: [m]), and c is a shaping coefficient (adimensional). If an intermediate node receives more than one copy of a packet, it makes the rebroadcast decision only upon the reception of the first copy of the packet. All the successive copies are automatically discarded to reduce the network traffic and avoid self-loops. All the nodes that receive a "fresh" packet by a node belonging to the 1st transmission domain contribute to form the 2nd transmission domain. This happens recursively, until the packet is not rebroadcast or reaches the physical network limit.

In [8], the performance of the IF protocol is investigated in a realistic IEEE 802.11 network environment, considering some important performance indicators, introduced in [9], such as the REachability (RE), the number of Saved ReBroadscast (SRB), and the end-to-end delay. The obtained results show that while the IF protocol significantly outperforms a simple flooding protocol in terms of rebroadcast (and energy) savings, it does not guarantee a sufficient reliability to warrant its use in safety-sensitive applications. A simple strategy to increase the reliability would consist of tuning the shaping parameter c in (1) in order to "artificially" increase the number of retransmissions. This approach allows to maintain a short the end-to-end delay, but unfortunately it is feasible and effective only when the traffic load is low. In fact, as shown in [8], when the traffic load is high, even an accurate tuning of the parameter c does not offer significant advantages. This has motivated the integration of the IF protocol with an efficient cluster-based architecture.

3 Cluster-Based Irresponsible Forwarding: The Idea

As previously mentioned, a multihop broadcast protocol can be evaluated according to three strictly correlated metrics: the end-to-end delay, the reliability (here expressed in terms of RE), and the Transmission Efficiency (TE). The goal of the CIF protocol is which of obtain a better tradeoff with respect to the IF protocol.

In order to increase the reliability of the IF protocol, still maintaining its low latency, we propose a hybrid approach that combines the probabilistic broadcast nature of the IF protocol with a "loosely clusterized" VANET structure. Our philosophy is to establish a weak artificial packet flow, having the task of discovering the presence of naturally formed clusters, exactly as the water flow in a river highlights the presence of underwater rocks thanks to the generation of a wave signature. Then, we exploit this informations in order to optimize the forwarding procedure, increasing the reliability and the transmission efficiency, but without building up a true clustered infrastructure. Therefore, we introduce the concept of *Ephemeral Cluster* (EC), that is a short-lived cluster of nodes that is recognized and exploited for a limited period of time (just the duration of a packet retransmission). To clarify

the concept, in Fig. 1 we show two typical transmission domains at a given hop. The upper transmission domain is *sparse*, since there is no node aggregation. In this situation, the IF protocol performs generally well, since the probability assignment function is sufficiently "steep" to effectively select the best retransmitting nodes. The lower transmission domain, instead, contains two ECs, one near to the source and one much farther. As shown in [8], in this scenario, the performance degrades since the probability assignment function tends to be similar for all nodes of a cluster, thus yielding to congestion and collisions. This problem is important especially when the cluster is far from the source, since the presence of several nodes, with roughly the same high retransmission probability (p_{th}), will probably lead to several retransmissions of the same packet.

After this preliminary introduction, we now present the forwarding procedure used by the CIF protocol. At the (generic) ith hop, it can be summarized in four steps, graphically represented in Fig. 2.

1. A packet transmission of a node of the $(i-1)$th transmission domain identifies the ith transmission domain.

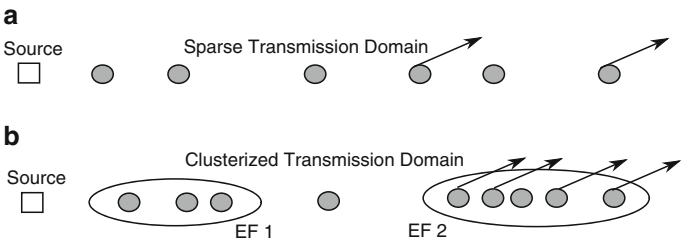


Fig. 1 Two transmission domains (**a**) the upper is *sparse*, while (**b**) the lower contains two ECs

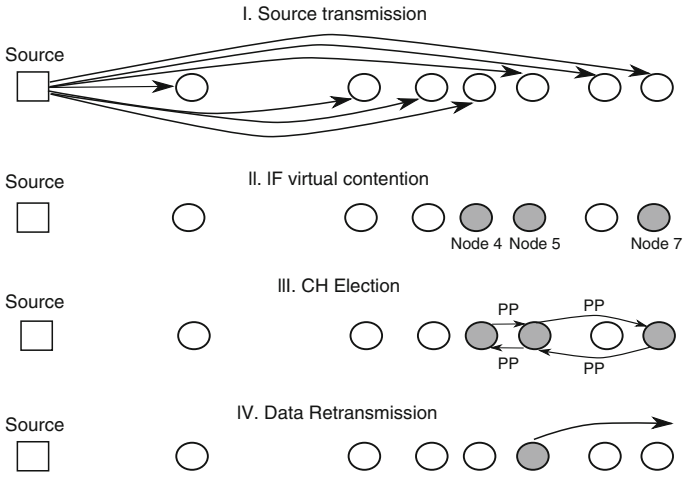


Fig. 2 Forwarding procedure of the CIF protocol

2. The second step derives directly from the IF protocol and is a sort of "virtual contention." In particular, every node in the ith transmission domain decides to become or not a potential forwarder performing the same election mechanism, described in Sect. 2, of the IF protocol. The winners of this contention will begin the third step, while the others will simply discard the packet.
3. The third step derives from the concept of "ephemeral cluster." Once a node wins the first virtual contention, it schedules the retransmission of a very short packet, denoted as probe packet (PP). A PP bears just two information (1) the unique identification (ID) number of the packet to be retransmitted; (2) the instantaneous difference $\Delta p = p_{\text{th}} - p$. The PPs are intrinsically single hop, i.e., they are not forwarded. A PP is transmitted with a low-transmission power, since a node is interested only in signaling its presence to its neighbors, and with a high priority, in order to reduce the overall latency. Moreover, a low transmission power allows to reduce channel interference. The specific power and priority setting of a PP have to be tuned according to the used medium access control (MAC) protocol, and their values in the scenarios of interest will be given in Sect. 5. After winning the virtual contention, every potential forwarder sends a PP. It then waits for a short prefixed interval, denoted as T_{w}: if, within this interval, it receives at least a PP containing a value of Δp larger than its own, it stops and discards the packet (in fact, there is some other better forwarder); conversely, it retransmits the packet. In the worst case, when a collision between two or more PPs happens, this selection mechanism fails and no node of the cluster is elected. In this case, all nodes will retransmit in order to guarantee a high reachability.
4. The fourth step corresponds to the transmission act from the designated forwarding nodes.

4 IEEE 802.11 Network Simulation Setup

We use the IEEE 802.11 model present in Network Simulator 2 (ns-2.31 [10]), sending small size packets (105 B) in order to prevent fragmentation, using the default values of the vanilla ns-2 installation that refers to the IEEE 802.11b standard. On top of the IEEE 802.11 MAC layer, we insert the IF forwarding protocol. Since this work does not focus on physical layer issues, we adopt a simple Friis free-space propagation model [11]. We also consider a static scenario, where the nodes are placed along a straight line of length L (dimension: [m]) and their positions are generated according to a Poisson distribution of parameter ρ_{s} (dimension: [veh/m]) – this is representative of highway scenarios with cars moving at similar speeds. In order to have a fair comparison between the results obtained with different values of z (dimension: [m]), we vary the network length proportionally to the transmission range, setting $L = 8z$. For instance, with a transmission range $z = 500$ m we consider a portion of road of 4 km in front of the source vehicle.

As in [6], there is a single source placed on the left vertex of the linear network, so that packets flow from left to right. For every (ρ_{s},z) pair, there is a nonzero

Frequency	2.4 GHz
Channel bandwidth	2 MHz
PLCPDataRate	1 Mbps
Data rate	1 Mbps
CW_{MIN}	31
SlotTime	$20\mu s$
SIFS	$10\mu s$
PreambleLength	144 bit
PLCPHeaderLength	48 bit

Fig. 3 Parameter of the IEEE 802.11b standard used in the simulations

probability of having a distance d between two consecutive nodes, say k and $k+1$, larger than the transmission range z, since d is exponentially distributed. If $d > z$, the $(k+1)$th node is unreachable and the kth one becomes the last reachable node (lrn) of that particular scenario. When lrn$\neq N$ the network is said topologically disconnected, whereas if lrn$= N$, the network is topologically connected.

The source sends a burst of 1,000 packets using a Poisson distribution with parameter λ (dimension: [pck/s]).[1] We use a value of λ equal to 100 pck/s, so that considering the packet size of 105 bytes leads to an average load of 84 kbps, that is significant with respect to the available data rate that is of 1 Mbps, as shown in Fig. 3. Two values of the parameter c, namely 1 and 5, are adopted, representing, respectively, weak and aggressive rebroadcasting policies. The results are obtained for a fixed node density value $\rho_s = 0.01$ vehicle/m, while the transmission range z assumes the values in the set $\{0.1, 0.3, 0.5, 0.75, 1, 1.5, 2, 3\}$ km, in order to have the desired value of the product $\rho_s z$. Clearly, the transmission range is obtained setting the suitable value of the transmission power. For small values of z, the network is rarely connected since $Pr\{d > {\rho_s}^{-1}\}$ is relatively high. On the other end, the network gets connected with a high probability (almost 1), if z is larger than 750 m.

Clearly, the IEEE 802.11 interfaces operate in the ad-hoc mode, and they send packets in a broadcast fashion. In this configuration, the distributed coordination function (DCF) cannot exploit the ready-to-send/clear-to-send (RTS/CTS) mechanism, since the latter is a viable strategy only for unicast communications. For the same reason, the ACK messages are also ineffective and, therefore, they are disabled. Hence, the hidden terminal problem is unsolved and retransmissions cannot happen at the MAC layer, since the sender cannot get information about the status of its communications. Without retransmissions, the contention window (CW) of the carrier sense multiple access with collision avoidance (CSMA/CA) MAC protocol is never increased and always assumes its initial value specified by the parameter CW_{MIN} of the IEEE 802.11 standard [7]. The parameters of the IEEE 802.11 standard relevant for the simulations are listed in Fig. 3.

[1] Our simulations show that the numbers of the generated scenarios (1,000) and of the transmitted packets (1,000) are sufficient to guarantee an interval of confidence greater than 95%; thus, we will omit any error considerations in our results analysis.

Finally, the CIF protocol foresees the use of two types of packets, data and probe, which require two different services from the lower layers. In particular, a PP requires a higher transmit priority and a lower power than a data packet . In order to obtain a higher priority, we set CW_{MIN} to 7, instead of the value of 31 used for data packets. On the other hand, as will be shown in Sect. 4, the transmit power can be set to different values to vary the transmission range. In all cases, PPs are transmitted with a power equal to 20% of the transmit power used to send data packets. The waiting time T_{w} is set to 10 ms.

5 Numerical Results

In Sect. 3, it was anticipated that three metrics will be used to assess the behavior of the CIF protocol: RE, TE, and end-to-end delay. The latter is the duration of the packet traveling time between its transmission instant at the source and its reception instant[2] at the lrn. The end-to-end delays of the IF and CIF protocols, with two different values of c (1 and 5), are shown in Fig. 4 as functions of the product $\rho_{\mathrm{s}}z$. As expected, the introduction of an election procedure increases the overall latency for both values of c. However, the delay still remains acceptable. In fact, according to [12], a good latency value for the prevention of chain car collision is about 0.1 s and from Fig. 4 one can observe that the overall latency for all $\rho_{\mathrm{s}}z$ values is lower or very close to this value. One has to keep in mind that the end-to-end delay is measured at the lrn that could be considerably far from the source, i.e., for a transmission range

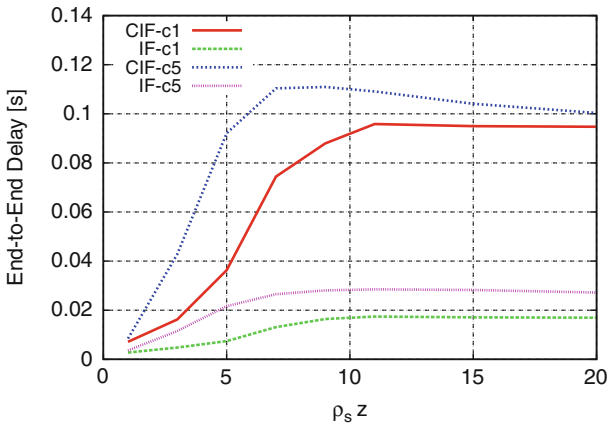


Fig. 4 End-to-end delay as function of the product $\rho_{\mathrm{s}}z$ for various combination of protocols and values of c. In particular, for both IF and CIF protocols two values of c are considered (1 and 5)

[2] We remark that only the packets received correctly at the lrn are considered in the end-to-end delay evaluation phase; hence, this metric is an upper bound of the end-to-end delay experienced by the network nodes.

of 1 km ($\rho_s z \leq 5$) the network dimension is $L = 8$ km. Finally, one can observe that for small values of the product $\rho_s z$ (e.g., 10), the latency is small. In fact, in this region, the network has a limited connectivity and the lrn could be in the proximity of the source, whereas in the other cases its distance is comparable with L.

The RE, introduced in [9], corresponds to the fraction of nodes that receive the source packet among the set of reachable nodes, i.e., those topologically connected to the source. In our case, the number of reachable nodes coincides with lrn. Intuitively, the RE is inversely proportional to the distance from the lrn, since the farther is the lrn, the larger is the number of hops required to reach it. Clearly, since IF is a probabilistic protocol, at every hop there is a nonzero probability of no effective transmission, thus cutting off the packet flow. This induces a singular consequence: with the same value of $\rho_s z$, less connected networks (with a smaller lrn) could have a higher RE since the number of hops required to reach the lrn reduces.

The TE is a novel metric, here introduced for the first time, that is somehow related to the concept of Saved REbroadcasts (SRB), originally introduced in [9]. In particular, for a given packet, we define the TE as the ratio between its RE and the overall number of retransmissions that is experienced during its trip towards the lrn. For instance, given a network with lrn $= 100$, a packet that reaches 80 nodes, through an overall number of retransmissions equal to 20, will lead to measure a value of RE equal to $80/100 = 0.8$ and a value of TE equal to $0.8/20 = 0.04$. Roughly speaking, an ideal forwarding protocol for safety-related vehicular applications should minimize the latency, still guaranteeing the highest possible RE. Instead, the TE is an indicator of the ability of the protocol of selecting the optimal forwarding node. Figures 5 and 6 show, respectively, the REs and TEs obtained with the IF and CIF protocols, using two different values of c (1 and 5) as functions of the product $\rho_s z$. From Fig. 5, one can observe that the classical IF protocol performs very poorly with the selected traffic load of 100 pck/s. In particular, when

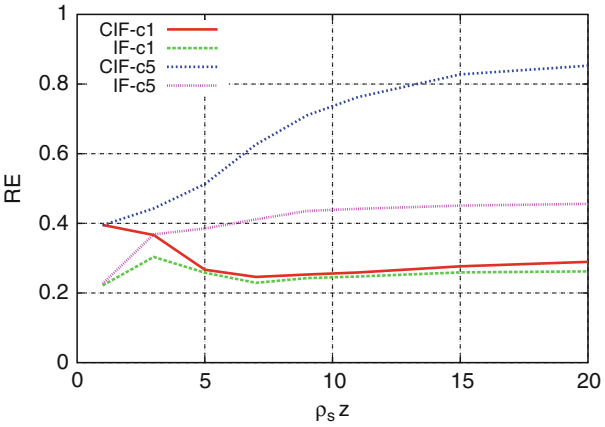


Fig. 5 RE as function of the product $\rho_s z$ for various combination of protocols and values of c. In particular, for both IF and CIF protocols two values of c are considered (1 and 5)

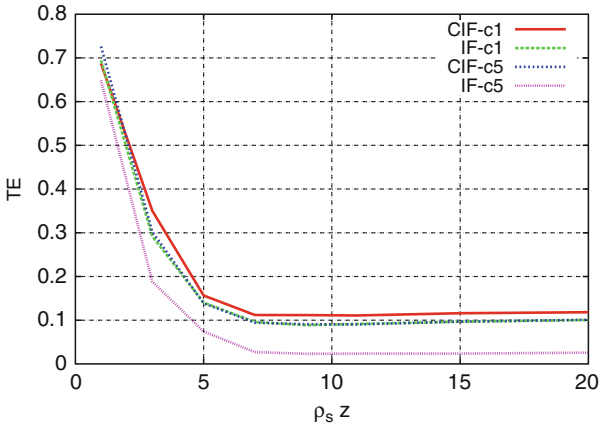


Fig. 6 TE as function of the product $\rho_s z$ for various combination of protocols and values of c. In particular, for both IF and CIF protocols two values of c are considered (1 and 5)

$c = 1$, the IF protocol cannot self-sustain itself and the RE remains around 0.2. With $c = 5$, the behavior is slightly better but the RE is still limited by collisions and congestion. The CIF protocol shows a similar behavior when $c = 1$. In fact, in this case, the improved relay selection mechanism does not offer a significant advantage, since the average number designated by the virtual contention is small (on average around 1, as shown in [6]). Conversely, CIF offers a significant improvement when $c = 5$, since it reduces the congestion of the channel, thus limiting the number of the data packet transmissions by substituting them with quick and low-power PP transmissions. Finally, Fig. 6 shows similar results. In particular, the CIF protocol outperforms the IF protocol for both values of c, but especially when c is equal to 5, where the improvement is particularly evident. Counterintuitively, the TE assumes higher values in the scarcely connected region. This is motivated by the fact that in this region, the single retransmission has a stronger impact on the RE since the lrn and the number of hops are typically small.

6 Conclusions

In previous works [6, 8], we have shown that the IF protocol can offer a high reachability when the shaping factor c is sufficiently high to self-sustain the protocol (typically, $c = 5$ is a good value). Unfortunately, the RE of the classical IF protocol degrades when the network traffic load increases, since a high value of c leads to a high number of retransmissions. In order to reduce the congestion, we have presented a novel approach to exploit the ephemeral clusterization that naturally emerges in VANETs, thus leading to a novel probabilistic forwarding protocol, denoted as CIF. The adopted approach brings significant benefits, in terms of RE and

of TE, with respect to the IF protocol. We also observed that the latency remains within the limit imposed by the target applications (safety-related applications in VANETs). Finally, we emphasize that one of the strengths of the CIF protocol is the limited required amount of information on the network topology. In particular, it requires the knowledge of just one topology parameter, the vehicular spatial density ρ_s, that could be obtained combining long-term statistics about the vehicular traffic and local estimations of the density. Hence, unlike other approaches (see, for example, [13]), the CIF protocol represents an efficient event-driven forwarding protocol, without the need of an auxiliary logical channel for exchanging local information between the nodes.

References

1. Li F, Wang Y (2007) Routing in vehicular ad hoc networks: a survey. IEEE Mag Vehicular Technol 2(2):12–22
2. Lin CR and Gerla M (1997) Adaptive clustering for mobile wireless networks. *IEEE J Sel Areas Commun* 15(7):1265–1275
3. Fiore M and Härri J (2008) The networking shape of vehicular mobility. In: Proceedings of the ACM international symposium on mobile ad hoc networking and computing (MobiHoc '08), New York, ACM, pp 261–272
4. Wang Z, Liu L, Zhou M, Ansari N (2008) A position-based clustering technique for ad hoc intervehicle communication. IEEE Trans Syst Man Cybern C: Appl Rev 38(2):201–208
5. Bononi L, Felice MD (2007) A cross layered MAC and clustering scheme for efficient broadcast in VANETs. In: IEEE internatonal conference on mobile adhoc and sensor systems (MASS 2007), Montreal, QC October 2007, pp 1–8
6. Panichpapiboon S, Ferrari G (2008) Irresponsbile forwarding. In: *Proceedings of the IEEE, 8th international conference on intelligent transport system telecommuniction* (ITST'08). Phuket, Thailand, October 2008, pp 311–316
7. Insitute of Electrical and Electronics Engineers (2007) IEEE Std 802.11TM-2007. Part 11: wireless LAN medium access ontrol (MAC) and physical layer (PHY) specifications
8. Busanelli S, Ferrari G, Panichpapiboon S (2009) Efficient Broadcasting in (IEEE) 802.11 Networks through Irresponsible Forwarding. Proc. IEEE Global Telecommun. Conf. (GLOBECOM), Honolulu, HI, USA, October 2009
9. Ni S, Tseng Y, Chen Y, Sheu J (1999) The broadcast storm problem in a mobile ad hoc network. In: Proceedings of the ACM international conference on mobile computing and networking (MOBICOM), Seattle, WA, pp 151–162
10. Network Simulator 2 (ns-2). [Online]. Available: http://isi.edu/nsnam/ns/
11. Rappaport TS (2002) Wireless communications. In: Principles & Practice, 2nd edn. Prentice-Hall, Upper Saddle River, NJ
12. Biswas S, Tatchikou R, Dion F (2006) Vehicle-to-vehicle wireless communication protocols for enhancing highway traffic safety. IEEE Commun Mag 44(1):74–82
13. Torrent-Moreno M, Mittag J, Santi P, Hartenstein H (2009) Vehicle-to-Vehicle Communication: Fair Transmit Power Control for Safety-Critical Information. IEEE Trans. Veh. Technol, September 2009, 58(7):3684–3707

Dynamic Spectrum Access Communications: Wavelet Modulation with Unequal Power Allocation

Marco Lixia and Maurizio Murroni

1 Introduction

Dynamic Spectrum Access (DSA) techniques are becoming a key issue for heterogeneous communication environment characterized by the sharing of the spectrum and the coexistence among various radio access nodes. These ideas gain in importance especially with respect to the vision of Internet of Things, where a multitude of different devices are expected to communicate and rearrange their network configuration in an autonomous way to route/exchange the information in the most efficient way. Finding new spectral resources or developing new techniques to assure a more flexible and efficient utilization of communication channels is a key issue for engineers. Nowadays, wireless systems are highly common, especially in urban and indoor environment, and the channel capacity is often shared by a large number of independent communications. The a priori knowledge of bandwidth (B) or time (T) availability of these channel portions is often unavailable due to their unpredictable fluctuations. In these particular conditions, complex adaptive schemes can be adopted to sense the channel and transmit when an opportunity is found. This approach is referred as DSA communication. A transmission opportunity can be described by the product $B \times T$, where B and T are unknown, in the sense that a channel can be available for a fixed elapse of time with unknown bandwidth or vice versa.

Wornell and Oppenheim first proposed Wavelet Modulation (WM) to transmit signal over a variable channel in time or frequency but with a fixed value of the product $B \times T$ [1–3]. WM exploits the properties of dy-homogeneous signals to transmit over multiple time scales and band frequencies. In this chapter, we propose a modified WM scheme, which allows the robust transmission of data within the framework of DSA networks. Modern communications often deal with the transmission of bitstream highly correlated. This is the case of multimedia data format, such as MPEG

M. Lixia (✉) and M. Murroni
Department of Electrical and Electronic Engineering, University of Cagliari,
09123 Cagliari, Italy
e-mail: marco.lixia@diee.unica.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_7,

or JPEG, where a hierarchical structure assigns different error sensitivity to data. Modulation with Unequal Power Allocation (MUPA) was proposed by Bruggen and Vary to improve the performance of conventional modulation schemes in the case of digital communication systems, which do not include, for some reason, channel coding (i.e., FEC codes) [4–6]. MUPA distributes the available budget power over the data, according to their sensitivities to channel error: it is assigned higher transmission power to more sensible data. As a result, it allows achieving improved final quality on transmitted data without any increase of the transmission bandwidth. In practice, MUPA is performed by assigning different weights to the data according to their "importance" (i.e., channel error sensitivities), whereas the average transmission power per symbol remain unchanged. As to this, it is a "continuous" process in the sense that weights are chosen in a real set with an accuracy that can be a priori selected and in theory infinite. This approach adds to the transmission scheme more flexibility compared to conventional FEC. Channel coding schemes add redundancy to the signals to be transmitted with a rating code that is constrained to assume only selected values [7]. Within a framework of DSA communications, in this chapter, we propose a technique to improve the performance of the WM based on an unequal distribution of the energy assigned at each bit in function of both channel availability ($B \times T$) and error sensitivity of transmitted data. As to the optimization criterion for the selection of the weights, we consider the minimum mean square error (MMSE) between transmitted and received data. We derive a numerical solution to the power weight optimization using Genetic Algorithms (GA) [8]. GA are search techniques used in computing to find true or approximate solutions to optimization problems. GA are extensively used in literature in different application fields of communication engineering such as for instance, network design, unicast and multicast routing [9, 10]. They allow finding iterated numerical solution to complex problems with an accuracy dependent on the number of iterations selected. GA can deal with highly nonlinear problems and nondifferentiable functions as well as functions with multiple local optima. They are also readily amenable to parallel implementation, which renders them suitable for real-time adaptive wireless communications. Test conducted show the effectiveness of the proposed method in case of transmission over DSA networks.

The rest of the chapter is organized as follows: first, Wavelet Modulation and Modulation with Unequal Power Allocation theory are reported in Sects. 2 and 3, respectively. Section 4 shows the Weights Optimization problem. Then, the proposed system model and simulation results are presented in Sect. 5.

2 Wavelet Modulation

Wornell and Oppenheim showed that dy-homogenous signals can be used to transmit information over multiple time scales and band frequencies. For a given orthonormal wavelet basis, each wavelet can be a dilation or translation of a single basic wavelet $\psi_n^m = 2^{m/2}\psi(2^m t - n)$, where m and n are the dilation and translation

indices, respectively. Using transmitted data as coefficient in the wavelet expansion is possible to generate a signal, which is able to transport data at different rate:

$$x(t) = \sum_m \sum_n \beta^{-m/2} q[n] \psi_n^m(t) \tag{1}$$

where $\beta = 2^{2H+1}$ and $q[n]$ are the data to be transmitted. The Hurst parameter H in WM controls the relative power distribution among frequency bands and, hence, the overall transmitted power spectrum.

Wornell and Oppenheim's research focuses on transmission over AWGN channel with unknown bandwidth or time availability. These characteristics implied that, at the receiver side, the system must be able to recover the bit-stream $q[n]$ using an arbitrarily narrow band given a sufficient period of time or, vice versa, using a arbitrarily short period of time given a sufficient receiver band. Furthermore, the system must be able to obtain better estimation of $q[n]$ the longer it listens or the greater the available bandwidth is. In WM, the signal is a binary finite bit-stream with $q[n] \in \left\{+\sqrt{E_\mathrm{b}}, -\sqrt{E_\mathrm{b}}\right\}$. At the receiver side, Discrete Wavelet Transform (DWT) algorithm is used to extract data. The duration–bandwidth characteristics of the channel in general determine the observation coefficients that can access and, hence, the available redundancy. If the channel is band-limited to 2^{T_U} Hz for some integer T_U, this precludes access to the coefficients at scales corresponding to $m > T_U$. Simultaneously, the duration constraint in the channel results in a minimum allowable decoding rate of 2^{T_L} symbols/s for some integer T_L, which precludes access to the coefficients at scales corresponding to $m < T_L$. Therefore, the available number of message measurements is $K = \sum_{m=T_L}^{T_U} 2^{m-T_L} = 2^{T_U - T_L + 1} - 1$. As a consequence, the message $q[n]$ can be recovered at a rate of 2^m by means of a band of 2^{m+1} Hz, that is, WM is characterized by a spectral efficiency of 0.5 symbol/s/Hz.

The error probability can be expressed as a function of both rate and bandwidth as:

$$\Pr(\varepsilon) = Q\left(\frac{1}{2}\sqrt{\sigma_\mathrm{c}^2\left[\frac{2\eta_\mathrm{F}}{R \backslash W} - 1\right]}\right) \tag{2}$$

where $\eta_\mathrm{F} \approx 1/2$ and σ_c^2 represents the signal-to-Noise Ratio (SNR).

Figure 1 shows the time–frequency portrait for WM. The biggest is the transmission time availability and the smallest is the bandwidth required to receive the signal and vice versa. Figure 2 shows the WM modulator and demodulator.

3 Modulation with Unequal Power Allocation

In MUPA, an unequal redistribution of the energy in accordance with the error sensibility of single bits by means of so-called weights w_i, $i = 1, 2, \ldots, M$, where M is the number of bits per frame, is performed.

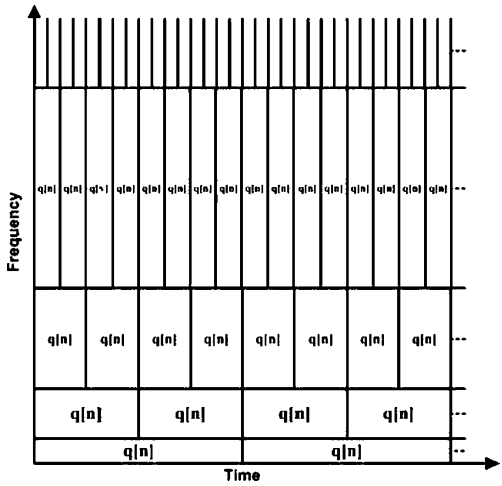


Fig. 1 Example of time-frequency portrait for WM with $H = 1/2$

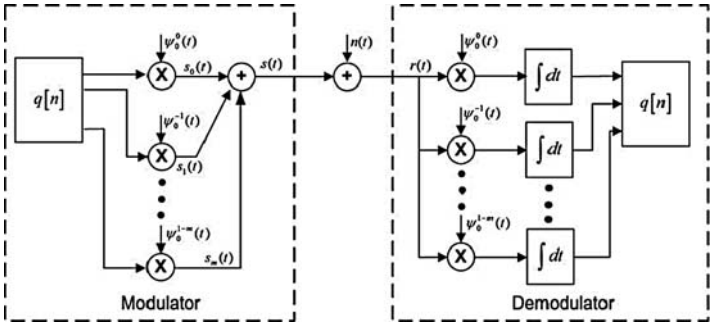


Fig. 2 WM communication system

Without loss of generality to model a generic bit-stream that exhibits different error sensitivities to channel conditions, we consider a discrete memoryless source (with period τ) $S : \forall\tau \rightarrow u(\tau)$ and an analog-to-digital process $AD : \forall u(\tau) \rightarrow \underline{x}(\tau)$ with $\underline{x}(\tau) \in (\underline{x}_k \,|k=1,2,\ldots,2^M)$, $\underline{x}_k = (x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(M)})$, $x_k^{(M)}$ being the LSB. Each $x_k^{(i)}$ is then multiplied with the specific weight $w_i \in \mathbb{R}^+$ of the diagonal matrix $\underline{W} = diag(w_1, w_2, \ldots, w_{\dot{M}})$ (computed off-line as shown in Sect. 4). The weighted bit pattern $\underline{y}(\tau) = \underline{W} \cdot \underline{x}(\tau)$ is then transmitted by a WM over a wireless channel affected by additive white Gaussian noise (AWGN) $n(t)$ with zero mean and variance $N_0/2$. After demodulation, the distributed vector is $\underline{z}(\tau) = \underline{y}(\tau) + \underline{n}_{\text{rel}}(\tau)$, where $\underline{n}_{\text{rel}} = (n_{\text{rel}}{}^{(1)}, n_{\text{rel}}{}^{(2)}, \ldots, n_{\text{rel}}{}^{(M)})$ represents the demodulated noise along the M signal message components (i.e., relevant noise). Following decision based on Maximum Likelihood (ML) criterion, the estimate $\hat{u}(\tau)$ is produced by inverse digital-to-analog (DA) process. A sketch of the system is depicted in Fig. 3.

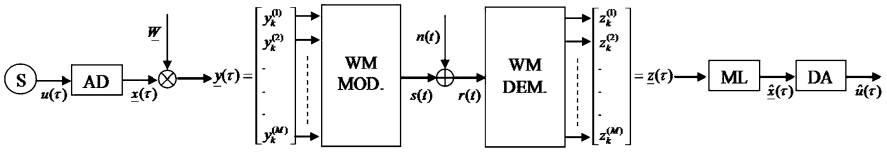


Fig. 3 System model for MUPA-WM

Considering bipolar binary representation $x_k^{(i)} = \pm 1$, if bits in $\underline{x}(\tau)$ are inverted due to channel impairment, a wrong decision $\hat{x}(\tau)$ is made at the receiver, thus producing a distortion $d(\tau)=[u(\tau) - \hat{u}(\tau)]$. Aim of the optimization process is to calculate optimal weights in the sense of a minimized expected value $E\{[d^2(\tau)]\}$. Assuming ergodicity, it is possible to calculate $E\{d^2\}$ as follows:

$$E\{d^2\} = \sum_{k=1}^{2^M} \sum_{h=1}^{2^M} d_{k,h}^2 P(\underline{x}_k) \cdot P(\underline{\hat{x}}_h | \underline{x}_k) \tag{3}$$

where $d_{\varsigma,\eta} = u_\varsigma - \hat{u}_\eta$ is the difference of possible parameter values, $P(\underline{x}_k)$ the occurrence of the reproduction levels u_k, and $P(\underline{\hat{x}}_h | \underline{x}_k)$ the transition probabilities between transmitted and received bit patterns. Due to the orthogonal properties of wavelet waveforms and to the independence of the noise samples, the transition probabilities are:

$$P(\underline{\hat{x}}_h | \underline{x}_k) = \left(\prod_{\substack{i=1 \\ x_k^{(i)} \neq \hat{x}_h^{(i)}}}^{M} P_{\mathrm{b}}^{(i)} \right) \cdot \left(\prod_{\substack{i=1 \\ x_k^{(i)} = \hat{x}_h^{(i)}}}^{M} \left(1 - P_{\mathrm{b}}^{(i)}\right) \right) \tag{4}$$

Within the MUPA framework, under the hypothesis of average energy normalized transmitted frames:

$$E_{\mathrm{b}} = \frac{1}{M} \sum_{i=1}^{M} E_{\mathrm{b}}^{(i)} = 1 \tag{5}$$

It is possible to write $E_{\mathrm{b}}^{(i)} = w_i^2 E_{\mathrm{b}}$ and impose the following constraint on the weights w_i:

$$\sum_{i=1}^{M} w_i^2 = M \tag{6}$$

where M is the number of bits in a frame and the bit probability is as in (2). Mathematically, the optimization problem is to minimize (3) under the constraint (6). In other words, MUPA raises ($w_i > 1$) the immunity to noise channel for more significant bits, paying as a counterpart lower robustness($w_i < 1$) on less

significant one, to achieve average improved performance on the transmission of parameter $u(\tau)$ in the sense of minimum expected distortion $d(\tau)=[u(\tau)-\hat{u}(\tau)]$. The complexity of the above optimization problem, which increases with the size of frames M, does not allow closed form solutions. Therefore, to identify the solution, we use a numerical approach based on GA.

4 Weights Optimization

GA are implemented as a computer's simulation in which a population of abstract representations (*chromosomes*) of candidate solutions (*genes*) to an optimization problem evolves toward better solutions. The evolution usually starts from a population of randomly generated chromosomes and happens in generations. In each generation, the fitness of every chromosome in the population is evaluated, then multiple chromosomes are stochastically selected from the current population (based on their fitness), and modified (mutated or recombined) to form a new population. The new population is then used in the next iteration of the algorithm. In the proposed system, the chromosomes are defined as arrays of $M=8$ genes $w_i \in \mathbb{R}^+$. The range of possible values of w_i is constrained by (6). An initial population $\{INIT\}$ of L chromosomes is randomly selected. The fitness function is as defined in (3). Two operations are allowed to determine the evolution of the initial population: *crossover* (with probability P_{cross}), used to interchange the elements of two chromosomes, and *mutation* (with probability P_{mut}), which modifys the value of one or more genes within a chromosome with the aim of leading the search out of local optimal. In particular, the most fitting part of the population $\{BEST\}$ is selected and directly inserted in the new generation, while the rest of the population $\{WORST\}$ is discarded and replaced by a subpopulation created by means of the crossover and mutation operators. The termination condition is satisfied once either the algorithm reaches a selected number of iterations (IT) or the fitness function maintains the same value for IT_{MAX} iterations. At the end of the process, the chromosome with low score in the fitness function (i.e., lower distortion on the reconstructed frame) will be selected for the transmission. The accuracy of such an approach is strictly dependent on the values of IT and IT_{MAX}, whereas the complexity of the algorithm depends also on the definition of chromosomes, on the size L of the initial population, and on the P_{cross} and P_{mut} probabilities. Chromosomes are arrays of genes that are real values. The higher the precision on the representation of the genes (i.e., the number of decimal digits used to approximate real values) the higher the accuracy achieved by the UPA, but also, the higher the complexity of the algorithm. Similarly, big size populations guarantee higher performance, but also lead to time-consuming processing. A critical matter is the selection of P_{cross} and P_{mut} probabilities: high values can determine instability of the GA that could diverge, whereas on the other side, low values are likely lead to slow convergence.

5 Simulation and Testing

Before testing the system, a preliminary analysis on the GA parameters has been performed. At each iteration, a maximum of 50% of best chromosome parents (i.e., $\{BEST\}$) have been selected to create the next generation. P_{cross} and P_{mut} were selected as to assure that maximum 20% of the $\{BEST\}$ population remained unchanged. According to this, L was varied in the set {8 16 32 64 128},IT_{MAX} was varied in the set {20 40 60 80 100}, P_{cross} and P_{mut} in the range 0.3–0.7 and 0.01–0.3, respectively. Finally for mutation, Δ varied within the range $0<\Delta \leq 0.5$.

The maximum difference in terms of fitness function value among all the solutions was observed to be less than 15%. Therefore, the following considerations can be made: huge size populations bring to better solutions at the expense of a higher processing time; the P_{mut} probability is suggested to be set equal to or higher than 0.1 to avoid an excessive number of iterations; the P_{cross} probability does not sort significant effects in the used range. Figures 4 and 5 show the fitness function performance depending on L and IT_{MAX} variations, respectively for different values of SNR.

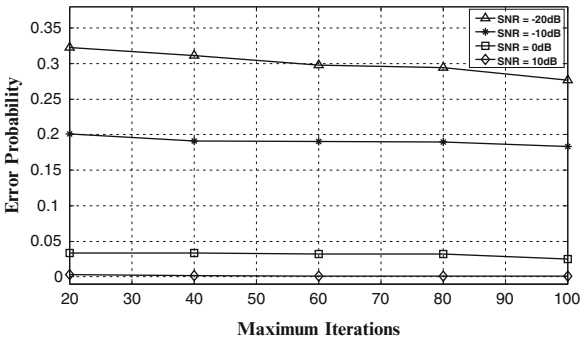


Fig. 4 Error probability as a function of IT_{MAX}

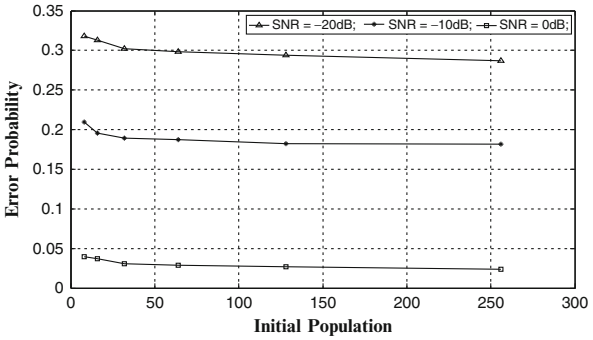


Fig. 5 Error probability as a function of the initial population

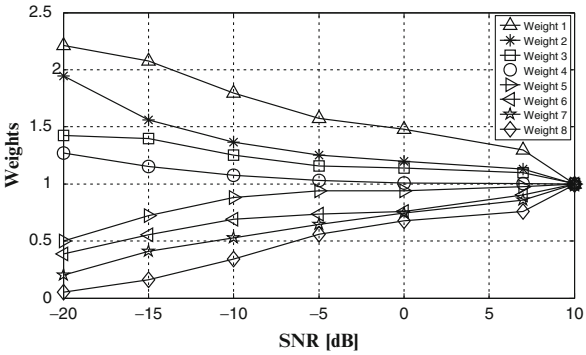


Fig. 6 Weights distribution in function of SNR variations. Weight 1 is associated to the MSB, consequently weight 8 is associated to the LSB

Figure 6 shows weights values as a function of SNR. It is possible to note that for low values of SNR the MSB's weight has a high value, about 2.2, whereas the value of the LSB's weight is about 0.2.

By increasing the value of SNR, it is possible to note that weights values converge to unity.

This happens because the most the channel condition are good, the less there is sensibility difference between MSB and LSB. As to the outcome from the preliminary tests on GA behavior applied to the MUPA problem, in the following experiments the $\{INIT\}$ population was composed by $L=256$ chromosomes, eight decimal digits were used to represent genes (i.e., the weights w_i), the probability $P_{\text{mut}}=0.3$, and $P_{\text{cross}}=0.5$, $IT_{\text{MAX}}=100$ and $IT=1000$. To simulate the system shown in Fig. 3, a source that generates sequences of random integers $u(\tau) \in \{0, 255\}$ has been used. These sequences are then sent into the A/D block where they are converted in frames of $M=8$ bits. Frames are then opportunely weighted and modulated by the WM transmitter. For WM, Daubechies wavelets of fourth order with dilation indices $m=\{10, 11, 12, 13\}$ have been deployed. Modulated signal is then transmitted over the channel using $\beta=1$ and so $H=-1/2$.

We simulated an AWGN channel with zero-mean and power spectral density $N_0/2$, but with discontinue availability. Figure 7 shows an example of $B \times T$ representation of such a channel: gray spaces represent unavailable resources that cannot be used to interfere with primary communications. On the other hand, transmission is allowed in the remaining white spaces.

The proposed model reproduces the channel conditions that are typical of DSA communication. Our system is based on the perfect knowledge of the channel opportunities statistic. This can be achieved in a real scenario by a spectrum sensing procedure to be periodically performed before transmission. WM allows to exploit the available opportunities and adapt the communication to periodically updated channel availability. To do this, wavelets corresponding to unavailable spaces are switched off. In other words, if $B \times T$ value does not allow to transmit, the signal at all the wavelet carriers that would disturb any already existing transmission are

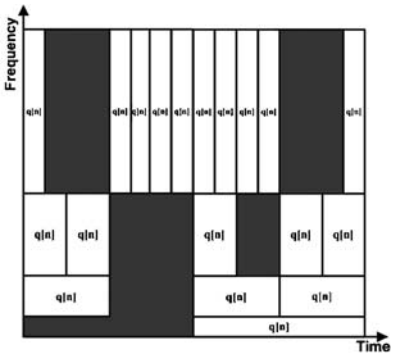


Fig. 7 Example of $B \times T$ channel representation: grey spaces represent unavailable portions of the channel whereas white spaces transmission opportunities

suppressed. This allows achieving both power saving and efficient communication since the unused power budget saved by the suppression of potentially interfering carriers can be redistributed to increase noise immunity of the wavelets carrier transmitted in the available opportunities.

In the band associated to the dilation parameter value $m=10$, data are transmitted at a rate of $R_{10}=1024$ bps, which corresponds to a bit transmission time of $T_{10}=0.977$ ms. At receiver side, a bandwidth $B_{10}=2^{(10+1)}=2048$ Hz is needed to extract the signal. Instead, for $m=13$, the transmission bit-rate is $R_{13}=8192$ bps and the bit transmission time is $T_{13}=0.122$ ms. A bandwidth of 16.384 kHz is necessary to receive the signal.

In order to evaluate performances of proposed system, Mean Square Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) have been calculated:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} \|u_i - \hat{u}_i\|^2 \tag{7}$$

where N is the size of the transmitted sequence.

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right) \tag{8}$$

where MAX_I^2 is maximum possible output value of u_i.

Several tests have been conducted with numerous different channel opportunities conditions randomly selected with fixed period. Figure 8 shows the PSNR values as a function of the SNR of WM with Unequal Power Allocation (WMUPA) respect to benchmark WM with Equal Power Allocation (WMEPA) systems. In this particular case $T=1$ ms. PSNR trend shows an improvement of the performance of the proposed system for the transmission of data, which exhibit different bit error sensibility to channel error. Precisely, WMUPA outperforms WMEPA along all the variation range of the SNR with a peak gain of 4.0564 dB for a value of SNR of -4 dB.

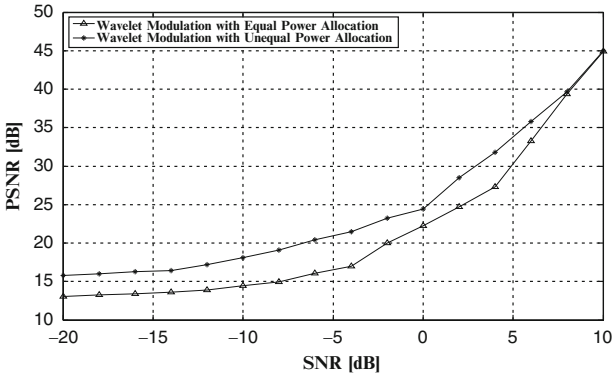


Fig. 8 PSNR vs. SNR for WMUPA and WMEPA

References

1. Wornell GW, Oppenheim AV (1992) Wavelet-based representations for a class of self-similar signals with application to fractal modulation. IEEE Trans Inf Theory 38:785–800
2. Wornell GW (1993) Wavelet-based representation for the 1/f family of fractal processes. Proc IEEE 81:1428–1450
3. Ptasinski HS, Fellman RD (1994) Performance analysis of a fractal modulation communication system. Proc SPIE Wavelet Appl Conf 2242:78–86
4. Bruggen T, Vary P (2005) Analysis of digital modulation with unequal power allocation IEEE Commun Lett 2:1143–1147
5. Bruggen T, Vary P (2005) Unequal error protection by modulation with unequal power allocation. IEEE Commun Lett 9(6):484–486
6. Bruggen T, Schulte-Hillen C, Vary P (2005) Soft demodulation and unequal error protection for digital modulation schemes. In: Proceedings of IEEE ICASSP, Philadelphia, PA, USA, Mar 2005
7. Benedetto S, Biglieri E (1999) Principles of digital transmission: with wireless applications. Kluwer Academic Publishers, New York
8. Whitley D (1994) A genetic algorithm tutorial. Stat Comput 4:65–85
9. Atzori L, Raccis A (2004) Network capacity assignment for multicast services using genetic algorithms. IEEE Commun Lett 8(6):403–405
10. Ahn CW, Ramakrishna RS (2002) A genetic algorithm for shortest path routing problem and the sizing of population. IEEE Trans Evol Comput 6(6):566–579

Optimal Cross-Layer Flow-Control for Wireless Maximum-Throughput Delivery of VBR Media Contents

Enzo Baccarelli, Mauro Biagi, Nicola Cordeschi, Tatiana Patriarca, and Valentina Polli

1 Introduction and Goals

Over the last years, a strong proliferation in the use of multimedia technology has been experienced, triggered by several reasons. First, wireless networking architectures (i.e. media CDNs and WLANs) are becoming an integral part of communication environments [1, Chaps. 5–7]. Second, the utilization of new hand-held wireless devices (such as media phones, PDAs, and laptops) is becoming common. Third, new multimedia applications (streaming stored/live audio/video) became popular first in Internet wired environments and now in wireless mobile environments [2, Chaps. 14 and 15]. However, these wireless media opportunities bring with them several technological challenges. In fact, wireless networks suffer from a large variation in connection conditions, mainly due to fading phenomena, mobility, co-channel and multiple-access interferences. Since the currently deployed network protocols (as, for example, the Internet [1, Chap. 4]) offer best-effort services to the higher layers of the protocol stack, the Application and/or Transport layers need to implement adaptive bandwidth control mechanisms, to setup reliable delay-sensitive media connections. The ultimate target is to optimize end-to-end system performance in terms of average sending-throughput under constraints on the energy-budget available at the PHY, buffer-capacity offered by the DL, and maximum bandwidth supported by the APP/transport layers. This is the focus of this paper, where a client–server networking architecture is considered [2, Chap. 8]. From an application point of view, this model well captures the main system features of the emerging multimedia wireless CDNs [2, Chap. 19, 3], where VBR-encoded media contents are provisioned to the end-user via client–server networking architectures that exploit a single-hop wireless link for the "last-mile" access.

E. Baccarelli (✉), M. Biagi, N. Cordeschi, T. Patriarca, and V. Polli
INFO-COM Dept, "Sapienza" University of Rome, via Eudossiana 18, 00184 Rome, Italy
e-mail: enzobac@infocom.uniromal.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_8,

1.1 The Tackled Problem and Related Works

Specifically, we model the transmit node (the server node) as a time-slotted fluid GI/GI/1 queuing system that is fed by a VBR encoder whose output rate can be controlled. In this operating context, we tackle the *closed-form* (no iterative) *cross-layer* (i.e. queue *and* channel-state aware) design of the controller *jointly* performing the optimal adaptive management of the connection bandwidth, transmit-energy, and throughput under four constraints.

The first one is on the available energy per slot *averaged* over the fading statistics and *conditioned* on the current queue-state and it is imposed by the energy-saving limitations typically arising at the PHY layer [4]. The second is on the allowed peak-energy per slot. It is still dictated by the PHY layer and arises from spectral compatibility issues and/or limitations on the maximum throughput allowed by the considered wireless standard [1, Chap. 6]. The third constraint involves the DL layer and (upper) limits the available buffer-capacity. It guarantees that the considered queueing system is stable and also provides an upper bound on the resulting average queue-delay (see the last part of Sect. 4). The fourth constraint arises from the APP layer and limits the maximum instantaneous bandwidth allowed by the connection. It is typically dictated by the finest granularity level allowed by the adopted encoder (such as the minimum size abiding by the quantization step) [2, Chaps. 5 and 6], and, together with the buffer-capacity, contributes to upper-bound the resulting average queue-delay (see Sect. 4).

According to the traditional "layered" perspective, the classic analysis of the steady-state performance offered in *wired* environments by *layered* congestion-control closed-loop systems, used to exploit various TCP versions (e.g. Reno and Vegas) in tandem with several queue-management policies are detailed in [5–7], while optimized joint energy/queue control policies derived by exploiting the analytical tool of the Markov Decision Process (MDP) and implemented via Dynamic Programming were presented, for example, in [8–12].

About the adopted notation, underlined letters denote vectors, scalar random variable are denoted by bold characters, while their outcomes are indicated by the corresponding no bold symbols. $\mathrm{E}\{.\}$ is the expectation operator, $\mathbb{R}_0^+$ is the set of the nonnegative real numbers, $\mathbb{R}^+$ is the set of the strictly positive real numbers, $\triangleq$ means "equal by definition", while $[x]^+$ indicates $\max\{x, 0\}$. $P(A)$ is the probability of the event A, $p_\sigma(\sigma)$ is the pdf of the r.v. σ, while the notation $\mathrm{E}\{f(\boldsymbol{\sigma}); a, b\} \triangleq \int_a^b f(\sigma)\, p_\sigma(\sigma) d\sigma$ is the expectation of the (scalar) function $f(\boldsymbol{\sigma})$ when the r.v. $\boldsymbol{\sigma}$ is limited to fall into the box $[a, b]$. Finally, $\mathrm{E}_\sigma\{\varphi(\boldsymbol{\sigma}; s)\} \triangleq \int \varphi(\sigma; s)\, p_\sigma(\sigma) d\sigma$ denote the expectation of the bi-argumental function $\varphi(\sigma; s)$ carried out only over the pdf of the r.v. $\boldsymbol{\sigma}$, while $[f(x)]_a^b$ indicates $\max\{a; \min\{f(x); b\}\}$.

2 System Architecture and Problem Setup

By referring to the server–client system architecture [2], we consider a GI/GI/1 fluid queue system fed by a VBR encoder. Time is slotted, the slot-length is unitary and the slot t spans the (semi-open) interval $[t,(t+1)), t \in \mathbb{N}_0^+$. The Information Units (IUs) to be sent over the (single-hop) wireless channel are delivered by the media encoder *at the end* of each slot, and they are buffered into a server's queue of *finite* capacity $N_{\max}$. Thus, we will refer to $\lambda(t) \in \mathbb{R}_0^+$ (IU/slot) as the number of IU arriving at the input of the queue at (the end of) the t-slot, i.e. the *connection bandwidth* at slot t.

The fading phenomena affecting the wireless link are assumed time-invariant over each slot (in accordance to a "block-fading" model [1]) and are considered i.i.d. from slot-to-slot. Thus, the channel-state $\boldsymbol{\sigma}(t) \in \mathbb{R}_0^+$ over slot t is modeled as a real nonnegative r.v.[1] with pdf $p_\sigma(\sigma)$. Furthermore, the channel-state value $\sigma(t)$ is assumed to be perfectly known at the transmitter at the beginning of each t-slot, so that, slot-by-slot, Perfect-Link-State-Information (PLSI) is available at the transmit scheduler. The overall system operates in the *steady-state* (e.g. it works under stationary and ergodic operating conditions). Let $s(t) \in \mathbb{R}_0^+$ be the number of IUs (i.e. the amount of fluid) buffered in the queue *at the beginning* of slot t. Thus, after denoting with $r(t)$(IU/slot) the number of IUs to be sent over the physical channel during the t-slot (the sending state at slot t), the following Lindley's equation:

$$s(t+1) = [s(t) + \lambda(t) - r(t)]^+ \quad , \quad t \geq 0 \tag{1}$$

describes the evolution of the discrete-time queue-length process $\{s(t) \in \mathbb{R}_0^+, t \geq 0\}$. Unless otherwise stated, in the sequel we assume that the queue operates under the stationary regime, so that $\{s(t)\}$ indicates the *stationary and ergodic* solution of the recursion in (1) and $p_s(s)$ is the corresponding *steady-state* pdf.

The cost of sending $r(t)$ units of fluid over slot t is the amount of energy $\mathcal{E}(t)$ (Joule) required for their transmission. Thus, we may assume that the corresponding number IU(t) of IUs sent over the channel in the t-slot depends on *both* $\mathcal{E}(t)$ and the channel-state $\sigma(t)$ via a rate-function $\mathcal{R}(\cdot;\cdot)$ adopted to measure the goodput-performance of the considered system, so that we can write

$$\mathrm{IU}(t) \triangleq \mathcal{R}\left(\mathcal{E}(t);\sigma(t)\right) \quad , \quad t \geq 1 \, . \tag{2}$$

The rate-function $\mathcal{R}(\cdot;\cdot)$ in (2) is a real nonnegative function depending on two nonnegative real arguments $\mathcal{E}(\cdot)$, $\sigma(\cdot)$, and it is measured in IU/slot. Roughly speaking, $\mathcal{R}(\cdot;\cdot)$ summarizes the goodput-performance of the DL/PHY of the considered system, so that its analytical properties and behavior may depend on several

[1] The meaning of the channel-state $\sigma(t)$ is application depending. Without loss of generality and for sake of concreteness, we may consider $\sigma(t)$ be the signal-to-disturbance (i.e. the signal to noise-plus-interference) ratio measured at the input of the client-terminal during slot t.

system parameters, such as the requested QoS, the FEC mechanisms implemented at the PHY layer, the fading statistics, the performance of the ARQ/fragmentation mechanisms utilized at the DL layer, and so on. We will assume that the rate-function $\mathcal{R}(\mathcal{E};\sigma)$ is nondecreasing both for $\mathcal{E} \geq 0$ and $\sigma \geq 0$, and strictly concave in the $\mathcal{E}$-variable.

2.1 Setup of the Optimization Problem

Let us indicate by $\underline{x}(t) \triangleq [\sigma(t), s(t)] \in \left(\mathbb{R}_0^+\right)^2$ the bi-argumental overall state of the system. Thus, our control problem focuses on the optimal design of both the number $r(t)$(IU/slot) of IUs to be sent over the channel at the beginning of slot t (i.e. the optimal design of the *sending-throughput*) *and* the number $\lambda(t)$ of IUs to be outputted by the VBR encoder at the end of slot t (i.e. the optimal design of the *connection bandwidth*). The ultimate target is the *maximization* of the *conditional* average sending throughput $\mathrm{E}\{r(t)|s(t)\}$ under constraints on the available average energy per slot $\mathcal{E}_{\mathrm{ave}}$ (Joule), allowed peak-energy $\mathcal{E}_{\mathrm{p}}$ (Joule), buffer-capacity $N_{\max}$(IU) and maximum connection bandwidth $\lambda_{\max}$ (IU/slot). After writing the energy requested to send $r(t)$ IUs when the channel-state is $\sigma(t)$ as

$$\mathcal{E}(t) \equiv \varepsilon(\sigma(t); r(t)) \triangleq \mathcal{R}^{-1}(\sigma(t); r(t)) \ , \quad t \geq 1 \ , \tag{3}$$

we can state the optimization problem as follows:

$$\max_{r(\cdot),\lambda(\cdot)} \mathrm{E}_{\sigma}\{r(\boldsymbol{\sigma}; s(t))\} \ , \quad s(t) \geq 0 \ , \quad t \geq 1 \ , \tag{4}$$

$$s.t.: \quad \mathrm{E}_{\sigma}\{\varepsilon(\boldsymbol{\sigma}; r(\boldsymbol{\sigma}; s(t)))\} \leq \mathcal{E}_{\mathrm{ave}} \ , \quad s(t) \geq 0 \ , \quad t \geq 1 \ , \tag{5}$$

$$0 \leq \varepsilon(\sigma(t); r(\sigma(t); s(t))) \leq \mathcal{E}_{\mathrm{p}} \ , \quad s(t) \geq 0 \ , \quad \sigma(t) \geq 0 \ , \quad t \geq 1 \ , \tag{6}$$

$$0 \leq s(t+1) \leq N_{\max} \ , \quad t \geq 0 \ , \tag{7}$$

$$0 \leq \lambda(\sigma(t); s(t)) \leq \lambda_{\max} \ , \quad s(t) \geq 0 \ , \quad \sigma(t) \geq 0 \ , \quad t \geq 1 \ . \tag{8}$$

Before proceeding, we stress that (4) points that we maximize (on a per slot-basis) the expected sending-throughput, *given* (i.e. conditioned on) the number $s(t)$ of currently buffered IUs. The number $r(t)$ of IUs to be sent must be searched over the overall set of nonnegative real-valued functions depending on *both* current queue *and* channel states, that is,

$$r(t) \equiv r(\sigma(t); s(t)) \ , \quad t \geq 1 \ . \tag{9}$$

3 The Controller Maximizing the Conditional Average Throughput

Let us indicate with $\varepsilon_r(r;\sigma) \triangleq \partial\varepsilon(r;\sigma)/\partial r$ the first-order derivative of the energy function in (3) carried out with respect to the r-argument. Thus, after recognizing that the optimization problem in (4)–(8) is an instance of *convex* optimization [13], the resulting optimal throughput/bandwidth solution $r^{\text{opt}}(\cdot;\cdot)$, $\lambda^{\text{opt}}(\cdot;\cdot)$ may be evaluated in *closed-form*, as detailed in the following *Proposition 1* (see the Appendix I of [14] for the proof):

Proposition 1. *Under the above reported assumptions, for the optimal controller solution of the constrained optimization problem in (4)–(8) we have that:*

- *The throughput is optimally scheduled according to*

$$r^{\text{opt}}(\sigma(t);s(t)) = \begin{cases} r_{\text{p}}(\sigma(t);s(t)) & s(t) < s_1 \\ \left[\varepsilon_r^{-1}\left(\sigma;\frac{1}{\mu(s(t))}\right)\right]_0^{r_{\text{p}}(\sigma(t),s(t))} & s(t) \geq s_1 \end{cases} \quad \text{(IU/slot)}, \tag{10}$$

where $\varepsilon_r^{-1}(\cdot;\cdot)$ denotes the inverse function of $\varepsilon_r(\cdot;\cdot)$ with respect to the $\mathcal{E}$-variable, while $r_{\text{p}}(\sigma(t);s(t)) \triangleq \min\left\{s(t);\mathcal{R}\left(\sigma(t);\mathcal{E}_{\text{p}}\right)\right\}$ is the peak value of the sending-throughput allowed at t-slot. The threshold s_1 in (10) dictates the boundary among the under-loaded and over-loaded operating regions of the server's buffer, and it may be computed by solving the following algebraic equation:

$$\int_{\sigma} \varepsilon(\sigma;r_{\text{p}}(\sigma;s_1))\ p(\sigma)\ \text{d}\sigma = \mathcal{E}_{\text{ave}}\ . \tag{11}$$

Furthermore, $\mu(s(t))$ in (10) is the optimal value of the dual variable of the tackled optimization problem, and it may be computed by solving the following (functional) equation:

$$\int_{\sigma} \varepsilon\left(\sigma;\left[\varepsilon_r^{-1}\left(\sigma;\frac{1}{\mu(s(t))}\right)\right]_0^{r_{\text{p}}(\sigma;s(t))}\right)\ p(\sigma)\ \text{d}\sigma = \mathcal{E}_{\text{ave}} \quad \text{for } s(t) > s_1\ . \tag{12}$$

- *The optimal bandwidth-management is dictated by the following relationship:*

$$\lambda^{\text{opt}}(\sigma(t),s(t)) \leq \min\left\{(N_{\max} - s(t)) + r^{\text{opt}}(\sigma(t),s(t));\ \lambda_{\max}\right\}. \ \square \tag{13}$$

About the implementation complexity of the optimal controller, we note that $\mu(s(t))$ in (12) and the threshold s_1 in (11) may be computed *off-line* on the basis of the system parameters and the pdf $p_\sigma(\sigma)$ of the channel-state. Thus, the *on-line* implementation of $r^{\text{opt}}(\sigma;s(t))$ in (10) may be accomplished via a simple *three-way memoryless threshold detector* whose input is the system state value (slot-by-slot) $\underline{x} = [\sigma;s]$. This means that the optimal controller we derived may be *actually implemented* without resorting to cumbersome DP-based iterative algorithms.

4 Conditional-vs.-Unconditional Average Throughput-Maximization and Performance Bounds

At first glance, it may be somewhat surprising that (13) imposes *only* an upper-bound on the optimal bandwidth. However, the examination of the proposition's proof in [14] unveils that this arises from two causes. Firstly, the expectation in (4) is conditioned on $s(t)$. Secondly, by assumption (see Sect. 2), arrivals dictated by $\lambda^{\text{opt}}(\sigma(t);s(t))$ occur at the input of the queue at the *end* of slot t, while the corresponding IUs sent over the channel are drained from the output of the queue at the *beginning* of slot t. As a matter of these facts, the value assumed by $\lambda^{\text{opt}}(\sigma(t);s(t))$ can influence *neither* the current throughput $r^{\text{opt}}(\sigma(t);s(t))$ *nor* the current queue-state $s(t)$ (see (1)).

However, the monotonic behavior of the rate-function points out that the conditional average optimal throughput $\overline{\text{IU}}^{\text{opt}}(s)$ does not decrease for increasing queue-state values. Hence, it could be expected that the bandwidth should assume the maximum values allowed by the bound in (13), in order to fill the queue as much as possible and, then, maximize the throughput averaged over *both* channel *and* queue states. By exploiting some basic results of the Theory of Ordered Statistics [15], the following *Proposition 2* formally confirms this expectation.

Proposition 2. *The following bandwidth-management policy:*

$$\lambda^{\text{opt}}(\sigma(t);s(t)) \equiv \min\left\{(N_{\max}-s(t))+r^{\text{opt}}(\sigma(t);s(t));\lambda_{\max}\right\} \tag{14}$$

maximizes the unconditional average throughput $\text{E}\{r(\boldsymbol{\sigma},\boldsymbol{s})\}$ *over the set of bandwidth schedulers meeting the bound in (13).* □

Unless otherwise stated, in the rest of this paper, we refer to the scheduler that jointly allocates throughput and bandwidth according to (10) *and* (14) as the *optimal controller*.

4.1 On the Average Throughput of the Optimal Controller

Let $\overline{r} \triangleq \text{E}\{r^{\text{opt}}(\boldsymbol{\sigma};\mathbf{s})\}$ (IU/slot) be the *unconditional* average sending-throughput generated by the optimal controller (that is, the optimal throughput averaged over *both* channel *and* queue-state statistics). The following *Proposition 3* points out its limit values and (statistical) behavior.

Proposition 3. *Let the client–server system work under the optimal controller policy. Thus,*

(i) For $\lambda_{\max} > N_{\max}$*, the queue-state identically equates* $N_{\max}$*, that is,* $s(t+1) = N_{\max}$ *for any* t*, and we have* $\overline{r}^{\text{opt}} = \overline{\text{IU}}^{\text{opt}}(N_{\max})$.

Furthermore, in this case the optimal controller in (10), (14) also maximizes the unconditional average throughput over the full set of schedulers meeting the constraints in (5)–(8).

(ii) For $\lambda_{\max} \leq N_{\max}$*, the steady-state pdf of the queue-state identically vanishes out of the open interval* $(\lambda_{\max}, N_{\max})$*, i.e.* $p_s(s) \equiv 0$ *for any* $s \notin (\lambda_{\max}, N_{\max})$*, and* $\mathrm{IU}^{\mathrm{opt}}(\lambda_{\max}) \leq \overline{r}^{\mathrm{opt}} \leq \mathrm{IU}^{\mathrm{opt}}(N_{\max})$. □

In order to optimize the throughput performance of the system, the joined optimal schedulers act so as to hold the buffer as full as possible. Hence, a proper choice for the buffer-size $N_{\max}$ is important in the (application-depending) trade-off between system throughput and system delay.

5 Performance Tests and Comparisons

In this section, we test the actual performance and robustness properties of the optimal controller of (10), (14) in terms of unconditional average bandwidth, queue length $\overline{s}^{\mathrm{opt}}$ and queue-delay $T_{\mathrm{q}}^{\mathrm{opt}}$. In the carried out numerical tests, the channel-state $\boldsymbol{\sigma}(\cdot)$ plays the role of instantaneous fading-affected Signal-to-Noise Ratio (SNR) measured (on a per-slot basis) at the client side, when the energy radiated by the server node over a slot time is unitary. We modeled $\boldsymbol{\sigma}(\cdot)$ as a central χ-squared r.v. with 4-degrees of freedom, i.e. $p_\sigma(\sigma) = \sigma \exp(-\sigma)$, for $\sigma \geq 0$ [16], while we chose a logarithmic rate-function $\mathcal{R}(\sigma; \mathcal{E})$ (commonly employed to measure the so-called Shannon Capacity of transmission links impaired by Gaussian-distributed additive noise-plus-interference disturbances [16]) to numerically evaluate the throughput performance: $\mathcal{R}(\sigma; \mathcal{E}) \equiv \log(1 + \sigma\mathcal{E})$, (IU/slot).

As already remarked in Sect. 1, at the best of the Authors' knowledge, neither "cross-layer" nor "layered" adaptive controlling architectures have been currently proposed in the open literature for the optimal (or even sub-optimal) allocation of connection bandwidth, sending-throughput and transmit-energy *under all* the constraints we considered in (5)–(8). So, since planning meaningful and fair performance comparisons proved to be troublesome, we focus on the unconditional average bandwidth and queue length performance of the optimal controller and give insight on the resulting trade-off. The plots in Figs. 1 and 2 report the behavior of $\overline{\lambda}^{\mathrm{opt}}$ and $\overline{\mathrm{T}}_q^{\mathrm{opt}}$ vs. the available average energy $\mathcal{E}_{\mathrm{ave}}$ and show the effects of the buffer capacity $N_{\max}$ on the average performance of the optimal controller. Specifically, in Fig. 1 when we consider $\lambda_{\max} = 2.5$ (IU/slot) and under the assumption of increasing buffer length's values $N_{\max}$ ranging from 3 to 100 IUs, we experience a flat zone beyond $\mathcal{E}_{\mathrm{ave}} > 10$ (Joule).

On the contrary when $\mathcal{E}_{\mathrm{ave}} < 10$ (Joule), we can recognize the impact of different values of $N_{\max}$ since for low values (e.g. 3, 7) we experience very low average throughput with respect to the case of high $N_{\max}$ values (e.g. 10, 100). Thus, we can ascertain that high $\mathcal{E}_{\mathrm{ave}}$ values do not require a heavy effort in terms of buffer length. On the other hand, if we pursue the way of considering high values of $N_{\max}$, then we verify high queue occupancy. This reflects (in Fig. 2) on the average

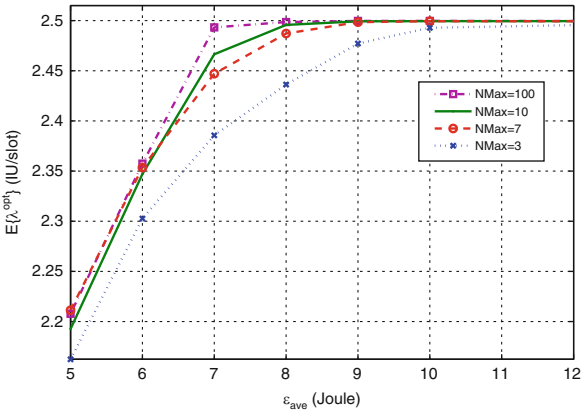
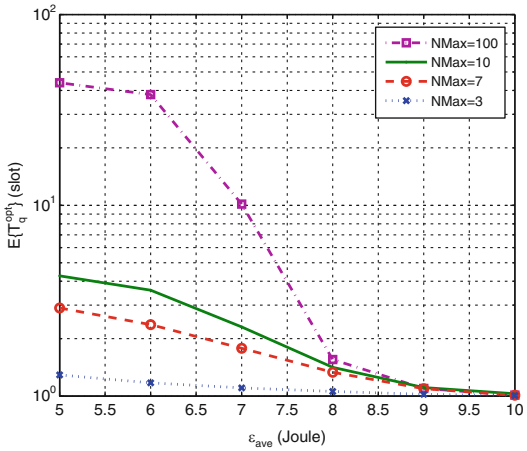


Fig. 1 Average bandwidth-vs.-average energy behavior of the optimal controller for $\mathcal{E}_{\mathrm{p}} = \infty$ (Joule) and $\lambda_{\max} = 2.5$ (IU/slot)

Fig. 2 Average queue-delay-vs.-average energy behavior of the optimal controller for $\mathcal{E}_{\mathrm{p}} = \infty$ (Joule) and $\lambda_{\max} = 2.5$ (IU/slot)



queue-delay that considerably increases when $N_{\max}$ is high. So a trade-off between average queue-delay and throughput should be found by traveling on a parallel track by paying attention to the available energy. Finally, this analysis allows to choice the buffer length when all constraints are assigned so to satisfy the user with a minimum end-to-end delay. Finally, Figs. 3 and 4 show the unconditional average bandwidth and queue-delay vs. the maximum connection bandwidth $\lambda_{\max}$, for ρ parameter values ranging from 0.8 to 1, being $(\rho/(1-\rho))$ the normalized SNR affecting the channel estimate $\{\widetilde{\sigma}(t)\}$. An examination of these figures confirms that even for values of the $(\rho/(1-\rho))$ ratio as low as 5–6, the performance loss remains, limited up to about 15% of the corresponding error-free case.

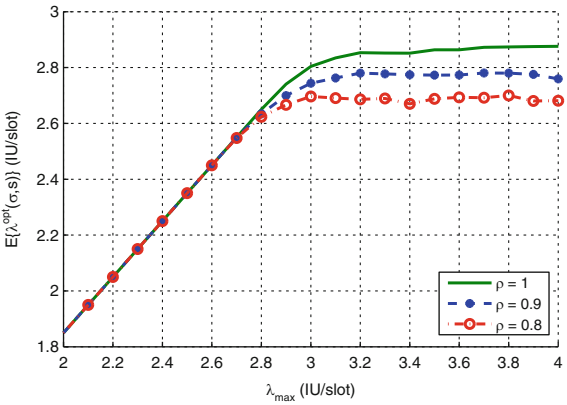


Fig. 3 Average bandwidth-vs.-maximum allowed bandwidth for some values of the channel-state estimation-error parameter ρ. $\mathcal{E}_{\mathrm{p}} = 21$ (Joule), $\mathcal{E}_{\mathrm{ave}} = 10$ (Joule) and $N_{\max} = 100$ (IU) are considered

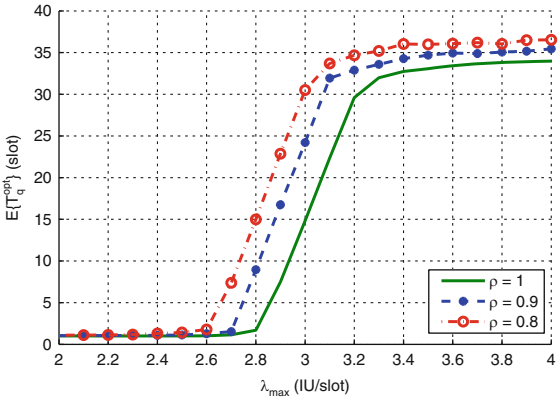


Fig. 4 Average queue-delay-vs.-maximum allowed bandwidth for some values of ρ parameter. $\mathcal{E}_{\mathrm{p}} = 21$ (Joule), $\mathcal{E}_{\mathrm{ave}} = 10$ (Joule) and $N_{\max} = 100$ (IU) are considered

References

1. Kurose JK, Ross KW (2007) Computer networking – a top down approach featuring the internet, 4th edn. Addison Wesley, Reading, MA
2. Van Der Schaar M, Chou PA (2007) Multimedia over IP and wireless networks. Academic, New York
3. Yates RD, Mandayam NB (2000) Challenges in low-cost wireless data transmission. IEEE Signal Process Mag 4(3):99–102
4. Schaar MVD, Shankar DS (2005) Cross-layer wireless multimedia transmission: challenges, principles and new paradigms. IEEE Wireless Commun 12(4):50–58
5. Padhye J, Firoiou V, Towsley DF, Kurose JF (2000) Modeling TCP Reno performance: a simple model and its empirical validation. IEEE/ACM Trans Netw 8(2):133–145
6. Fall K, Floyd S (1996) Simulation-based comparisons of Tahoe, Reno and SAKC TCP. Comput Commun Rev 26(3):5–21
7. Low SH (2003) A duality model of TCP and queue managment algorithms. IEEE/ACM Trans Netw 11(4):525–536
8. Berry RA, Gallager RG (2002) Communication over fading channels with delay constraints. IEEE Trans Inf Theory 48(5):1135–1149

9. Goyal M, Kumar A, Sharma V (2003) Power constrained and delay optimal policies for scheduling transmission over a fading channel. In: Proceedings of IEEE INFOCOM 2003, San Francisco, Mar 30–Apr 3, 2003
10. Rajan D, Sabharwal A, Aazhang B (2004) Delay-bounded packet scheduling of bursty traffic over wireless channels. IEEE Trans Inf Theory 50(1):125–144
11. Prabhakar B, Biyikoglu EV, Gamal AE (2001) Energy-efficient transmission over a wireless link via lazy packet scheduling. In: Proceedings of IEEE INFOCOM 2001
12. Shakkottai S, Srikant R (2002) Scheduling real-time traffic with deadlines over a wireless channel. Wireless Netw 8(1):13–26
13. Collins B, Cruz R (1999) Transmission policies for time-varying channels with average delay constraints. In: Proceedings of 1999 Allerton conference on communication, control and computing, Monticello, IL, Oct 1999, pp 1–9
14. Cordeschi N (2008) Adaptive QoS transport of multimedia over wireless connections – a cross layer approach based on calculus of variations. PhD Thesis, http://infocom.uniroma1.it/∼cordeschi/phdthesis.pdf
15. Baccelli F, Bremaud P (2003) Elements of queueing theory, 2nd edn. Springer, Berlin
16. Tse D, Viswanath P (2006) Fundamentals of wireless communications. Cambridge University Press, Cambridge, UK

A Secure MPLS VPN Infrastructure for Complex Geodata Sensor Network

Mirko Luca Lobina and Tatiana Onali

1 Introduction

A complex geodata sensor network (i.e., GSN) is aimed at providing the forecasts of geo-conditions. A complex GSN is the combination of several independent sensor networks (i.e., SN), each of which geographically monitors different environmental factors by means of specific sensors and sends the information to a central data-storage system for collection and further analysis purposes (Fig. 1).

The overall goodness of the forecasting provided by a complex GSN relies on the observational accuracy of the SN and the statistical and mathematical core models. Limiting our discussion to the first aspect, more data are available from different SN and different points of times, and more accurate are the results. To this aim, each independent SN should ideally perform the communication of geo-conditions to the same complex GSN together with hundreds/thousands of other distributed SN in almost real-time fashion.

With respect to this ideal condition, the real world presents three common practices: few local and independent sensors; few SN connected to a complex GSN by means of leased single lines; and several SN connected to a complex GSN by means of large backbone network (owned by a service provider), which offers different regional services. The first practice is not applicable to complex GSN and beyond the scope of this chapter. The second solution can be stable and technically feasible in geographically restrained contexts, but it is not convenient, for economic reasons, in wide scenarios. The third perspective is attractive, but only under certain technical constraints regarding implementation and security. This chapter focuses on the facets of the last approach proposing a Multi Protocol Label Switching Virtual Private Network (i.e., MPLS VPN) infrastructure as a steady technology and widespread service enabler.

This chapter is organized as follows: Section 1 provides an essential background on MPLS VPN technology and related features. Section 2 proposes a

M.L. Lobina (✉) and T. Onali
Department of Electrical and Electronic Engineering, University of Cagliari, Italy
e-mail: m.lobina@diee.unica.it; tonali@diee.unica.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_9,

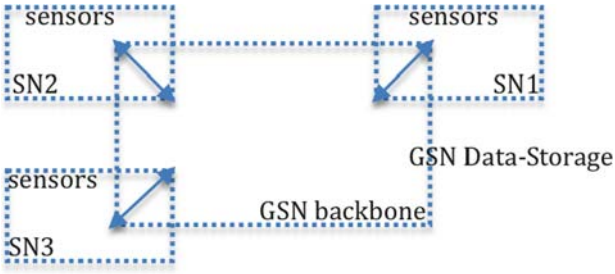


Fig. 1 Sketch of GSN

possible MPLS VPN implementation for a complex GSN. Section 3 addresses the security aspects of the proposed implementation. Section 4 concludes the chapter.

2 MPLS VPN Infrastructure

MPLS [1] is a label switching mechanism. The label, a short and fixed-length data segment, is inserted in the packet header and used to forward the traffic to the right direction in the MPLS network in a complete path (called LSP, Label-Switched Path) between two edge routers (called LER, Label Edge Router). The forwarding is performed by the internal nodes (called LSR, Label-Switched Router) of the MPLS network only reading the label and without examining the packet header itself. This examination is performed at a control-path level, leaving to the data-path only a forwarding task. This is the reason why the MPLS is a real fast switching approach.

MPLS presents several major benefits and features. First of all, MPLS is a straightforward switching technology with only two basic distinct components: a signaling protocol, used to setup a LSP, and a data plane to forward the packets based on the labels (a LSR, receiving a packet, executes two operations: removes the old label and inserts on its own; forward the packet to the right egress interface). MPLS is independent from a particular Data Link Layer technology: the two above cited components could be implemented in different technologies (IP, ATM, or FR). MPLS supports several label distribution and resource reservation protocols specifically designed to sustain traffic engineering (e.g., usually referred to as MPLS TE and defined as the process of steering traffic across the backbone to facilitate efficient use of available bandwidth between a pair of routers [1]). MPLS can easily employ multidimensional classification (i.e., many different types of Forwarding Equivalent Classes (FEC)). In this way, all the packets belonging to the same FEC receive an identical treatment in the forwarding process.

The last feature of MPLS, for some aspects related to the multidimensional classification, is the ability to build connectionless VPN, here referred to as MPLS VPN. The term “connectionless” means that a VPN user is connected to the MPLS service as a “cloud,” which ensures that packets are forwarded correctly to the other

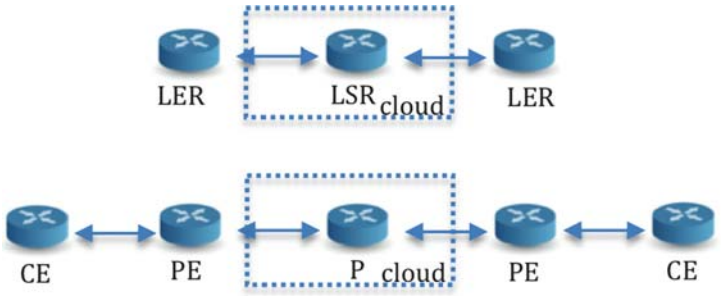


Fig. 2 Sketch of MPLS (up) and MPLS VPN (down)

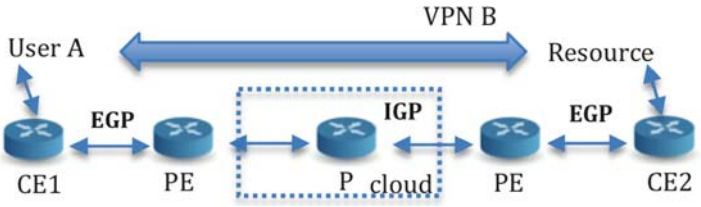


Fig. 3 User A and "the resource" connected via VPN B

VPN user site. The nomenclature and definitions used in a MPLS VPN framework are introduced in [2, 3]. Figure 2 sketches such a context underlying the difference with pure MPLS. A MPLS VPN network consists of three main components: PE (Provider Edge Router), P (Provider Router), and CE (Customer Edge Router).

The first two elements belong to the MPLS core while the last is typically located at a customer site. The role of these routers is described in the following with a generic user A trying to access to a resource via VPN B (see Fig. 3).

User A and the resource are published respectively by CE1 and CE2. First of all, under the MPLS layer it is necessary that all the playing routers have enabled appropriate routing protocols to provide basic route publishing and reciprocal knowledge. The link CE-PE can be served in principle by any external gateway protocol (EGP), both static and dynamic, while the MPLS core can be provided with an internal gateway protocol (IGP). As an example, we can consider a RIP for link CE1-PE and PE-CE2, and OSPF for the MPLS core (PE and P routers). When user A wants access to a certain resource, the PE must be able to distinguish the right VPN from the resource, and user A belongs to and performs the forwarding. To this aim the PE employs VPN routing/forwarding instances (VRF). Each VRF is connected with one or more interfaces to the one or more CE, all belonging to the same VPN. In our example, we suppose PE has only one interface to the CE1 and only one VRF A, corresponding to VPN B (this configuration is specular for PE-CE2). When a route comes from CE1, the PE is able to distinguish VRF Aimmediately. Once this

operation is performed, it remains the forwarding to the other PE inside the core. A route distinguisher (RD) of 6 bytes is associated with the VRF and prepended to every VPN route received from the CE. Such addressing scheme is called VPN-IPv4 (or VPN-IPv6) and distributed between the PEs by means of the Multi-Protocol Border Gateway Protocol (MP-BGP). The PE, receiving, for example, IPv4 and producing VPN-v4 routes, creates the ideal ground for pure forwarding inside the MPLS core. Obviously, these basic operations are repeated in the opposite order at the egress side of PE-CE2-resource.

3 Proposed Infrastructure for Complex GSN

Several works that exist present aspects and features related to SN aimed to forecast geo and environmental conditions (as an example: [4–6] and [7]). In this paper a generic SN is presented (see Fig. 4), as composed by a minimal set of wireless-wired[1] sensors for a 1 km^2 of monitored heterogeneous ground. The details about such sensors are presented in Table 1.

Every SN is connected to the MPLS core by the chain CE*SN*-PE, running a generic EGP. The connection CE*SN*-PE must be provided, at a minimum, by a fast Ethernet link. Such kind of connection is enough for covering the SN bandwidth necessities.

From the PE, the data are transferred through the MPLS core (running IGP) to another PE and then to an operation center (i.e., OC) or a CE*Storage* (these last links run EGP). Inside the core, a Gigabit connection is always supposed (see Figs. 5 and 6).

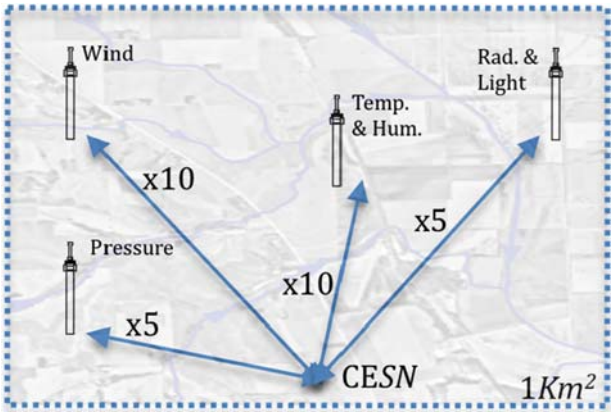


Fig. 4 The sensor network

[1] Pure Ethernet or RS232-to-Ethernet.

Table 1 Details about sensors

Monitored factor	Typology of connection (wireless/wired)	Precision per sample (bit)	Number per network (1 km^2)	Sampling time (/hour)
Wind	Wireless/wired	8 bit	10	4
Temperature and humidity	Wireless/wired	7 bit	10	4
Radiation and light	Wireless/wired	7 bit	5	2
Air pressure	Wireless/wired	7 bit	5	4

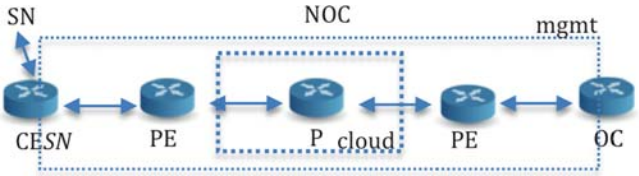


Fig. 5 SN and OC

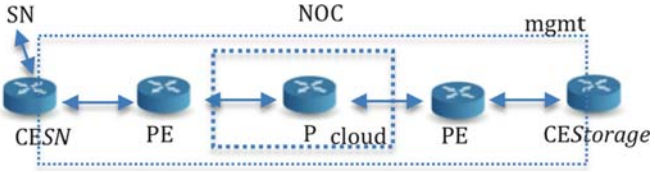


Fig. 6 SN and CE*Storage*

The proposed strategy employs one ad-hoc VRF for every monitored environmental factor in Table 1. This configuration is replicated in every PE in a way that every factor is distinct from the source (the sensor) to destination (the storage). As an example, in the following the detail of configuration for a VRF-wind is proposed for a generic PE, here called "router1" (see Listing 1 [8]).

Listing 1

```
Listing 1
hostname router1
ip cef
ip vrf wind
   rd 1000:1
! 1000 refers to BGP instance
  route-target both 1000:1
ip vrf forwarding wind
  ip address XXX SUBNET
! XXX & SUBNET refers to a specific
! interface of router!
```

Another VRF-mgmt is established to monitor the in-band CE*SN* and if it is always connected to the CE*Storage* (see Fig. 6). In the following the detail of configuration for a VRF-mgmt is also proposed for router1 (see Listing 2 [8]).

Listing 2

```
hostname router1
ip cef
ip vrf wind
   rd 1000:2
! 1000 refers to BGP instance
  route-target both 1000:1
ip vrf forwarding wind
  ip address XXY SUBNET
! XXX & SUBNET refers to a specific
! interface of router!
```

The MPLS infrastructure has no direct connection to Internet or other external sites, but the MPLS core obviously could have through other sites or VRF.

4 Securing MPLS VPN Infrastructure

As introduced in [1], security is composed at least by three key factors: architecture, implementation, and operation. Architecture is the formal specification of the applied technology. Implementation refers to how the architecture is currently implemented. Here we can find common programming mistakes. Finally, operation includes operation issues (as introducing weak passwords or losing/sharing digital certificates or signatures). There are no 100% secure systems, however, starting from a good knowledge of the three cited factors for a certain technology, it is possible to fix as much threats as required by the given SLAs.

The ideas beyond the proposed architecture are two: guaranteeing separation between monitored environmental factors using distinct VRF; creating an island isolated in terms of traffic from external sites. Our analysis is strongly focused on what was introduced in the last section, deepening key weaknesses and proposing countermeasures. Here the threats are basically classifiable as intrusion and denial-of-service (DoS), each against VPN, MPLS core, Internet or Extranet site. In this work we refer directly to the first two attacks, underlying that an attack against the MPLS core is also against a VPN. Note the distinction: intrusions allow full access to internal data, whereas DoS does not give such access, but prevents access for all users. Thus, DoS against the core can affect indirectly a VPN.

The intrusion and DoS vectors in a VPN could come from another VPN, the MPLS core, or the Internet. Keeping in mind that the principles of separation against outside networks (including Internet in our model) and the control of inbound traffic into a VPN prevail as well, we can restrict our analysis only to the PE router. Such a device really seems to be the weak link in the chain, with specific focus on the PE interfaces that connect the CEs on a VPN. Note that by default a CE,

always considered untrusted (apart from the case where it is directly managed by the service provider), can see only its belonging interface and cannot ping other PE interfaces. Obviously, if a PE is jeopardized, via intrusion or DoS, this is not guaranteed. Starting from a spoofed customer's VPN identity, the external PE interface can be attacked with the insertion of:

- Fake IP packet – The PE has a VRF configured on the interface CE-PE and from this receives an IP packet for a destination that is not comprised in the VRF. The PE simply discards it, in fact there are no routing entries for other destination included in the VRF.
- Fake MPLS packet – The operative situation is same as previous, but in this case a fake MPLS frame is injected. The router simply discards the packet, as it knows that no MPLS-labeled packets would arrive from that interface.
- Modified IP packet – Same situation as previous, but with the PE that has a static IP route to a certain site (not served by MPLS). In this case the router could analyze the traffic via software, and this can lead to a possible flooding of the PE.

The conclusion is that the customer, who is malicious, can see only the internal resource of the VPN he belongs to. Furthermore, only the third attack, among the above cited, can be considered as really dangerous. In all cases, the resources published in the VPN are visible and potentially weak at a service level.

Another attack attempting can be performed via password guessing to Telnet/SSH sessions during maintenance sessions of the PE. Two constraints must be underlined: first of all a maintainer connects to the PE internally with respect to the MPLS core (thus, it is not possible a man in the middle (MiM) scenario); secondly, this attack is really risky only in case of weak passwords.

The last possible attack scenario is an identity spoofing in an External site. This situation is similar to the first (spoofing of a customer's VPN identity).

Note, for all the cases, that the core is safe as it is configured with another class of IP addresses (the internal interface of the PE and other intra-core router are not visible) than the VPN.

5 Conclusions

This paper has presented a possible MPLS VPN implementation for a complex GSN, deepening the aspects related to the links specifications, security and traffic separation features.

The proposed solution adopts one VRF for each single monitored environmental factor. In this optic, two sample VRF configurations has been also proposed together with a schema of the network in terms of links and adopted routing protocols.

The last section faced in detail some security aspects of the proposed implementation. The conclusion is that the proposed realization is secure both for the MPLS core and the VPN. The only Achilles' heel is in the external interface of the PE router, but this weakness can also be solved by working on the basic global configuration of the router itself.

References

1. Behringer MH, Morrow MJ (2005) MPLS VPN security. Cisco Press
2. RFC3031 (2001) Multiprotocol label switching architecture. http://www.ietf.org/rfc/rfc3031.txt
3. RFC 2547bis BGP/MPLS IP (2001) VPNs. www.ietf.org/proceedings/04mar/I-D/draft-ietf-l3vpn-rfc2547bis-01.txt
4. Barrick John DW, Ritter John A, Watson Catherine E, Wynkoop Mark W, Quinn John K (1996) Calibration of NASA turbulent air motion measurement system. Technical Report: NASA-96-tp3610, NASA Langley Technical Report Server
5. Cerpa A, Elson J et al (2001) Habitat monitoring: application driver for wireless communications technology. In: 2001 ACM SIGCOMM workshop on data communications in Latin America and the Caribbean, ACM Press, New York, NY, pp 20–41
6. Mainwaring A, Culler D et al (2002) Wireless sensor networks for habitat monitoring. In: Proceedings of the 1st ACM international workshop on wireless sensor networks and applications, pp 88–97
7. ITU-T (2008) Ubiquitous Sensor Networks (USN)-ITU-T Technology Watch Briefing Report Series, No. 4. - www.itu.int/dms_pub/itu-t/oth/23/01/T230100000040001PDFE.pdf
8. Lakshman U, Lobo L (2005) MPLS configuration on Cisco IOS software. Cisco Press

Part II
The Role of the Middleware

Making the Internet of Things a Reality: The WhereX Solution

Antonio Puliafito, Angelo Cucinotta, Antonino Longo Minnolo, and Angelo Zaia

1 Introduction

During the last decade, technology was mainly dominated by solutions targeted to improve communication among people (personal computers, mobile phones, PDA and social networks). The trend is now slightly changing in the sense that information exchange is now requested also with objects, animals and plants, thus creating a more complex reference scenario generally known as the Internet of things [1], i.e. the paradigm that embodies the data interchange between objects.

This technology shift is taking place in several areas, making the need of information more evident "anytime" and "anywhere". The ability to respond to such needs, and hence to generate business, depends very much on the timely identification and location of objects and on the prompt management of the events they generate.

RFID technology, and its logical extensions, such as Sensor Network, GPS and WiFi, are well fitted to identify people and objects without any human intervention and to understand their specific functioning state [2, 3]. A standardised system for objects identification is needed, and *EPC* (*Electronic Product Code*) is moving towards this direction [4, 5]. In addition, a system to foster technology integration with a seamless access directly from the application level would be very beneficial to speed up the development process. We believe that the development of an RFID middleware to address these challenging items is a unique possibility to successfully ensure data interchange and technology integration [6].

This solution is known by the term *middleware.* This is a software layer interposed between the technological and the application sub-levels. Its functions are mentioned above.

A. Puliafito (✉), A. Cucinotta, and A.L. Minnolo
University of Messina, Engineering Faculty, Contrada Papardo, S. Sperone, 98166 Messina, Italy
e-mail: apuliafito@unime.it; a.cucinotta@unime.it; antonino.longo@unime.it

A. Zaia
Inquadro s.r.l., Via Nino Bixio, 77 – 98123 Messina, Italy
azaia@inquadro.com

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_10,

In this paper, we are going to describe the work done for the creation of an "event based" middleware for the automatic identification of different objects in different contexts. We briefly describe the technologies used. However, we wish to outline that the middleware is able to implement an effective abstraction of the technological level, adapting the logic of management of each device.

The *RFID* (*Radio Frequency Identification*) technology is becoming more and more widespread, thanks to the evident advantages that may come from its use. In fact, the Radio Frequency Identification technique allows to identify objects without a visual contact. This improves the effectiveness of company systems and processes in several areas such as logistics, asset management, the identification of patients in hospitals [7], and many other areas.

The growth trend observed for the RFID technology has also been observed for *Sensor Networks* [8]. This is largely due to their intrinsic characteristics of independence and cooperation among nodes and to the wireless communication system that provides a high potential, even without the existence of previous communication infrastructures. Thanks to the ability of nodes to cooperate with each other and to make the self-configuration of the network, the data detected are correctly routed, even if communication infrastructures are totally absent. Several areas of application are possible, from military to the monitoring of environment, precision agriculture, intelligent buildings, and many other areas. S*ensor networks* typically consist of a considerable number of sensors with minimum processing and communication abilities, which are connected to actuators, if necessary.

GPS (*Global Positioning System*) was developed by the USA Department of Defence that still controls it, even if the use of the system is free. The needs were to provide the GPS user (whether still or moving) with very accurate information about his/her 3-D position, speed and time, anywhere on earth or near the earth.

Besides the technologies mentioned above, it can be useful to use WiFi devices to do monitoring and location work even in mobility. In particular, by using a system of triangulation and through the assessment of the *Received Signal Strength Indication* (RSSI), we can make an estimate of the position.

Some recent research has outlined how the implementation of a middleware that can integrate some of the technologies mentioned above[9, 10], or other technologies even in mobility [6, 11], leads to benefits arising from a better use of each potential.

It is therefore evident that the availability of a middleware is fundamental for the connection between the different technologies of identification and location (RFID, Sensor Network, GPS, WiFi, etc.) [12] and the application layer. However, the matter is not limited to technological integration. In fact, the availability of powerful and reliable mechanisms is crucial for the management of big amounts of information received by the devices, to clean the data affected by errors, and to filter them according to precise rules. Last but not least, it becomes essential to make data available (dispatching) to company applications (CRM, ERP, Data Warehouse, etc.) in a simple and reliable way [13–15], by favouring the integration towards the application level.

The middleware layer interposes between the technological and the application sub-levels. This way, the middleware layer must assure flexibility and scalability and favour the processes of integration by reducing costs and improving the effectiveness of business processes. Such need is so strong, that some recent researches mention the measurement of the index of customer satisfaction [16]. However, it is evident that the creation of a middleware is not a painless process. Two fundamental issues must be solved: the continuous technological evolution (RFID, sensors, etc.) on one hand and the integration with company IT systems on the other hand. In order to solve these two open issues, the middleware must have some peculiarities. The main characteristic is to be "*event based*" [17, 18]. This assures an effective process of technological abstraction, switching from a vision linked to the specific issues of interfacing of devices (tags/RFID reader, ZigBee, Sensor Network, GPS, WiFi, etc.) to the management of the *event* generated by each device. This approach allows to integrate the logic of the management of different devices.

Furthermore, the middleware must include a rule engine that can detect whether a specific situation occurs and can respond appropriately. All the events generated by the devices must be filtered and cleaned as well.

Finally, a new-concept middleware, which can provide a prompt response to technological evolution, must be equipped with a very flexible and scalable software architecture. In order to need such requirements, SOA architecture solutions and XMPP [21] communication models must be used. Enterprise systems are shifting from a monolithic applications to an ecosystem of services organised according to a *Service Oriented Architecture* (*SOA*)[19, 20, 22].

The purpose of this paper is to show how the solutions mentioned above and the outlined technological trends must be taken in due consideration when a middleware is designed for the management of traceability and location in high-mobility environments. The core of this middleware must respond to the following state-of-the-art technologies: *SOA (WS, REST, and HTTP), Enterprise Service Bus (XMPP, JMS), and Multichannel Communication (Jabber, VoIP, SMS, and MMS)*.

In order to give the highest level of flexibility and scalability to the system, we need to make a wide use of SOA principles in the organisation of both the external integration processes (towards enterprise applications) [23] and the exchange processes within the middleware, by favouring the addition/modification of functions and services (scalability). As we appropriately point out below, the use of the XMPP communication model is very important because, since this is a real-time communication protocol, it can provide a high level of interactivity, and can simplify the integration with enterprise applications. Saying that a middleware must simplify the integration processes means to favour the exchange of information. For this purpose, two different solutions can be used. In the first case, data are dispatched according to the SOA philosophy: the exchange of data is made according to a specific request that the system responds to. In the second solution, the XMPP communication protocol is used. It makes the communication system independent, by turning the middleware into a push server that notifies events. This aspect paves the way for the evolution of the system towards the Web. In fact, the purpose is to equip the middleware with an appropriate component that is integrated in static HTML pages,

and enables them to update their contents automatically as soon as an event occurs (e.g., push data to a browser by protocol XMPP).

All these characteristics turn out to be flexibile. In fact, according to what we say in this paper, very different issues (from *document management, to supply chain,* to *safety in work environments, to system maintenance*, *etc.*) can be solved at the same time through a single-structured middleware [24]. This also allows to reduce costs because the middleware simplifies the processes of integration with the enterprise IT systems. Furthermore, the developer no longer needs to focus on the specific issues related to the technology (RFID, Sensor Network, ZigBee, etc.), but he/she can focus his/her attention on the final application. The processing ability of such a middleware authorises us to talk about *business intelligence* because this solution can exchange data with external applications, analyse the data at the same time, in order to point out any margin of improvement in the effectiveness of the systems.

Our work has therefore appeared as the creation of a finished product: the WhereX® middleware.

The description of a real use case is useful to understand how the Internet of things is the direction to go to guarantee interoperability and effective interaction between people and objects. In this sense, ITU has already identified RFID as the technology to be used to connect objects. WhereArt® is an application based on the WhereX® RFID middleware that can enhance the interaction mechanisms to offer services based on the concepts of user mobility and identity, in areas such as museums, art galleries, parks, cities, etc.

Our paper is organised as follows. In sect. 2 we describe the architecture of the designed middleware. We describe its general organisation, as well as the technical choices made for the creation of each component. In the third section we describe the WhereX® middleware, by also showing some screenshots about specific operating functions. The fourth section contains the description of some real cases that can help to understand how this middleware can be used by one or more companies in totally different contexts. The difference of each application described outlines the high level of flexibility achieved. Finally, we show the conclusions in the last section.

2 Design Principles of an RFID Middleware

The role of middleware is to offer the best solution to easily deal with the integration. A good definition of middleware could be “a layer of software that resides between the business application and the devices layer of heterogeneous platforms and protocols”. It decouples the business applications from any dependencies on the plumbing layer, which consists of heterogeneous operating systems, hardware platforms and communication protocols [25]. In other words, the presence of a middleware to build distributed systems relieves developers from the implementation of low-level details related to the network, such as concurrency control, transaction management and network communication, so that they can focus on application

requirements. Some examples of middleware successfully used are: CORBA from the Object Management Group (OMG) [26], Microsoft COM [27], SUN Java RMI [28], IBM's MQSeries [29], and Remote Procedure Calls (RPCs) introduced by SUN in '80s.

The effort taken in the design of a middleware layer goes towards technological generalisation anyway.

2.1 Architecture Description

As shown in Fig. 1(a), the middleware we intend to design should specifically fit the management of RFID-based infrastructures (RFID technology, Sensor Net-works, GPS, Bluetooth, WiFi, etc.).

It introduces significant perspectives to enterprise system integration, by extending the interoperability concept, by promoting new forms of collaboration and by a careful management of issues such as context awareness, mobility management and seamless connectivity. Its final goal is to provide a solution compliant with the use of Web services and able to provide multichannel communications.

The purposes of the middleware can be summarised in simultaneous management of many objects and readers, even with different technologies; management of many sensors placed on wide surfaces, not covered by previous infrastructures; high scalability to meet the requirements of modern and dynamic companies and on- and off-line interfacing with mobile devices.

The architecture has been created by paying attention to functional modularity, and for this reason the use of the SOA philosophy has been fundamental also for

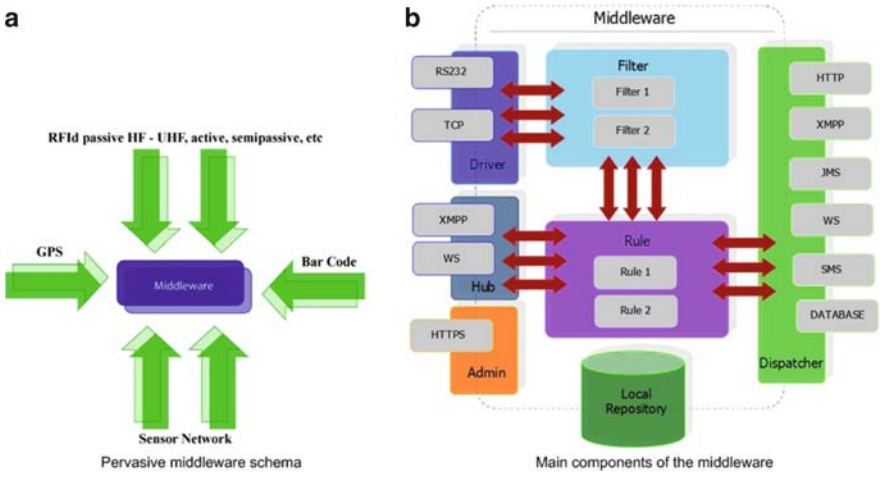


Fig. 1 Middleware architecture (**a**) Pervasive middleware schema (**b**) Main components of the middleware

the internal processes of the middleware. This gives a high level of flexibility and scalability to the system. As shown in Fig. 1(b), these are the main functional components of the architecture:

- *Data Collection (Drivers and Hub)*. They are two distinct components that implement two solutions (*Direct* and *Logical entities*), both aimed at data acquisition from physical devices. In the *direct solution* the physical devices are directly interfaced/connected to the middleware through different interfaces (Ethernet, RS232, Bluetooth, etc.). In the *logical entities* solution the data reach the middleware through other software modules that can generate events.
- *Data Filtering (Filter)*. It is the component that filters data, removes the data affected by errors, removes redundant data, and aggregates information.
- *Events Management (Rule)*. Its task is to generate customisable events, depending on specific situations. This way, the aspect related to the event management can be separated from the specific technical issues related to the physical device that generates the event.
- *Data Dispatching (Dispatcher)*. It dispatches data to and from the components of the system, and to the applications outside the system. It consists of sending *reconstructed information* to the different applications interacting with middleware. The most recent communication protocols that allow middleware to be integrated in enterprise systems are HTTP, XMPP, JMS, WS, SMS and Database. They can even be simultaneously adopted, thus providing a powerful multichannel communication feature.
- *Data Storage (Local Repository)*. It plays two important roles. The first role is to store the configuration parameters of the system. The second role is to keep the history of the data detected to perform statistic processing or to create business intelligence.
- *Middleware Configuration (Admin)*. It provides the interaction tools for the system administration by a console made through a user-friendly web interface.

3 The WhereX® Middleware

In this section, we present WhereX®, an implementation of the middleware described in the previous section. WhereX® is a mobile middleware being used in enterprise systems, produced in Wireless RFID Laboratory of University of Messina and commercialised by Inquadro S.r.l., a spin-off company of the University of Messina. WhereX® is the middleware designed for the management of infrastructures for automatic identification based on the modern RFID (Radio Frequency Identification) technology, Sensor Networks, GPS (Global Positioning System), Bluetooth and WiFi. WhereX® introduces significant perspectives to enterprise system integration, by extending the interoperability concept and by promoting new forms of collaboration and a careful management of issues such as context awareness, mobility management, and seamless connectivity. WhereX® is based on a J2EE architecture, is integrated with Oracle Application Server 10g and can adapt

its behaviour according to the customer's needs. It is possible because there is a Rule Component that integrates the Rhino engine. *Rhino is an open-source implementation of JavaScript, written entirely in Java. It is typically embedded into Java applications to provide scripting to end users*. The benefit is to enable the user to interact with the system. By using the potential of Java, the user can enter new rules and create different workflows. In brief, the user can customise the logic of the middleware according to his/her specific needs.

WhereX® also has a JavaScript component that, once it is integrated in the static HTML pages, allows to create Web-oriented platforms that can monitor the status of the devices almost in real-time. In particular, platforms can be created for the geolocation of devices, or panels for monitoring several physical quantities such as temperature, brightness, noise, etc. The JavaScript component mentioned above is a sort of bridge used by the middleware to notify events. It allows to update the contents of HTML pages automatically with no action to be made by the user. WhereX® totally reflects the architecture shown in the previous section. Of course, WhereX® is a finished product.

4 Case Study

As we have outlined several times, the main purposes that must be achieved by the middleware are flexibility, scalability, integration and reduction of costs. All these turn out to be an improvement in the effectiveness of business processes. Such application is a clear example of the great potentialities of the WhereX® middleware, which allows the programmer to focus on the specific problem to be solved, taking care of all the aspects related to the integration with legacy systems and enterprise systems in general, providing very powerful features to manage the user and components mobility, thus reducing the time to market the final solution as well as its quality and performance.

WhereArt® intends to provide a valuable solution to enrich the cultural offer of public and private realities (museums, parks and cultural events) introducing a new way to access and distribute information to the users.

The proposed solution is an audio–video virtual guide executed on handheld devices or tablet PCs to be provided to visitors of a cultural event in general and able to suggest personalised pathways according to age, preferences or cultural level and dynamically adapt the information on the basis of the position of the user. WhereArt® exploits the services offered by the middleware WhereX® and makes wide use of RFID and GPS technologies.

The main features of the developed solution are: *multimedia content* accessed via handheld computers (PDA) or tablet PC; *personalised user profiles* (adult, children, students, archaeologist, etc.) which allow information filtering and customisation; *generation and combination of multimedia contents* through a web-based back-end panel; *automatic conversion* of text into synthesised audio; *location of users* nearby a point of interest and automatic delivery of multimedia contents.

Thanks to the capability to easily integrate and manage technologies such as RFID and GPS and to exploit several wireless communication protocols such as ZigBee and Bluetooth, WhereArt® allows to customise multimedia contents according to the user's needs, even in highly mobile conditions. Making use of several communication technologies, WhereArt® can operate in indoor environments (museums, art galleries, etc.) as well as in outdoor spaces (parks, cities, etc.). WhereArt® offers all the modules and tools needed to manage the entire information flow from content production to fruition. All of these provide effective mechanisms to address the problems of technological integration and interoperability management between mobile devices and heterogeneous RFID systems.

The architecture of WhereArt® is basically composed of four elements: **WhereArt® CMS** allows to locate the points of interest in the map, to manage content and user profiling; **WhereX®** implements the integration and interaction among different components; **WhereArt® Mobile** allows the user to receive multimedia content and, therefore, to interact with the objects of interest; *Technology (RFID, Sensors, ZigBee, GPS, Bluetooth, etc.)*

Internet represents the underlying communication technology that guides the system behaviour.

5 Conclusions

In this paper, we have presented some principles for the creation of a RFID middleware, very flexible and scalable, that well integrates with the concept of Internet of things and paves the way towards the concrete exploitation of this new challenging reference scenario.

Our middleware is based on an SOA architecture and operates according to an event-based mode, which allows to implement an effective abstraction of the RFID technology used, by adapting the management logic of the devices. This way, the operator can avoid dealing with the issues regarding the physical interfacing of devices and focussing on the management of the events generated by the devices. An internal system of filters and rules allows to interpret the events and to customise the functional logic of the middleware.

Moving towards the direction of service integration and the well-known Internet of things, we highlighted how the proposed solution favours data exchange and simplifies the integration between technology and applications. Web-oriented applications can be easily created and can monitor the status of sensor networks or RFID devices in real-time.

We have presented the WhereX® middleware and the WhereArt® application, which by using services and basic mechanisms for mobility, location and content adaptation allow to effectively realise a personalised audio–video virtual guide to be used in both indoor and outdoor environments.

References

1. Pujolle G, (2006) An autonomic-oriented architecture for the internet of things. In: Modern Computing 2006. JVA '06. IEEE John Vincent Atanasoff 2006 International Symposium on 3–6 Oct, 163–168.
2. Kong N, Li X, Yan B (2008) A model supporting any product code standard for resource adressing in the internet of things. In: Intelligent Networks and Intelligent Systems, 2008. ICINIS '08. First International Workshop on 1–3 Nov. 233–238.
3. Michael MP (2008) Architectural solutions for mobile RFID services for the internet of things. In: Congress on Services - Part I, 2008. SERVICES '08. IEEE 6–11 July, 71–74.
4. Ning H, Ning N, Qu S (2007) Layered structure and management I internet of things. In: Future generation communication and networking (fgcn 2007) vol 2, 6–8 Dec, 386–389.
5. Gao J, Liu F, Ning H, Wang B (2007) RFID coding, name and information service for internet of things. In: Wireless, Mobile and Sensor Networks, 2007. (CCWMSN07). IET Conference on 12–14 Dec. 36–39.
6. To appear: Bruneo D, Puliafito A, Scarpa M, Zaia A (2009) Mobile middleware and its integration in enterprise systems. In: Published on Handbook of entreprise integration, Mostafa Hashem Sherif, 3rd edition, Taylor & Francis chap.6, New York - 10016 - USA, October.
7. Wu B, Liu Z, George R, Shujaee KA (2005) eWellness: building a smart hospital by leveraging RFID networks. In: Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 1–4.
8. Stankovic JA (2008) When sensor and actuator networks cover the world. In: ETRI Journal, vol 30(5), 627–633.
9. López TS, Kim D (2007) A context middleware based on sensor and RFID information. In: Proceedings of the Fifth Annual IEEE International Conference on Pervasive Computing and Communications Workshops (PerComW'07) 2007 IEEE.
10. Clauberg R (2004) IBM Research, Zurich Research Laboratory, "RFID and Sensor Networks". In: RFID Workshop, University of St. Gallen, Switzerland, Sept. 27, 2004.
11. Wu J, Wang D, Sheng H (2007) Design an OSGi extension service for mobile RFID applications. In: IEEE International Conference on e-Business Engineering.
12. Kim M, Lee JW, Lee YJ, Ryou J (2008) COSMOS: A Middleware for integrated data processing over heterogeneous sensor networks. In: ETRI Journal, Vol. 30(5), 696–706.
13. Ghayal A, Khan MZ, Moona R (2008) SmartRF: a flexible and light-weight RFID middleware. In: IEEE International Conference on e-Business Engineering, IEEE.
14. Hsu C, Mei H, Lee C, Lee D (2008) An Intelligent High Available RFID Middleware. In: Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12–15 July 2008 IEEE.
15. Wang W, Sung J, Kim D (2008) Complex event processing in EPC sensor network middleware for both RFID and WSN. In: 11th IEEE Symposium on Object Oriented Real-Time Distributed Computing (ISORC) IEEE.
16. Sheng QZ, Li X, Zeadally S (2008) Enabling next-generation RFID applications: solutions and challeges. In: Sept. 2008 IEEE Computer Society.
17. Park Y, Heo P, Rim M (2008) Measurement of a Customer Satisfaction Index for improvement of mobile RFID services in Korea. In: ETRI Journal, vol 30(5).
18. Hu W, Ye W, Huang Y, Zhang S (2008) Complex event processing in RFID Middleware: a three layer perspective, in Third 2008 International Conference on Convergence and Hybrid Information Technology, IEEE.
19. Foster (2005) Service-oriented science. Science 308(5723):814–817.
20. MacKenzie CM, Laskey K, McCabe F, Brown PF, Metz R, Hamilton BA (2006) Reference model for service oriented architecture 1.0. In: OASIS SOA Reference Model Technical Committee, http://docs.oasis-open.org/soa-rm/v1.0/.
21. XMPP Standard Foundation, http://www.xmpp.org/.
22. Pautasso C, Zimmermann O, Leymann F, RESTful Web Services vs. Big Web Services: making the right architectural decision. In: Proc. of the 17th International World Wide Web Conference (WWW2008) (Beijing, China).

23. Weijie C, Weiping L (2008) Study of integrating RFID middleware with enterprise applications based on SOA. In: 978–1–4244–2108–4/08/$25.00© IEEE.
24. Aberer M Hauswirth, Salehi A (2006) A middleware for fast and flexible sensor network deployment. In: VLDB '06, Sept. 1215, Seoul, Korea.
25. Kanoc T (1999) Mobile middleware: the next frontier in enterprise application integration. In: White paper, Nettech Systems Inc.
26. Pope A (1998) The corba reference guide: understanding the common object request broker architecture. In: Addison-Wesley.
27. Box D (1998) Essential COM. In: Addison Wesley Longman edn.
28. Java Soft. Java remote method invocation specification, revision 1.5, jdk 1.2 edition, Oct. 1992.
29. Gilman L, Schreiber R (1996) Distributed computing with IBM MQSeries, Wiley.

A Scalable Grid and Service-Oriented Middleware for Distributed Heterogeneous Data and System Integration in Context-Awareness-Oriented Domains

David Parlanti, Federica Paganelli, Dino Giuli, and Agostino Longo

1 Introduction

Context awareness deals with the capability of applications and services to react to specific events characterizing a target situation. The picture of such situation may be built by means of context information provided by sensors, i.e., physical and/or virtual sensors. Applications may exploit such situation awareness to react consequently, for instance, by enforcing adaptation actions at a logical layer (e.g., by sending notifications to interested users) and/or at a physical layer (e.g., by reconfiguring the physical environment by means of actuators).

This vision implies the integration of the physical world with the digital one and therefore requires proper instruments easing the integration of heterogeneous-embedded devices in ubiquitous computing environment (local scale) with enterprise-level services and business processes (global scale).

At present, we are investigating the adoption of an SOA (Service Oriented Architecture) approach for easing the integration of heterogeneous resources (e.g., sensors, actuators, enterprise information systems) for the development of context-aware applications in enterprise domains. We take as reference the domains of maritime surveillance and dangerous goods monitoring, where situation awareness pictures integrating local- and global-scale information resources are needed to develop added-value decision-support enterprise applications.

The SOA approach interprets distributed systems mainly as a problem of service specification, implementation, and composition. A "service" may be defined as a computational entity endowed with an open and addressable specification of

D. Parlanti (✉) and F. Paganelli
National Interuniversity Consortium for Telecommunications (CNIT), Italy
e-mail: david.parlanti@gmail.com; federica.paganelli@unifi.it

D. Giuli
Department of Electronics and Telecommunications, University of Florence, Italy
e-mail: dino.giuli@unifi.it

A. Longo
SELEX Sistemi Integrati SpA, Rome, Italy
e-mail: alongo@selex-si.com

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_11,

its expected behavior. We thus extend the definition of the "computation entity" to include software components encapsulating sensors/actuators functionalities. Integration of such real-world devices and business systems usually requires decoupling between service consumers and providers, thus demanding support also for one-way, notification response and solicit response interaction patterns. In order to address such invocation requirements, SOA's implementation solutions should be correlated also with "message-oriented" approaches. Message orientation gives new insights on service provision/consumption as well as on the overall SOA architectural style. More specifically, services can now be simply defined as "message-processors," while a service-oriented system can be consequently interpreted as a "network of connected message processors." Under this perspective, we present the SAI – Service Application Integration – system as a working example of a message-oriented SOA solution empowered with GRID scalability for data and system integration.

2 Related Work

This paragraph presents a few relevant works proposing middleware solutions for integrating sensor and actuator networks with web and enterprise applications.

Karnouskos et al. [4] propose a web-service device-to-business integration infrastructure by applying SOA principles to networked embedded devices. The solution is based on a web service approach, where each device offers its functionality as a set of services. Devices are attached to the middleware directly as web services using DPWS (via DPWS-enabled controllers) or by means of legacy system controllers.

The Global Sensor Network (GSN) middleware [1] is a solution aiming at providing a uniform interface for easing the integration and deployment of heterogeneous sensor networks. It is based on the abstraction of "virtual sensor," representing real sensors or software components aggregating different sensors in terms of input streams and one output stream. GSN adopts a container-based architecture. The container provides services for virtual sensors management, including remote access, security, persistence, and concurrency.

In [6], RESTful principles have been applied to elaborate a logical architecture of a middleware for enabling plug and play access to heterogeneous sensor and actuator networks, including addressing, discovery, and controlling mechanisms.

A proposal for a Service-Oriented Device Architecture (SODA) is presented in [3]. The objective of SODA is to integrate a wide range of physical devices into distributed enterprise systems, by providing the capabilities for accessing sensors and actuators as business services. A SODA implementation includes three main components: the device adapter, which translates proprietary and industry-based standard interfaces of devices into the device service abstract model; the bus adapter which maps the device service abstract model to the enterprise-level SOA binding mechanism; a device service registry providing discovery capabilities.

3 SAI Middleware Overview

The SAI middleware is targeted at enabling information search and data mash-up in complex, distributed enterprise environments characterized by strong technological heterogeneity. By "heterogeneity," we mean here differences in data-representation formats, data schema structure and semantics, communication protocols, security requirements and security-credentials management, and processing capabilities. The goal is that of making possible for clients to search data without any knowledge about its physical location in the distributed environment, while not worrying either about the consistency of data, even in case of failure of system components or of connected legacy information systems, nor about how the system internally distributes computing power or load-balances service requests.

3.1 SAI Architecture

The SAI architecture accounts for its primary system-level objectives through the adoption of solid patterns in distributed systems design, including: the "Message Broker" pattern as regards interaction among the system's heterogeneous components; the "Adaptor" pattern for enabling uniform access to the orchestrated legacy systems; the "Master/Worker" pattern for enabling distribution and load-balancing of the system's computational workload. A snapshot of the currently implemented SAI architecture is shown in Fig. 1. The logical structure of the architecture and the functionalities provided by each component will be presented in the following subsections.

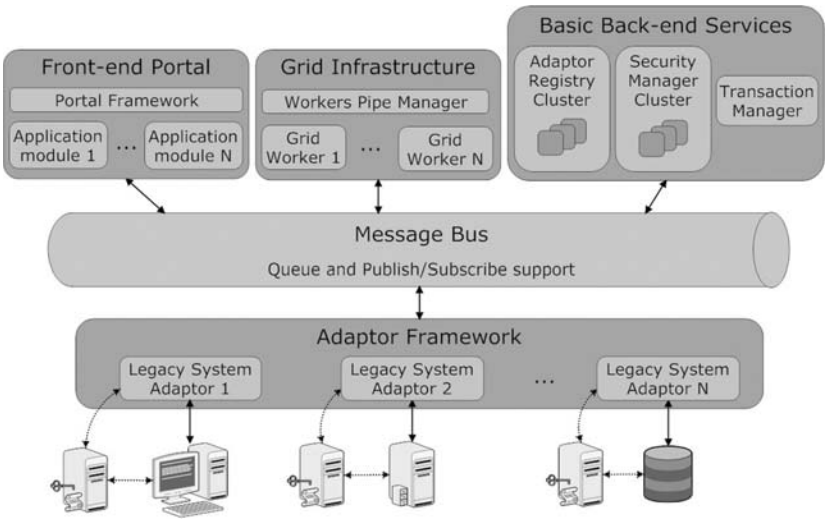


Fig. 1 SAI middleware architecture

3.1.1 Front-End Portal

The Front-End Portal infrastructure enables the uniform development of SAI front-end applications. The Portal Framework is a container establishing requirements for valid front-end "application modules," which can be composed at run time and delivered as a fully integrated web application and accessed via Desktop PC or mobile phone. At present, the Portal Framework component is powered by Liferay, an open-source implementation of the Portlet specification.

3.1.2 Grid Infrastructure

It is well-known that context aware applications usually need to process large amounts of data and that such processing needs parallelization to improve overall system performance, resilience and throughput. Accordingly, the SAI Grid Infrastructure provides workload distribution to applications and system services. SAI's Grid has been developed from scratch in the Java Standard Edition for achieving full control and configurability of load splitting over system nodes. Our implementation strictly follows the Master/Worker design pattern: at its heart, the grid simply consists of three entities: a master (that is the "client" of the grid), a channel for enabling master to worker communication and a set of one or more worker instances. According to such pattern, the master starts parallelization by defining a set of "jobs" which are then distributed (or "mapped") to worker processes, then waiting for scheduled task to be completed. The final step then implies the master to organize (or to "reduce") collected results into a "single" meaningful unit which shall be coherent with the semantics of the distributed work.

The SAI grid also provides capabilities for intelligent and configurable routing of jobs to workers and dynamic jobs reassignment in case of workers failures. To provide such capabilities, the "PipesManager" component associates each worker instance to a unique "pipe" composed of a "pending jobs" and of a "completed jobs" queue. Being each worker linked to a unique pipe, jobs contention is minimized, while masters – which act as the clients of the grid infrastructure – can exploit pipes information to enable configurable jobs-routing algorithms. Moreover, the PipesManager monitors the "liveliness" of each enabled worker and reacts to possible failures by reassigning pending jobs on "live" executors: of course, jobs dynamic reassignment is transparent both to masters and to live workers.

3.1.3 Basic Back-End Services

This group of components contains services providing cross-cutting capabilities supporting the functioning of the whole SAI infrastructure.

The Security Manager is the SAI basic back-end service providing mechanisms for validating the identity of system's principals, for granting or denying authorization for actions performed on SAI-managed resources, for guaranteeing the integrity and confidentiality of messages, and for supporting the evidence of actions performed by system's principals.

The "Adaptors Registry" component manages the "functional profile" of each information system which is connected to the SAI system by means of a dedicated adaptor. A functional profile describes the message-processing capabilities of an adaptor through ordered input-output pairs of XML message-type identifiers and "meta properties" (unordered name-value pairs) describing the nonfunctional capabilities of the mediated legacy system. The Adaptors Registry enables querying of registered functional profiles by authorized clients and monitors the operational status of activated adaptors to keep up-to-date the profiles registry status. It also listens for notifications concerning variations in the adaptors functional profiles.

The Transaction Manager component is charged of coordinating global (distributed) transactions in the SAI distributed environment to ensure consistency of data access and manipulation operations. At present, the Transaction Manager component is powered by JOTM, which is an open and standalone implementation of the JTA (Java Transaction API) specification.

3.1.4 Message Bus

The Message Bus is the infrastructure providing application-level messaging capabilities to the SAI system components. The SAI interfaces are completely decoupled from any concrete messaging broker implementation, in order to enable maximum flexibility and scalability of the middleware. Indeed, messaging capabilities can be provided either by a full-fledged Message Oriented Middleware (MOM) or by lighter solutions based on IP multicast, such as the JGroups library. At present, the Message Bus component is powered by ActiveMQ, an open-source implementation of the Java Message Service (JMS) specification.

3.1.5 The Adaptors Framework

The SAI Adaptors Framework enables the interfacing of the SAI with heterogeneous information systems. Interfacing happens adaptation of the proprietary data format supported by legacy systems to the shared XML data model possibly used within an SAI-enabled application domain.

An SAI adaptor should be considered as a lightweight and configurable XML processor. Indeed, each Adaptor is a microcontainer for a pluggable service implementation. An adaptor-managed service embeds the integration logic with an external information source (e.g., an embedded device or an enterprise-level system). In this regard, the service is a client of the legacy system, and it is charged of parsing received requests into the proprietary "dialect" spoken by the legacy system over the supported network communication protocol. Each service can be completely described through its "functional profile." The service functional profile describes the service's processing capabilities by means of ordered input-output pairs of XML message types. The microcontainer manages the life cycle of the service, while being capable to filter both incoming and outgoing XML documents

through inbound and outbound interceptor chains. By exploiting such "interceptor pattern," it is possible to augment the microcontainer capabilities at runtime, simply by injecting additional interceptors in the inbound/outbound chains through configuration. As an example, interceptors have been implemented to enable compression of network streams, or to provide support for WS-* standards in order to free service implementations from unnecessary Web Services plumbing. The Adaptor is not bound to any specific XML envelope or format (e.g., SOAP), while support for XML standards can be configured at runtime through dedicated inbound and outbound interceptors. Special security interceptors can also be developed to condition request processing or messages dispatching to authentication/authorization policies of legacy security systems. Finally, an Adaptor can be configured to listen for request messages on a variety of network transport protocols and can support synchronous request/response, asynchronous one-way, asynchronous notification-response, solicit-response messaging patterns.

3.2 System Dependability

System dependability can be defined as "the ability to avoid service failures that are more frequent and more severe than is acceptable" [2]. At present, the SAI framework achieves satisfactory dependability levels by means of its Grid Infrastructure (which natively handles load-balancing and jobs-failover, as shown in the previous section), by the clustering of Basic Back-End Services and Adaptors and by the implementation of some basic autonomic capabilities for preserving the security state of the whole system. Indeed, the SAI architecture achieves basic support for autonomicity by exploiting the well-known "heartbeat" technique, since each SAI component can produce heartbeats that can be audited by other active components. Heartbeats help monitoring the overall system state and also enable automatic reaction of adaptors with regard to the state of the Security Manager service cluster. Hence, when adaptors stop hearing heartbeats from the security cluster, then they stop accepting requests and dispatching of outgoing messages until security heartbeats are resumed. This capability, although very simple to implement, allow for fail-over strategies that are capable to preserve the security state of the system.

3.3 Service Invocation and Composition in the SAI Framework

Each SAI information system adaptor is described by a functional profile that is stored in the Adaptors Registry cluster. Each adaptor is a message processor: its functional profile can be completely described by a set of unique XML input and output message-type pairs. Given such assumptions, Fig. 2a shows that message "M1" can be transformed into message "M2" through adaptor "A1" and that message "M3" can be transformed into message "M4" by means of adaptor "A2."

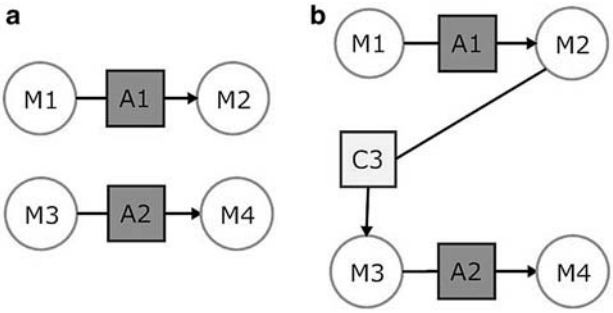


Fig. 2 Graph-based representation of service invocation and composition: (**a**) message "M1" can be transformed into message "M2" through adaptor "A1" and that message "M3" can be transformed into message "M4" by means of adaptor "A2"; (**b**) the connector "C3" links "M2" to "M3" through a syntactical transformation and it makes it possible to reach "M4" starting from "M1"

We can equivalently state that there is a "path" from M1 to M2 and another path from M3 to M4. When considered as a whole, the set of message nodes M1, M2, M3, and M4 and edges A1 and A2 together define a disconnected and directed graph "G," which is just an alternative representation for the information stored in the Adaptors Registry cluster state. It can be established whether a specific message can be transformed into another type of message by checking if a path between the two message types is found to exist in the graph-based representation provided by the Adaptors Registry. Such a path can be considered as the special "workflow" to execute in order to provide clients with the requested information.

Of course, situations may also occur where a same type of message can be produced by multiple adapters. For instance, we can suppose that M2 can be produced from M1 either by means of the A1, A2, or A3 adapter. SAI clients are then required to select among the available path from M1 to M2 by maximizing an objective function "f." The objective function is specified so as to take into consideration client preferences, as represented in their issued data-aggregation requests, and quality of service metrics as collected during repeated interaction with registered adapters.

In the SAI system, a "service invocation" simply consists of two elements: (1) a "request message," a message embedding information which is known or can be arbitrarily specified by a message producer; (2) the specification of a "target" message type to be produced by the SAI system, given the "request message" as input. Hence, the simplest request that clients can make to the system consists of a "target" message for which a direct path from the provided "request" message is found to exist in the adaptors registry. A complex request consists of a "target" message for which no direct path via a single adaptor invocation from the provided "request" message is found to exist in the adaptors registry. For instance, it turns out that a request for M4 given M1 is a complex request since there is no direct single connection between M1 and M4. A complex request could thus be handled by a path composed of multiple adaptors. In case that there are no registered adapters enabling M4 to be reachable from M1, from a logical point of view, a special entity ("connector") needs to be introduced in order to enable the system to respond to these types of requests.

A connector is an entity which is specialized in syntactical transformations (e.g., message-format adaptations) among message types. It is allowed to link message nodes with no outgoing links with message nodes with no incoming links. In Fig. 2b, we can see that connector "C3" transforms "G" into a connected graph, because it links "M2" to "M3" through a syntactical transformation: thanks to the connector, it is now possible to reach "M4" starting from "M1."

The injection of connectors into the SAI system, and the capability of clients to select one off multiple competing paths by means of special optimization algorithms, together allow the SAI system to solve complex data-aggregation scenarios. Moreover, since invocation workflows are created "on-the-fly" through simple operations on the graph-representation of adaptors profiles, there is no need for predefined hard-coding of invocation sequences into the system. Hence, as more connectors can be added progressively during the life-cycle of the system, SAI information retrieval capabilities can be said to be "evolutionary."

4 Case Study for Dangerous Goods Monitoring

We are currently performing some experimentation activities for evaluating the use of the SAI middleware for developing context-aware applications.

A preliminary case study for the SAI middleware has been designed in the framework of a research project in the domain of *european maritime surveillance*. The goal of the project is the provision of a secured information exchange platform capable of bringing together the existing monitoring and tracking systems used for maritime safety and security, protection of the marine environment, fisheries control, control of external borders, and other law enforcement activities in order to satisfy information needs and information production profiles of heterogeneous actors (such as European Level Agencies Layer, Member State Ministries, Member State Operating Bodies).

Current research and prototyping activities are focused on a case study for *monitoring dangerous goods shipping across intermodal transport routes*.

Complexity of this application domain is due to several factors: the kind of transported goods and related risks for the surrounding physical and social environments; the heterogeneity of transportation means that are usually required for end-to-end delivery, the wide range of users which is involved to different extent in the shipping process. Most important user categories include [5]: transporters; final users (sender and consignee); multimodal operator; administration, authorities, traffic control services; emergency services. These users may be characterized by different regulations, specification, and technological infrastructure and pose different information needs requirements in terms of content type, information granularity, timing constraints.

We are currently designing and developing a service platform which, by leveraging on the capabilities provided by the SAI middleware, aims at providing information services to such different user categories. The middleware will support

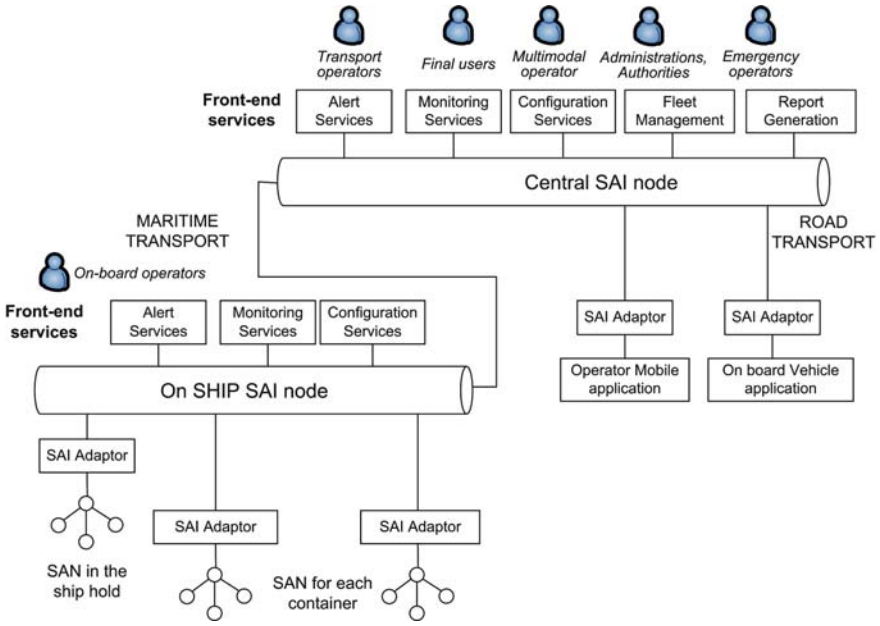


Fig. 3 Architecture of the dangerous goods monitoring case study

interoperability among the involved organizations and public authorities, while also hiding to clients the heterogeneities of the underlying technological infrastructures (e.g., sensors and actuators, on-board and mobile devices, enterprise-level and web-based application, etc.).

The prototype of the service platform has been customized to fit both the maritime transportation of dangerous goods and the interconnection with road transport segments. The service platform architecture is shown in Fig. 3.

The monitoring and control infrastructure is based on a set of RFIDs, sensors and actuator networks (SANs) which could be deployed in the container (or other unit loads) and/or in the ship hold in order to monitor and control the environmental conditions. The SANs expose their features as services via the SAI Adaptor (encapsulating SAN features as services to be exposed in the SAI architecture).

The back-end and front-end application logic which is deployed locally (e.g., on the ship on a local instance of the SAI node) and/or remotely (on a central SAI node) receives messages from the SANs. The on-ship monitoring service processes the information and if required, may trigger alert services to the on-board operators and/or to the remote monitoring system (e.g., via satellite communication). Analogously, SANs may receive reconfiguration commands for adapting their behavior according to a remote or local configuration adaptive logic (e.g., to increase the frequency of message delivery in case of abnormal conditions).

5 Conclusions

The SAI middleware shows that also enterprise-class technologies can be exploited in the development of context-aware applications. The proposed middleware has been empowered with a Grid infrastructure to offer workload distribution and load-balancing for the processing of context data, while still preserving the security state of the system and data consistency. Ongoing activities on the case study for dangerous goods monitoring in intermodal transport will help assessing the added value of the proposed integration system in a Web of Things scenario.

Acknowledgments This work was partially supported by SELEX Sistemi Integrati. Technical assistance from Luca Capannesi, Department of Electronics and Telecommunications of the University of Florence, is gratefully acknowledged.

References

1. Aberer K, Hauswirth M, Salehi A (2006) Middleware support for the "Internet of Things". In: 5th GI/ITG KuVS Fachgespräch "Drahtlose Sensornetze", Stuttgart, Germany
2. Avizienis A, Laprie J, Randell B, Landwehr C (2004) Basic concepts and taxonomy of dependable and secure computing. IEEE Trans Dependable Secure Comput 1(1):11–33
3. de Deugd S, Carroll R, Kelly K, Millett B, Ricker J (2006) SODA: Service Oriented Device Architecture. IEEE Pervasive Comput 5(3):94–96
4. Karnouskos S, Baecker O, de Souza LMS, Spiess P (2007) Integration of SOA-ready networked embedded devices in enterprise systems via a cross-layered web service infrastructure. Proceedings of 2007 Emerging Technologies and Factory Automation Conference, pp 293–300
5. Nathanail T (1995) Architectural design for the monitoring of intermodal transportation of hazardous goods. Proceedings of the 1995 TransTech Conference, pp 69–73
6. Stirbu V (2008) Towards a restful plug and play experience in the web of things. Proceedings of the 2008 International Conference on Semantic Computing, pp 512–517

InterDataNet: A Scalable Middleware Infrastructure for Smart Data Integration

Franco Pirri, Maria Chiara Pettenati, Samuele Innocenti, Davide Chini, and Lucia Ciofi

1 Introduction

The Internet of Things (IoT) is a vision in which the Internet extends into our everyday lives through a wireless network of uniquely identifiable "objects" or "things" [22]. While the "Internet of Things" perspective has been focused on enabling the networking capabilities of the Things, the "Web of Things" perspective proposes to integrate the networked "Things" into the Web [20], thus making them available as resources, i.e., documents. The advantage of this integration approach is that the "Things" can be treated like any other Web resources, leveraging on the REST architectural style to provide low entry barrier. This would eventually enable the loosely coupled "Things" to be reused in different contexts and applications. In the meantime, the Web itself would be evolving [9] and undergoing significant transition reflected by the massive effort to make the WWW evolve from a file-server to a database paradigm, with the purpose in mind of creating an "interlinked data Web" doing for data what the World Wide Web did for documents; this vision is known as the Linked Data approach [1, 2]. In the words of the creator of the Web itself, "The Linked Data approach is a Web of Things, conceptually, because it gives the thing a notion of its own identity in the Web. Tying an actual thing down to a part of the Web is the last link – the last mile" [13]. Taken together the two approaches, the Internet/Web-of-Things and the Web of Data concur in the vision of a data-web (lower case is deliberate) evolving to become part of a world-wide database to be both exploited and enriched by internetworking "things" on a Web-wide scale. In order for such view to come true, the realization of "ubiquitous network environment," that is the integration of heterogeneous sensing devices, heterogeneous actuators, and Web of Data is required. At the state of the art few convincing architectural styles for such integration exist, while the academic and industry interest in this domain is growing [16, 19]. Some projects, for instance [22], though proposing innovative applications, focus on IoT applications developed as

F. Pirri (✉), M.C. Pettenati, S. Innocenti, D. Chini, and L. Ciofi
Electronics and Telecommunications Department, University of Florence, Italy
e-mail: franco.pirri@unifi.it; mariachiara.pettenati@unifi.it; samuele.innocenti@unifi.it; davide.chini@gmail.com; lucia.ciofi@unifi.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_12,

dedicated systems with specific environments in mind with low extensibility and reuse capabilities. Other works, aiming at offering objects integration architecture on a Web-wide scale [3, 8, 11, 20] though targeting the scalability as a main driving principle, focus on making all resources available via standard Web mechanisms, dealing with "things" much likely as other Web resources that are documents. InterDataNet approach brings scalable interoperability to IoT leveraging on the Web of Data paradigm. Such an approach entangles using flexible identification mechanism allowing rich description, management, and interoperability of heterogeneous objects and things. A key step in this direction is making data "smarter." Indeed, the smarter are the data describing the objects, the easier is sharing, linking, and processing between distributed applications. Making data smarter and using such data at an infrastructural level, means moving smarts into the data itself rather than hard coding them into software applications and/or smart objects.

In this chapter, we provide an overview of the InterDataNet architecture designed to enable heterogeneous objects networks to expose and integrate their smart data. At the core of the system sits the InterDataNet middleware that defines an object Information Model and the related Service Architecture operating on it in order to provide: (a) *a scalable and open service to support a consistent reuse of objects identifiers*, that is a global reference and addressing mechanism for locating and retrieving objects in a Web-wide scale; (b) *a set of transparent application-services functions*, namely historic data management and replica management.

This chapter focuses on discussion of architectural elements rather than implementation details. For more complete technical and implementation details, see [10] and forthcoming papers.

2 Grounding Principles and Design Paradigms of the IDN Framework

To fulfill the main driving design criteria [6, 11, 18–20] of extensibility, low entry barrier, interoperability, and scalability of the architecture we propose, we adopt the following set of conceptual and technological design paradigms:

- *The system has to be layered*; layering [23] is the architectural pattern to pursue scalability and legacy data integration at infrastructural level. This is achieved, implementing the functionalities at different IDN layers as network devices.
- *The design of middleware following Service Oriented Architecture (SOA) approach* [12, 15] to allow the development of loosely coupled and interoperable infrastructural services that can be combined into more complex systems.
- *The system has to adopt a REST style (Representational State Transfer) paradigm* to make InterDataNet an explicit resource-centric infrastructure [7, 21]. REST architectural style that emphasizes scalability of component interactions, generality of interfaces, independent deployment of components, and intermediary components to reduce interaction latency, enforce security, and encapsulates

legacy systems. WWW is one instance of a real distributed system based on REST. WWW enjoyed the scalability and growth as a result of a several design principles defined in REST. In REST approach, every resource (sources of specific information) is uniquely addressable, each of which can be referred to using a global identifier (a URI). In order to manipulate these resources, components of the network (clients and servers) communicate via a standardized interface (e.g., HTTP, a protocol that is Client/Server, stateless, cacheable and layered) and exchange representations of these resources (the actual documents conveying the information) [11].

3 IDN Main Views and IDN Naming System

IDN is described through the ensemble of concepts, models and technologies pertaining to the following three views Fig. 1:

IDN-IM (InterDataNet Information Model). It is the shared information model representing a generic document model which is independent from specific contexts and technologies. It defines the requirements, desirable properties, principles and structure of the document to be managed by IDN. In the IDN-IM, information is highly structured and endowed of specific metadata. Moreover the Information Model defines the conceptual operations that the applications perform on the pieces of information and documents. Generic information modeled in IDN is formalized as an aggregation of elementary data units, named Primitive Information Unit (PIU). Each PIU contains data and metadata. All data and metadata are handled, by the IDN-Service Architecture. Among those, each metadata element may be visible to applications; it can be defined by the application itself (in this case it is named "application properties") or by the Service Architecture and, if needed, used by applications. Another class of metadata elements exist for internal Service Architecture use only.

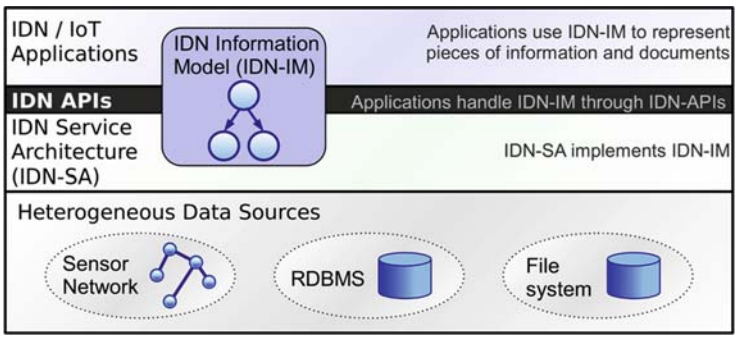


Fig. 1 IDN framework

Each node is addressed by one or more URIs related to one or more names (aliases), as it will be detailed in Section IDN Naming System. An IDN-document structures information units and it is composed by nodes related to each other through directed "links." Moreover, IDN-documents can be inter-linked, so two main link types are defined in the Information Model: *aggregation links*, to express relations among nodes inside an IDN-document; *reference links*, to express relations between distinct IDN-documents. Each PIU belonging to the document can also be addressed as a document root node increasing information granularity and reuse. IDN-IM documents express data contents and relation between contents. Data and metadata are structured following the name-value representation and embedded inside the node.

IDN-SA (InterDataNet Service Architecture). It is the architectural layered model handling IDN-IM documents (it manages the IDN-IM concrete instances allowing the application to "act" on pieces of information and documents). The IDN-SA implements the reference functionalities defining subsystems, protocols and interfaces for IDN document management. The IDN-SA exposes an IDN-API (Application Programming Interface) on top of which IDN-compliant Applications can be developed. The IDN-SA encompasses the IDN Naming Systems handling the structured document addressability and resolution of IDN-documents names.

IDN-Applications. IDN-compliant applications implement the context-dependent logic and store/manage information using the IDN-API to exploit the IDN-SA services. IDN applications use the shared IDN-IM to represent and handle their pieces of information and documents.

In accordance to the REST approach, *IDN naming system* adopts an HTTP:URI-based naming conventions to address IDN-nodes. IDN-nodes have, at least, one canonical name which unambiguously identifies the node for each state of its lifecycle; the canonical name has to be globally unique and is assigned once when the node is created. The uniqueness of the name is achieved by the use of URI: the creator of the node has to handle local unique identifiers only, while the LDNS is entitled to do the rest of the job. More than one name can be assigned to IDN-nodes. Such names are also URI-based and are referred to as "aliases." These are "logical names" of the resource. For instance, the resource describing Temperature Sensor identified by the canonical name:

"http://interdatanet.example.org/examples/sensors/temperature/tz04b."

To refer to the same information in a different context, we can add the alias name: "http://mydomain.example.org/my_home/kitchen_temp." Starting from these logical names, mechanisms to unambiguously physically locate and access the specific resource are needed. To this end, IDN architecture is composed of three logical conceptual layers (see Fig. 2 left part): (a) *the upper layer handles HFN (Human Friendly Identifiers)* and it is used by applications to identify IDN nodes. The node's canonical name and its aliases are located in this class; (b) *the middle layer handles URN (Universal Resource Names)* to unambiguously, univocally and persistently identify the IDN nodes independently of their physical locations; (c) *the lower layer handles URL (Uniform Resource Locators)* to identify and access Storage Interfaces adapting the object data model to the IDN-Information Model. As any resource can be replicated in several locations, each URN can correspond to many URLs.

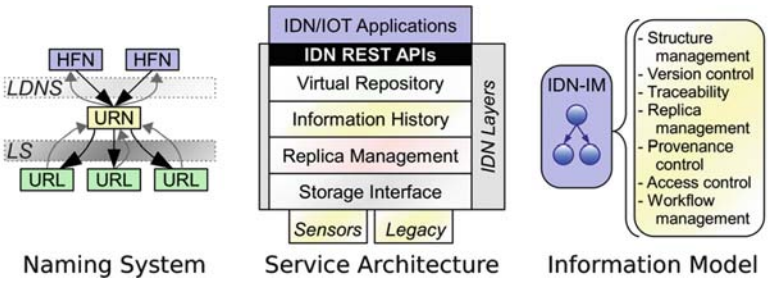


Fig. 2 IDN three layers Naming System, Service Architecture and Information Model

As HFNs, URNs, and URLs are sub-classes of URIs, they are hierarchical and their direct and inverse resolution is possible using DNS (Domain Name System) system [14] and a REST-based approach. By construction, the IDN naming system is compliant with Linked Data principles [4] through the choice to adopt HTTP:URI-based naming conventions and management of identifiers.

3.1 The IDN Service Architecture

The IDN-SA provides an effective and efficient infrastructural solution for IDN-IM implementation. IDN-SA is a layered service-oriented architecture and it is composed of four layers (see Fig. 2, middle): Storage Interface Layer; Replica Management Layer; Information History Layer; Virtual Repository Layer. IDN-SA layers functions are hereafter briefly specified, starting the description from the bottom of the stack. For space constraints, in this section we will not detail on two aspects related to versioning and replica management. Their integration in the IDN architecture is fundamental in order to provide collaboration-enabling functions, but their detailed description goes beyond the scope of the present paper.

Storage Interface Layer (SI). The SI layer provides a REST-like uniform view over distributed heterogeneous data independently from their location and physical data source. The SI provides physical addressability to resources through URLs addresses.

Replica Management Layer (RM). This layer provides a delocalized view of the resources to the upper layer offering URN (Universal Resource Name) which are used here to identify resource, to URL address resolution through the LS (Localization Service). As such the upper layer handles only resources' univocal and persistent identifiers and allows the association of several physical resources (replicas) to the same identifier.

Information History Layer (IH). This layer manages PIUs history allowing navigation and traversing into the versioned information space. At this layer, PIU are identified through URN plus an optional version parameter identifying the time-ordered position. This URI identifies a version of the node.

Virtual Repository Layer (VR). It exposes the IDN APIs to the IDN-compliant Applications exploiting lower layers services. VR provides the maximum abstraction of structured information to the application. Indeed, this layer is seen from the application as the container-repository of all PIUs. The resolution of Human Friendly Names (HFN, logical resource identifiers) into unique identifiers (URN) is realized in this layer exploiting the LDNS (Logical Domain Name System) service which is logically located inside the VR layer (see Fig. 2, middle). A sub-service of the VR layer, namely the Resource Aggregation Service (RAS), is entitled to receive all document requests from the applications. When it receives a request, it collects all the content by traversing the links listed in the PIUs, starting from the root node (addressed through the HFN). Each PIU contributes as a building-block to construct the requested information. In agreement with the request, the RAS applies a conversion in the desirable format and replies.

4 IoT Objects Representations Supported by the IDN Framework

In a typical sensor networks [5], sensor data generated by each sensor node is routed toward a sink node. Then, a system attached to the sink node processes the data format into appropriate data format to be provided to the application [3]. The sink node accepts a query from a client and serves as an interface between the sensor network and applications. The IDN approach does not make any assumptions on the internal of a sensor network other than that the sink node is accessible by the Storage Interface and allows an abstraction of sensor data offering a document representation exposed by IDN-API. The IDN-SI can also be layered into the sink node itself. The IDN-API provides what we name the "IDN Virtual Object" (see Fig. 3) (IDN-VO) document description compliant with the IDN-IM (Information Model) abstracting objects data from implementation details. An IDN Virtual Object can be any kind of data producer, for example, a real sensor, a wireless camera, a desktop computer, a cell phone, or any combination of them.

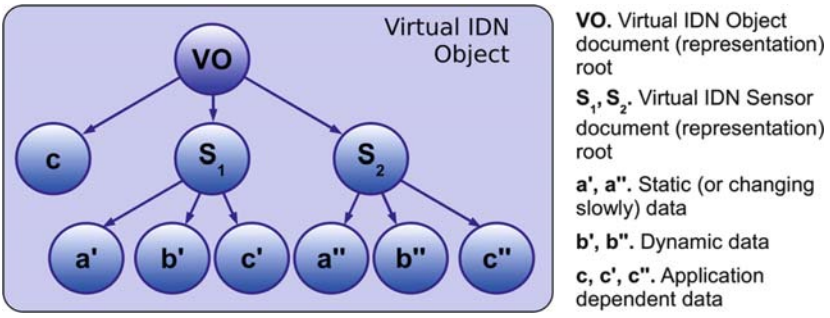


Fig. 3 Example of IDN Virtual Object composed of two sensors (S1 and S2)

An IDN-VO exposed through the IDN-API provides a rich, globally accessible virtual object description composed of three elements: (a) *information related to the object*; i.e., a description of the object itself in terms of parameters which are static (or changing slowly), and it is mainly conforming to the existing standards (e.g., sensorML markup language); (b) *information related to the value/measures performed by the object*; these are couples name-value (e.g., temperatures measures over time), (c) *ad-hoc description of the object created by the application*; this type of information brings into the document model the data which are relevant for the specific application (e.g., metadata for specific smart spaces applications).

The three elements of the IDN-VO are nodes in an IDN document; as such, they are uniquely addressable and can be separately managed by the IDN infrastructure benefiting also from its versioning and replica capabilities. Data handled by the layers of IDN are consequently transformed from low-level data into meaningful, high-level information to be exploited to build applications. Application-level services (such as context-aware services) may use information provided by several sensor networks or interact with actuators located in different networks in order to complete their task. The sensor networks can provide back into the system high-level data or complex actuation tasks that can be used by other application aware services to produce higher level of abstraction or more complex task.

5 Conclusions and Research Challenges

This paper discusses the architectural requirements of a scalable networking architecture that seamlessly integrates heterogeneous information sources to provide rich context aware services. The paper is focused on discussion of architectural elements rather than implementation details. For more complete technical details, see [10] and forthcoming papers.

At the core of the system sits the IDN middleware an infrastructural solution supporting a decentralized and scalable publication space for smart data integration and interoperability. In IDN, the things/object is represented by an IDN document, and IDN Virtual Object, which is a virtual abstraction of the thing (or "virtual thing" [3]) that is interlinked nodes of data and related metadata; it is described in compliancy with the IDN Information Model and it is univocally addressed by an HTTP URI (PRI). Thanks to the properties of the IDN-IM described above, each node of the "virtual thing" is univocally identified and addressed through an URI.

The definition of the IDN Virtual Object, together with the functions provided by the IDN Service Architecture and, in particular, by the IDN Naming System, brings a set of advantages which place the IDN approach beyond the state of the art:

- *The global addressability*. Attained through the IDN three-layers naming system providing scalable and consistent reuse of URI identifiers for reference, addressing, and retrieval of resources on a Web-wide scale

- *The transformation of low-level object information into more meaningful, high-level information, i.e., smarter data*. Attained by the management of the IDN Virtual Objects by the different layers of the IDN Service Architecture
- *The inherent possibility to maximize the efficiency in the creation of "virtual objects"* reusing parts of the objects descriptions of the IDN VO, that is handling the effective orchestration of different objects at an infrastructural level, through the IDN stack, an consequently easing the applications development. Virtual Objects and their data streams can be combined in arbitrary ways, and thus enable the applications to use a data-oriented IoT consisting of an objects networks connected via the IDN
- *The possibility to augment the data management with semantic elements*; this is leveraged by the direct compliancy of the IDN Naming System with the Linked Data [4] approach envisaging the application of the URI naming and HTTP URI addressing to the semantic description of the resources to enable building Web-wide semantic applications [17].

Within the InterdataNet perspective on the Web of Thing (WoT) vision, a number of relevant issues should deserve more deep discussion to validate the approach, however we just mention those that are objects of active research and development in our team. General level issues concern the implementation details of the IDN Service Architecture and Naming System [10], and demonstrating how the IDN fits into the Web of Data and Linked Data view [17]. As for the IDN perspective on the IoT, the following issues are relevant: how IDN controls objects in the IoT, that is how it interacts with actuators, how the IDN makes or enables the discovery of Objects, how the IDN is able to provide notifications from Objects, how the IDN addresses the security issue, how IDN handles real-time availability requirements of some IoT applications, how IDN satisfies the objective to augment its features with semantic capabilities.

Like WWW provided a global space for the seamless integration of local "webs of documents" into a global, open, decentralized and scalable publication space thanks to TCP/IP and internetworking layered solutions, we claim that the realization of the grand vision of the interlinked smart data would be much easier and faster if we could count on an "interdataworking" infrastructure. The IDN layered middleware aims to provide an attempt in the direction of an interdataworking vision.

Acknowledgments We would like to acknowledge the valuable support of Prof. Dino Giuli for the material and scientific support to this research activity. Moreover we acknowledge the precious work of Luca Capannesi for the technical support in the development activities.

References

1. AA.VV. (2008) Linked Data on the Web (LDOW2008) Proceedings WWW 2008 Workshop, April 22, 2008 Beijing, China. http://events.linkeddata.org/ldow2008/
2. AA.VV. (2009) Linked Data on the Web (LDOW2009) Proceedings WWW 2009 Workshop, April 20th, 2009 Madrid, Spain. http://events.linkeddata.org/ldow2009/

3. Aberer K, Hauswirth M, Salehi A (2007) Infrastructure for data processing in large-scale interconnected sensor networks. In: Mobile Data Management (MDM'07), Mannheim, Germany
4. Berners-Lee T (2006) Linked data – design issues. W3C http://www.w3.org/DesignIssues-/LinkedData.html
5. Culler D, Estrin D, Srivastava M (2004) Guest editors' introduction: overview of sensor networks. Computer 37(8):41–49
6. Dolin RA (2006) Deploying the "Internet of Things". In: Proceedings of the international symposium on applications on Internet, IEEE Computer Society, pp 216–219
7. Fielding RT (2000) Architectural styles and the design of network-based software architectures. Unpublished Doctoral dissertation, University of California, Irvine
8. Gronbaek I (2008) Architecture for the Internet of Things (IoT): API and Interconnect. In: Sensor technologies and applications, 2008. SENSORCOMM '08, Second International Conference on, pp 802–807
9. Hendler J, Shadbolt N, Hall W, Berners-Lee T, Weitzner D (2008) Web science: an interdisciplinary approach to understanding the web. Commun ACM 51(7):60–69. doi: 10.1145/1364782.1364798
10. Innocenti S (2008) InterDataNet: nuove frontiere per l'integrazione e l'elaborazione dei dati. visione e progettazione di un modello infrastrutturale per l'interdataworking. Ph.D. Dissertation, Electronics and Telecommunications Department, University of Florence, Italy, Dec 2008
11. Kawahara Y, Kawanishi N, Ozawa M, Morikawa H, Asami T (2007) Designing a framework for scalable coordination of wireless sensor networks, context information and web services. In: Distributed computing systems workshops, 2007, ICDCSW '07. 27th international conference on, p 44
12. Lund K, Eggen A, Hadzic D, Hafsoe T, Johnsen F (2007) Using web services to realize service oriented architecture in military communication networks. Commun Magazine, IEEE, 45(10):47–53. doi:10.1109/MCOM.2007.4342822
13. McManus R (2009) Interview with Tim Berners-Lee, Part 2: Search Engines, User Interfaces for Data, Wolfram Alpha, and More... ReadWriteWeb, 10 July 2009
14. Mockapetris PV, Dunlap KJ (1995) Development of the domain name system. SIGCOMM Comput Commun Rev 25(1):112–122. doi: 10.1145/205447.205459
15. Matthew MacKenzie, Laskey C, McCabe K, Brown F, Peter F, Metz R, Hamilton BA. Reference Model for Service Oriented Architecture 1.0 OASIS Standard, 12 October 2006 http://docs.oasis-open.org/soa-rm/v1.0/
16. Pachube: Building a platform for Internet-enabled environments. http://community.pachube.com/
17. Pettenati MC, Chini D, Parlanti D, Pirri F. InterDataNet: A Web of Data foundation for the Semantic Web vision IADIS International Journal on WWW/Internet 6(2):16–30 ISSN: 1645–7641
18. Prehofer C, van Gurp J, di Flora C (2007) Towards the web as a platform for ubiquitous applications in smart spaces. In: Second workshop on Requirements and Solutions for Pervasive Software Infrastructures (RSPSI), Innsbruck
19. Presser M, Barnaghi P, Eurich M, Villalonga C (2009) The SENSEI project: integrating the physical world with the digital world of the network of the future – [global communications newsletter]. Commun Mag IEEE 47(4):1–4
20. Stirbu V (2008) Towards a RESTful plug and play experience in the web of things. In: Semantic computing, 2008 IEEE international conference on, pp 512–517
21. Vinoski S (2007) REST eye for the SOA guy. IEEE Internet Comput 11(1):82–84
22. Welbourne E, Battle L, Cole G, Gould K, Rector K, Raymer S et al (2009) Building the Internet of things using RFID: the RFID ecosystem experience. Internet Comput IEEE 13(3):48–55. doi: 10.1109/MIC.2009.52
23. Zweben SH, Edwards S, Weide B, Hollingsworth J (1995) The effects of layering and encapsulation on software development cost and quality. IEEE Trans Softw Eng 21(3):200–208

CONVERGENCE: Extending the Media Concept to Include Representations of Real World Objects

Nicola Blefari Melazzi

1 Introduction

One of the key enablers of today's media revolution has been the emergence of broadly accepted multimedia standards – in particular, the standards produced by the MPEG community. A large part of the current MPEG standards are centered on "classical" needs of the media industry. Today, however, the distinction between media products and other digital resources is increasingly blurred. In both cases, users need information resources to be constantly available, up to date, and sharable. They also need to synchronize information over a complex mesh of devices to guarantee its integrity and to exert control over the way it is accessed and used. In both cases, they need the ability to "revoke" information that is no longer true or valid or which they no longer wish to make available to others. In fact, these needs extend to digital information describing Real World Objects (RWOs) – products, companies, people, and locations that users wish to buy, meet, or visit.

In brief, there is a convergence between the requirements of media users and those of other consumers and producers of digital information.

Several of these needs are addressed by existing MPEG standards. MPEG-21 already defines standard ways of providing meta-information and standard ways of describing the content and structure of complex "Digital Items" [1]. The CONVERGENCE framework aims at extending the ability to manage and trade digital objects to a broader range of digital objects, including descriptors for RWOs and to extend the possible actions that we can perform on digital objects. We call these new, more adaptable, digital objects Versatile Digital Items (VDIs).

To achieve these goals, we will start from the Digital Item Declaration standard (ISO/IEC 21000–2) from MPEG, extend it to cope with requirements derived from

N.B. Melazzi (✉)
University of Rome, Tor Vergata, Rome, Italy
e-mail: blefari@uniroma2.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_13,

new application scenarios, and develop a complete environment to handle VDIs. More specifically, the CONVERGENCE framework aims at:

1. Handling new needs associated with the emergence of an "Internet of Things." In the Internet of Things, RWOs are enhanced with machine-readable digital identifiers, such as barcodes and RFID tags which link to additional information describing their properties, dynamic state, and context. The CONVERGENCE framework will define common mechanisms for handling descriptors for different classes of RWO. Such mechanisms will lay the basis to enhance existing information services that reference RWOs, by enabling a huge potential for integrating services on the basis of common data items. Example services include e-auction sites, such as E-Bay, location based services, such as "Friend Finder," and even multiplayer games.
2. Being intrinsically dynamic. In today's networked world, the content of the information exchanged between providers and consumers is increasingly volatile. Catalogs, CVs, technical specifications, descriptions of locations, or metadata for physical items such as books refer to a world that is continuously changing. The producers of information need the ability to update the information they have released and consumers need mechanisms to check if a digital resource is up to date, to request an update, and to select between several versions of the same item. Automatic updating of distributed information will thus be one feature of the CONVERGENCE framework. The update mechanism will include mechanisms allowing producers to "push" updates to consumers and enabling "consumers" to pull their own updates.
3. Supporting "digital forgetting," i.e., the guarantee that content generated at one period of a user's life does not come back to haunt him/her. To meet this need, the CONVERGENCE framework will provide mechanisms allowing users to "unpublish" VDIs. A second mechanism will allow users to define expiry dates for whole VDIs or for specific items of information. Together these mechanisms will allow sites and services to perform automatic garbage collection deleting expired information. In this way, the CONVERGENCE framework will act as an enabling technology for site owners and regulators wishing to provide or enforce protection for user privacy.
4. Providing novel security and privacy mechanisms supporting the needs of all users of VDIs (information providers and information consumers). The design will take account of the differing needs of professional information providers, whose products are designed for use by the general public, often for a fee, and non-professional providers, who will be enabled to express privacy requirements concerning the use of specific kinds of information and exert expiry dates for the validity of the information. An example is the protection of the privacy of users making queries about VDIs or requesting updates of VDIs. Similarly to a DNS server, which, by knowing the queries, knows the sites visited by users, it is technically possible that the creator or the maintainer of a VDI would know all the actions and therefore the interests of people requesting or updating her VDIs, a feature which may be helpful in some scenarios or has to be avoided in other scenarios. Additional mechanism will protect the privacy of information

consumers, preventing unwanted profiling of their searches, requests for updates etc. Protecting the user information at the source (i.e., the VDI) is better than delegating this function to applications and makes the big difference between trustworthiness and trust.
5. Incorporating metadata technologies from multimedia standards and Semantic Web technology, providing a homogeneous way of searching and handling digital items of all kinds. The framework will be integrated closely with RDF (Resource Description Framework) technology and will support ontologies for describing metadata. To be compatible with other popular solutions, it will also include a standard mechanism for folksonomy tagging, allowing users to "tag" (or search for) items of digital information across multiple web sites, services, and applications.
6. Facilitating users to contribute to the production of media. Many digital consumers have become producers of digital media, which they increasingly share with other users in Web 2.0 applications. The CONVERGENCE framework will provide users who generate their own content with a globally unique identity for their work together with standard mechanisms for bundling it with meta-data. In addition, the CONVERGENCE framework will provide users with a way to easily share information between heterogeneous devices and with other users.

2 The CONVERGENCE Framework

The framework consists of four main elements:

1. VDI: The basic unit of transaction and exchange within the framework.
2. Middleware: Software allowing CONVERGENCE and third party applications to define and perform operations on VDIs. This includes network entities implementing the distributed CONVERGENCE logical architecture.
3. Tools and applications: Software tools and applications, based on the middleware and allowing end-users to manipulate VDIs.
4. A Community Dictionary Service (CDS): An RDF-based semantic framework for ontology sharing.

Figure 1 shows the relationships among these elements and the way in which the CONVERGENCE approach extends the existing MPEG-21 framework. We describe the first three elements of our proposed framework below.

2.1 Versatile Digital Items

Within the framework, the fundamental unit of distribution and transaction will be the VDI. The VDI is technically an extension of the MPEG-21 concept of a Digital Item. It is meant as a very general data format holding information related

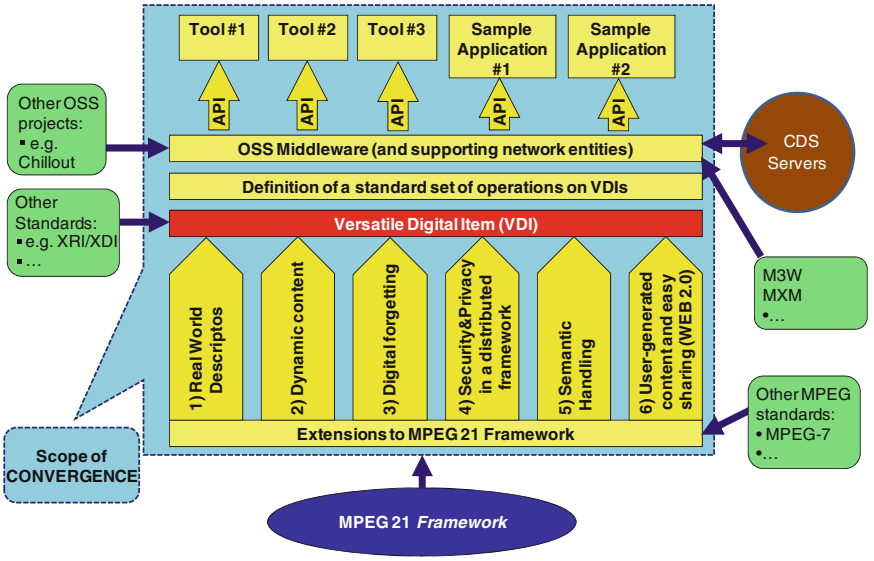


Fig. 1 CONVERGENCE framework

to any virtual or physical item. Like MPEG-21 DIs, VDIs will bind together meta-information (describing the content and structure of the item) and resources (other VDIs, audio, images, video, text, complex descriptors of RWOs etc.). Like MPEG-21 DIs, they will include identifiers (e.g., Digital Item Identifiers as specified by ISO/IEC 21000–3) and statements expressed in a Rights Expression Language (REL) allowing users to define the ways in which they can be used. Unlike DIs, they will be deliberately designed to take account of the dynamic, infinitely variable nature of the information world beyond classical multimedia.

VDIs will be designed to provide broad support to references to RWOs such as products, locations and people, making it possible to store, synchronize and certify RWO descriptors in the same way as conventional media objects. In addition, CONVERGENCE will define the metadata included in MPEG-21 DIs to support the new functionality provided by the framework, for instance, synchronization and deletion of VDIs that have already been released and semantically enhanced search and interpretation of VDI metadata.

2.2 The CONVERGENCE Middleware

Like MPEG-21 DIs, VDIs are self-contained: all the information necessary to handle a VDI is contained in the VDI itself. This is analogous to the situation in connectionless packet-switching protocols where all the information necessary to transport packets is contained in the packets. The CONVERGENCE partners

believe that, just as the state-less nature of IP packets have contributed to the success of the IP protocol, so, self-contained VDIs can contribute to the success of the CONVERGENCE framework.

But a data format on its own is merely a facilitator for operations on DIs. These operations have to be provided by a separate layer. The MPEG community has understood this need, and has already published the Committee Draft of MPEG eXtensible Middleware (MXM, ISO/IEC 23006) [2]. MXM's goal is to specify an architecture and a set of standardized APIs for handling a variety of MPEG technologies, including DIs. The goal of the CONVERGENCE middleware is to play a similar role with respect to VDIs.

The CONVERGENCE middleware will provide APIs to dynamically define and encapsulate new classes of content and related meta-information, to create VDIs packaging different classes of information resource, to guarantee their security and privacy and integrity, to name them, to support semantic interpretation of metadata and tags using CDSs, to search for them, filter them, read and write their attributes and content, adapt them for use on different machines, copy them, test their validity and efficiently synchronize them across multiple machines.

Compared to MPEG-21 and MXM, the CONVERGENCE middleware will significantly extend opportunities for interactions among different actors along the value chain.

Figure 2 describes a classical value chain for an MPEG-21 DI, in this case a DVD. Actor 1 creates the DI, defining the title of the video, Rights and Permissions etc. (meta-information), and inserting video tracks (resources). Actor 2, a dubbing company, adds sound tracks (additional resources). Actor 3, the censorship board, adds information on the video's certification (new meta-information). The chain continues until DI reaches the final Actor (i.e., the consumer). The only way to edit or read a DI is locally, when it is physically available on a given actor's premises. Thus, a downstream actor has no way of modifying choices that have been made upstream. And once the DI has been delivered to the next actor down the chain there is no way the upstream actor can update the information he or she has provided.

In CONVERGENCE, by contrast (see Fig. 3), VDIs are living objects that can always be modified by any authorized actor of the value chain. In other words, CONVERGENCE extends the life of the value chain beyond the actions performed in the first delivery phase (i.e., when the VDI physically reach the consumer for the first time); as a consequence, CONVERGENCE enables new forms of business models. What is more, actors can operate on VDIs remotely, even when they are physically located on the premises of other actors. In brief, we can consider the CONVERGENCE value chain as a "bus" allowing any station attached to the bus (any actors) to interact with any other station.

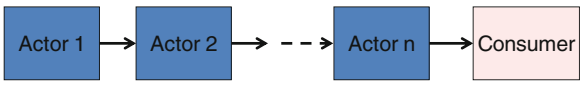


Fig. 2 Sequence of actors involved in a classical value chain

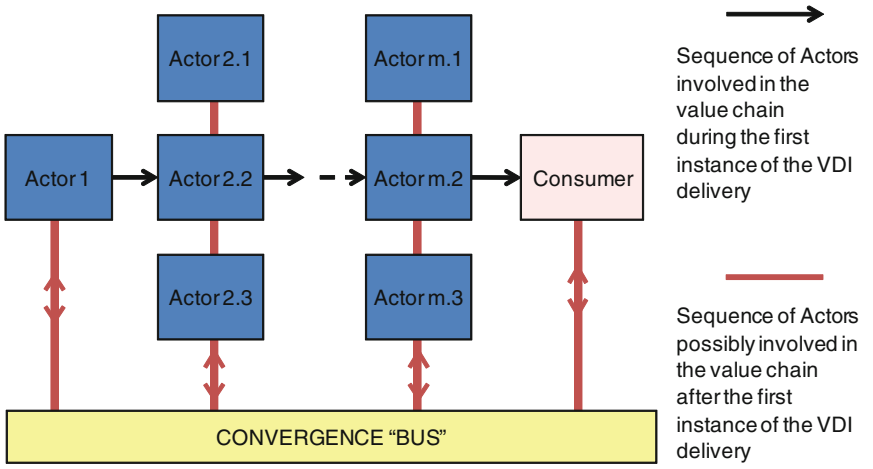


Fig. 3 Sequence of actors in the CONVERGENCE value chain

The CONVERGENCE "bus" enables a broad range of operations that are not straightforward with conventional DIs. For example:

- Distributors (with appropriate rights) can remove certain portions of a movie they do not wish to distribute to their2 market
- Producers and government agencies can inform customers who have bought a product (and its associated VDI) that the product has been found to be dangerous and should not be used
- The author of a CV (distributed as a VDI) can update the copy she has sent to perspective employers
- The buyer of a movie in English can sell it to an Italian user, acquiring and inserting the Italian audio track
- A team of doctors can handle a patient's medical record (a VDI) collaboratively with each doctor updating the record locally and propagating the updates to the other copies

As regards the technical approach for middleware, below we present some key features of our proposal.

2.2.1 Middleware Model

Dynamism of information is easily managed when the involved resources are always available on the network. Nevertheless, we cannot assume that the devices storing VDIs will always be connected. To remove the "always-on" assumption, CONVERGENCE will develop an Asynchronous Middleware that "remembers" when an operation has been requested and ensures the request is executed when networking facilities allow it, that is when the middleware detects a specified event.

To support this functionality, the middleware will adopt an event-based approach, implemented via Publish/Subscribe technologies.

In the publish/subscribe paradigm, "subscribers" are interested in particular events generated by "publishers." Depending on the way subscribers express their interests, we can distinguish two main classes of publish/subscribe systems, namely topic-based and content-based systems. In topicbased systems, subscribers join a group interested in a specific topic of interest, and publications on the topic are broadcast to all group members. In other words, publishers and subscribers have to explicitly specify the group they wish to join. In contentbased publish/subscribe systems, on the other hand, matching of subscriptions and publications is based on content and no prior knowledge is required. These systems also have the advantage that they allow users to set up permanently active queries referring to large numbers of different "terms." For a system such as CONVERGENCE, which provides descriptors for highly heterogeneous RWOs, this is an extremely important feature. A second distinction is between centrally managed and distributed systems. The major disadvantage of centralized systems is their lack of scalability and faulttolerance. Distributed systems overcome these limitations. Given the enormous scale of the "Internet of Things" that CONVERGENCE seeks to support, the CONVERGENCE middleware will adopt a distributed content-based strategy.

2.2.2 Networking Approach

The CONVERGENCE publish/subscribe mechanism will exploit distributed P2P overlay infrastructures. More specifically, the CONVERGENCE middleware will implement a content-based publish/subscribe system based on a Distributed Hash Table (DHT). DHT approaches to network look-up are described in [3–5] and have been shown to be very effective in performing scalable, fault tolerant resource lookup on large peertopeer networks. Within the DHT framework, CONVERGENCE will implement content-based publish/subscribe mechanisms specifically designed to operate on VDIs. The approach chosen will allow CONVERGENCE to design effective subscription-peer mapping schemas to store subscriptions into the network, to define predicatebased expressive query semantics to efficiently match and deliver events to subscribers, and to ensure a uniform distribution of load among peers.

To improve scalability, P2P connectivity in CONVERGENCE will be organized as a hierarchically-structured federation of organizations and communities of interest (e.g., a virtual marketplace). In this way, a limited set of nodes will be responsible for managing the federation for performing DHT-based publication/subscriptions and searches in a hierarchical way.

The DHT approach will enable the CONVERGENCE middleware to perform semantically enabled search for VDIs and to react to specific VDI related events. However, the middleware also needs functionality for remote handling of the VDIs that the DHT lookup has found. CONVERGENCE will thus exploit invocation protocols to communicate between the two middleware entities concerned. Candidate

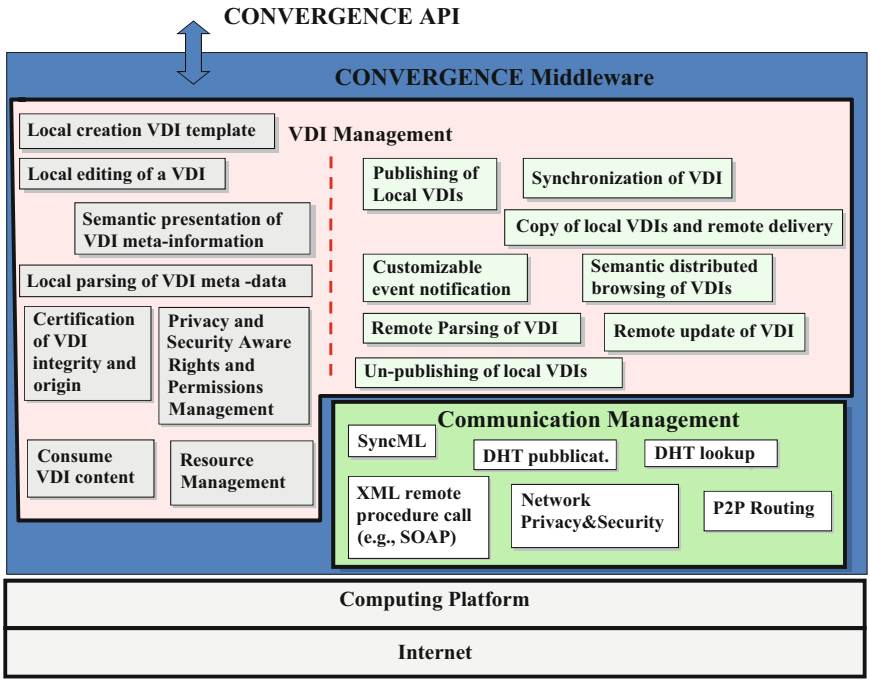


Fig. 4 Middleware functionality

solutions for implementing this functionality include the XML Remote Procedure Call, JSON-RPC, SOAP as well as REST protocols. Figure 4 summarizes the functionality to be provided by the CONVERGENCE Middleware.

The middleware will offer two classes of functions: (i) VDI management, enabling operations on local and remote VDIs; (ii) Communication management, supporting distributed communication among middleware modules. The dashed line in Fig. 4 separates VDI Management modules that require the Communication Function (on the right) from those that do not require such functionality (on the left). A subset of these functions will be exported to the upper layer by means of standardized CONVERGENCE APIs.

2.3 Tools and Applications

To demonstrate the usefulness and usability of the CONVERGENCE framework for developers and end-users, CONVERGENCE will design three general purpose tools that exploit the CONVERGENCE middleware to operate on VDIs.

- VDI Creator: A tool which takes a collection of digital information and wraps it into the VDI format

- VDI Manager: A tool to operate on VDIs, where the operations allowed depend on the invoker and the rights defined in the VDI
- VDI Search: A tool to find VDIs in a worldwide distributed storage infrastructure

A key goal will be to offer an effective implementation of user security, rights and privacy, defined inside the VDI. The operations allowed in VDI Manager will depend on the rights of the invoker defined in the VDI. For example, a typical publisher of VDIs would have the right to push updates to subscribers, change the metadata etc.; a typical subscriber would have the right to view VDIs, create links to them and pull updates from the publisher. There are many possible variants to this model. It is quite possible, for instance, that an online artist might authorize other (unknown) artists to modify and redistribute her work.

The CONVERGENCE framework will provide many different ways of creating new applications. The simplest and most flexible way will be to use CONVERGENCE tools in combination with other standard software such as a browser. Alternatively, application software can use scripts or make direct calls to the APIs provided by the middleware. These will take the form of Web services, providing loose coupling with operating systems, programming languages and other related technologies, and facilitating the creation of Mashups.

3 Use Cases

The CONVERGENCE framework and the VDI concept offer novel features and functionality of clear benefit to end-users. Below we describe use cases that illustrate these features.

3.1 Dynamic Logbook

This use case illustrates the use of VDIs as a "dynamic logbook" for a technological device. In this example, the device is a handset but it could also be a washing machine, a dishwasher, a digital camera, a car. The functionality provided would be useful in a broad range of business to consumer and business to business applications.

Every handset that Big Telco sells to its customers comes with a pre-installed "digital logbook" – a VDI associated with the serial number for the phone. The logbook stores details of the date of purchase, the warranty, the firmware, pre-installed software, a user guide, and a maintenance record; when the user purchases a SIM from an operator, the logbook can be updated to include the SIM code and the owner's tariff plan (another VDI). Optionally, if the user agrees, the VDI may also contain personal identity information, which she can make available to some users and not to others.

Copies of the logbook are stored at several different locations. The manufacturer has a copy, Big Telco has a copy, the distributor has a copy, the operator has a copy, and a copy is pre-installed on the phone. When a customer buys and takes the phone home, she waves it over a reader and transfers a copy of the logbook to her local PC.

From this point on, each of the actors along the value chain can change the parts of the logbook for which it is responsible. The manufacturer can update the user manual. The phone operator can update the information on the user's tariff plan. If the user sells the phone to a friend she can change the ownership data. The different actors can use CONVERGENCE search facilities to identify phones with certain characteristics, and send messages which will appear to their owners (even if the identity of the owner is not known). Thus, a manufacturer who discovers a dangerous battery fault, can send a recall message to owners of phones with the defective part; an operator who has cut charges for certain calls can target owners on a certain tariff plan; a distributor can warn a customer when her warranty is about to wear out.

When a customer experiences problems with a phone, she sends a copy of the logbook to the call-center operator. The operator feeds the logbook to special simulation and diagnostic tools, which rapidly reproduce the problem and propose a solution.

It is important to stress that the functionality exemplified in this use case (i.e., a communication link between the producer of an object and the user(s) of that object), could be very useful in numerous situations.

3.2 Digital Forgetting – Automatic Garbage Collection

In the past when we wrote a diary or a letter to a friend, when we stayed up till the small hours talking politics over wine or beer, what we wrote and said remained in private drawers and the memory of friends. While today, when we write a blog or an email, or "chat" with a friend, we cannot know who will read it 20 years later. To fully enjoy the freedom offered by modern communications, to use it to the full, we need a right to "digital forgetting." In this use case, we illustrate how CONVERGENCE can contribute to this goal.

Jaap is 20, very bright, cynical, and a natural rebel. When his father comes home with his friends from the board room, he listens to their conversations. He hears how easily they talk about billions of dollars, how indifferent they are to the effects of what they are doing. But what strikes him is the way they talk, the tone, the language. Jaap has a good ear and a quick pen. His blog, "The Bank," – is a huge success – not least among his father's junior staff who think it is very true and very funny…

But now many years have passed and Jaap needs a job. He is no less bright, no less cynical and no less of a rebel than before, but it would be better if future employers did not know. "The Bank" is only one of his blogs. Going through everything he has ever posted takes 2 days, and for some of the postings there is no remedy. He can't remember the web sites address, or he's lost the password,

or the service no longer exists. Fortunately, Safeblog – his blogging service – is CONVERGENCE-based. VDI Search gives him his old postings. Thankfully he clicks on the "unpublish" button. A second later, the postings are flagged for deletion, not just on Safeblog, but on every other site that has a copy of the postings: friends' home pages, search engines that have cached copies, even the WayBack Machine. What happens next will depend on the policy of the site, but most respectable sites respect users' wishes, and in the Netherlands, it is now a legal offence for companies to use "unpublished" digital information on job candidates.

Jaap is relieved that he has managed to unpublish "the Bank," but he also knows that he has run a huge risk. For his next very personal postings, he sets an expiry date just 3 months after publication. CONVERGENCE makes him feel safe.

3.3 VDIs in a Store

Goods entering the store will be stamped with a RFID tag and associated with a VDI. The VDI will store information about the item's type, model/version/features/firmware. The VDI will also store additional information, such as stock entry date, expiration date (if applicable), inventory number, or location in the storage area. This will make it easy to find the item via simple search processes.

Once the item is on the shelf, employees will be able to scan the RFID tag to obtain the associated VDI and use it to check/update the shelf stock or consult the inventory. Customers will be able to interact with the VDI on the spot, either by using their own CONVERGENCE-enabled smart-phones or by using an interactive display mounted on the shelf. They will be able to check the item's features, query for additional information, and search for other compatible/complementary items, thus enabling cross-selling.

Once the item is checked-out, the information contained in the associated VDI will be updated in several ways. The store internal information (ex. inventory number) will be deleted or encrypted (in case of item returns), the receipt number will be stored, and the ownership of the VDI will be changed. Other possible operations include notification of sales to the producer/supplier, and automated generation of warranties, stored directly on the VDI. Such warranties avoid loss of the warranty certificate by the customer.

Once home, customers will be able to search the web for further products details, manuals, new versions, firmware updates, product-related forums or users that bought similar items, related to the VDI. If the user brings the item to a CONVERGENCE-enabled repair shop, the technicians will use the associated VDI to check for warranty issues or search for technical information (schemes, parts details, repair methods, etc.). The digitally signed details of the intervention will be stored in the VDI, thus keeping a complete history of the item lifecycle.

4 Conclusions

We believe that the CONVERGENCE concept can be a fundamental enabler for the Internet of Things as it provides both a common and flexible format for representing information related to RWOs and sophisticated functionality to handle such representations.

In addition, we remark that all the information necessary to handle a VDI is contained in the VDI itself. This makes it possible to handle networking functions at the "VDI layer," i.e., at the application layer, following the trend of overlay communications. In this setting, networked VDIs can support innovative scenarios and applications, helping boosting Internet of Things exploitation.

Acknowledgments This position paper is based on a project proposal. The author thanks all the partners that contributed to the proposal and acknowledges specific contributions and suggestions to this paper from the following individuals: Maria Teresa Andrade, Leonardo Chiariglione, Andrea de Polo, Heinrich Hussmann, Dimitra Kaklamani, Stelios Pantelopoulos, José Ribas, Carsten Rust, Peter Stockinger, Mihai Tanase, Richard Walker, Andrea Detti.

References

1. ISO/IEC 21000 (MPEG-21) series of International Standards, ISO/IEC 21000–2:2005. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41112
2. http://www.chiariglione.org/mpeg/working_documents/explorations/mxm/MXM-Reqs.zip
3. Tam D, Azimi R, Jacobsen H-A (2003) Building content-based publish/subscribe systems with distributed hash tables. In: Proceedings of the 1st international workshop on Databases, Information Systems, and P2P Computing (DBISP2P), Berlin, Germany, Sept 2003
4. Terpstra WW, Behnel S, Fiege L, Zeidler A, Buchmann AP (2003) A peer-to-peer approach to content-based publish/subscribe. In: Proceedings of workshop on DEBS, San Diego, CA
5. Yang X, Hu Y (2007) A DHT-based infrastructure for content-based publish/subscribe services. In: Proceedings of the 7th IEEE international conference on peer-to-peer computing (P2P'07), Galway, Ireland, 2–5 Sept 2007

Service Oriented Middleware Solutions for Emergency Communication Networks

Fabrizio Ronci and Marco Listanti

1 Introduction

In response to civil protection and public safety emergency situations, a number of operators, decision-makers, and institutional and commercial service providers are usually supposed to cooperate in order to assist involved population and environment, to overcome the crisis and to start reconstruction. As far as conventional, and possibly inadequate, communication services, basically relying on radio voice calls, yielding to a wide gamut of real-time, interactive, and multimedia data-oriented information flows, new viewpoints on network architectures arise from integrating available Information, Communication, and Media Technologies (ICMTs) [1].

Accepting the reasonable surmise that a modern and effective Emergency Management Network should be designed as a Next Generation Network paradigm result, it comes out that it has to be conceived, integrated, and operated over an all-IP networking platform, with an easily increasable number of networked application domains, each of them able to cope with a different civil protection stakeholder service or class of service requirement [2].

Nevertheless, such an extent of integration has to be guaranteed over a plethora of different technologies, ranging from narrowband radio transmissions to wireless incumbent and emerging standards, from traditional network architectures shaped mainly by simple interactions of homogeneous devices to extremely articulated networking schemes, that need to unify and bridge diverse communication resources and components [3].

In few words, well explained by Fig. 1, this networking platform, dedicated to planning for reacting to large disasters and to emergency preventing, managing, and overcoming, has to be implemented as a heterogeneous, pervasive, and distributed one, where computational clusters, environmental sensor networks, desktop workstations, wired broadband and wireless networks, wearable and pervasive devices,

F. Ronci (✉) and M. Listanti
INFOCOM Department, University of Rome "La Sapienza", Via Eudossiana, 18-00184 Rome, Italy
e-mail: ronci@net.infocom.uniroma1.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_14,

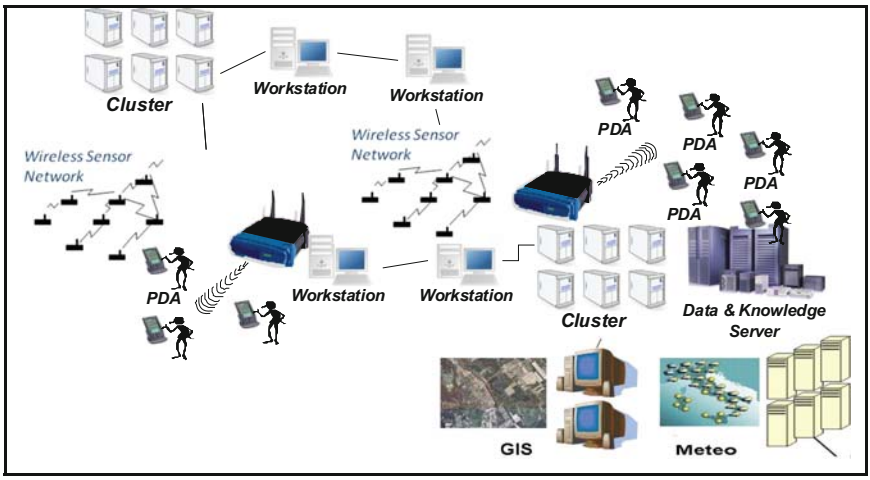


Fig. 1 Heterogeneous, pervasive and distributed platform [4]

mainframe computers, and GIS/Meteo service data farms constitute an unique control, communication, and elaboration support to operators, fully virtualizing different technologies, protocols, and systems [4].

All of these network resources and components, including today's mobile devices that as per performance are a lot more powerful than those of the early days, are able to host very sophisticated and versatile software [5]. This fact not only enables the usage and exploitation of middleware solutions to integrate knowledge [4, 6], but also, in our vision, enhances reliability, provides transparency, and guarantees scalability with respect to physical, link, routing, and transport technologies and schemes: the latter issue is the main focus of this paper and of our incoming work.

More precisely, in this article, the main novelty is represented by the assessment that standard middleware can provide a kind of feasibility study in applying Publish/Subscribe paradigm to Emergency Communication Network, along with an experiment implementation proving that, by the same means, applications such as interactive delay-sensitive communications can be supported, which is not so intuitive using Pub/Sub scheme middlewares.

Hereafter some sections follow. In Sect. 2, a sketch of Italian National Civil Protection Service implementation in a case of major disaster will be presented. In Sect. 3, a number of related works and projects, with stress on MIUR-FIRB INSYEME Project and an its use case, will be resumed. In Sect. 4, operating principles and main structures of a standard specified middleware, OMG DDS, are introduced, along with some discussion about its applicability and profitability to Emergency Communication Services. Section 5 follows as exposition of our first service solution implementation, namely an emulation of SIP for voice-calls-over-IP-establishment by means of OMG DDS. Section 6 contains concluding remarks and future possible directions in related research.

2 Organizational Framework

The Italian Civil Protection National Service, the Italian organization devoted to prevent and to face with emergency events, involves a large number of actors: government central and local authorities, operational intervention forces, civilian volunteers associations, etc. Because of unparalleled territory features, Italy gathers many risks, different by nature, but all contemporarily present and with comparable probability of occurrence: hydrogeologic, volcanic, seismic, sanitary, humanitarian, anthropic risks, risk of wild fires, as well as similar risks in bordering or remote countries.

Among them the hydrogeologic risk is second only to the seismic one, in terms of loss of human lives and of damages caused to structures, but can be considered the prevailing one in terms of composition of probability of occurrence, spatial distribution and social impact.

Italian law recognizes local authorities, such as Regions, Provinces, and Municipalities, as legislative, planning, and implementation bodies for civil protection actions, as well as government nominees in-loco, such as Prefects, as responsible for control and coordination. The Head of Municipality, the Mayor, is the first responding civil protection authority, in his/her own territory. Regions are responsible, among other duties, for offering central state agencies, first of all the Civil Protection Department, all the means necessary to exert their competencies in large emergencies.

In [1] a good sketch of emergency management command and control chain in Italy is exposed; there, a number of C2 centers are individuated:

1. SOUR (Regional Unified Room for Operations)
2. CCSs (Centers for Coordination of Aids)
3. COMs (Mixed Operational Centers)

In the following, we will refer to the same entities with only two exceptions: we will consider also the COCs (Municipality – in Italian *Comunale* – Operational Centers) and instead of the SOUR, because of considering a large event, the DICOMAC (Command and Control Direction), which usually is built up within the SOUR by central Civil Protection Department staff personnel.

These authorities can take advantage from the availability and work of some Operational Structures: the National Corps of Fire Fighters, the fundamental component of the Service; but also Armed Forces, State Police (PS) Corps, Local Police (PL) Corps, etc., and, last but not least, Associations of Volunteers (OVL).

3 Related Work and Projects

In the broader context of Public Protection and Disaster Relief (PPDR) Communications, a key point of reference is constituted by results from MESA Project: it is an international effort, driven by ETSI and TIA, aimed to individuate general

public safety and emergency services and relevant communication architecture or structure, along with common components in that architecture. Most fundamental MESA deliverables are dedicated to assess specific user requirements [7], to categorize classes of sub-networks in the overall architecture and classes of connections among these sub-networks [8] and to deepen into some interoperability issues about devices interworking [9].

From MESA point of view, there might be classified four (plus one) types of communication networks for civil protection communications, services and applications:

(a) *Personal Area Network* (PAN)
(b) *Incident Area Network* (IAN)
(c) *Jurisdiction Area Network* (JAN)
(d) *Extended Area Network* (EAN)
(e) *Ancillary Wireless Network* (AWN)

WORKPAD, an European research Project, identifies, to its aims, an architecture formed by a front-end and a back-end layer [1]. Within WORKPAD it has been developed also RESCUE, an open source middleware for service oriented communications in mobile disaster response environments [5].

A series of works have been dedicated to wireless communications applied in the emergency response context [3, 10, 11], where Wireless Mesh Networks, Mobile Ad-hoc Networks, and Wireless Sensor Networks as well, have to be considered not abdicable classes of components in Emergency Management Architectures.

Besides this, importance of considering middleware and software interacting components based on middleware was recently assessed [12, 13].

Middleware, on its own, has been recognized as a convenient method to provide services that hide, in favor of applications, the complexity of the underlying network, which is a typical situation in heterogeneous Emergency Communication Networks [14]; meanwhile Publish/Subscribe paradigm [15] appears to fit nicely and gracefully in a communication model, typical of the case as well, where information flows are mainly driven by users, selecting all and only data topics or attributes they are interested in.

MIUR-FIRB INSYEME (INtegrated SYstems for EMErgency) is an Italian research Project concerned to realize a robust and flexible communication structure that facilitates coordination among human operators, systems, and a variety of fixed and mobile equipment in an environment featured by high level of ubiquity, heterogeneity, dynamicity, and adaptivity. Some characteristics are considered crucial in INSYEME architecture, such as QoS-driven approach at all layers, performance and functionality adaptivity, context and location awareness, and resources visibility to users and applications [16].

INSYEME architecture is thought as formed by the integration of three key elements: a communication infrastructure, a distributed processing environment, and a middleware layer for the integration of knowledge. In other words, the elements constituting the system architecture form a three-layer structure, where the communication infrastructure handles wireless communications among operators and connects sensor networks with the control centers [6]. The information knowledge

layer forms an intermediate layer aiming to correlate all the information exchanged by the communication infrastructure, and the grid computing layer manages a distributed processing, exploiting the communication infrastructure to support decision and data mining applications.

Since the first drawing of the INSYEME Project statement of work, a Mobile Grid approach was stated to be pursued, aiming to implement a novel combination of the disciplines of pervasive computing and of high performance and adaptive programming models, in particular by means of an adaptive and context-aware programming model. This model has the declared objective of providing adaptation mechanisms as completely mapped at the application level, while the supporting environment only delivers resource information through reflective mechanisms [4].

Present article, re-affirming the validity of the above mentioned programming model approach, points to explore some not well-enlightened mechanisms among system layers, e.g., cross-layer interactions among applications and services, middleware layer and well known or novel protocols operating at underlying layers. This effort has been accomplished considering standard middleware implementations, with the perspective of feeding programming model development with a kind of feasibility study.

For the INSYEME Project, 10 Scenario Clauses have been generated [17]. Two out of these ten clauses are related closely to the command and control chain, taking in account also a condition of high variability level. Out of the remaining eight clauses, one has been subdivided into two sub-clauses, in order to take advantage from one of them as an explicative sub-scenario. This sub-scenario has been used throughout this paper.

It describes a population evacuation situation forced by a flooding:

(a) Many operational forces (Fire Fighters, Local Police, Local Volunteer Associations) cooperate, coordinated by a COC and the parent COM
(b) Parent CCS exercises a Mobility Check Point (CP)
(c) There are PANs, as the MESA concept, including various kinds of sensors
(d) There are control centers and operational forces at an upper hierarchical level, exchanging information with operating structures on-the-ground
(e) There are some databases, with different level of security assurance on them

4 OMG Data Distribution Service

Object Management Group Data Distribution Service (OMG DDS) utilizes the well-known Pub/Sub paradigm to achieve highly efficient information dissemination among multiple publishers and subscribers that share interest in so-called Topics. Topics are basic data structures expressed in the OMG IDL language. This specification also includes a QoS framework allowing the middleware to match requested and offered (RxO) Quality of Service parameters, on the basis of attributes like reliability, ordering and urgency.

Selection of such a specification for our work originates from considering integration and interoperability requirements: data centricity, in respect of actors and

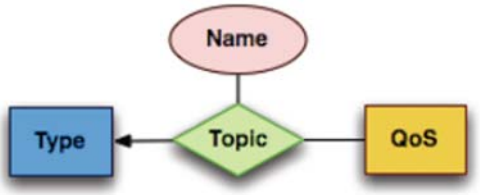


Fig. 2 OMG DDS topic general format

technologies, and transparency, up to layers as higher as possible. OMG DDS offers such features as it allows methods directly at data level, through an interoperability protocol, and compatibility among technologies up to the transport layer.

OMG DDS also presents, at various degrees of extent, all the characteristics already considered mandatory for middleware employable in Emergency Management Networks [14]: Distributed Data Storage, Pub/Sub Service, Contextualization Service, Localization Service, Neutrality about complexity of underlying network(s) and, above all, Affordability of extreme heterogeneity in terms of devices, bandwidth, connectivity and robustness.

In few terms, OMG DDS enables the realization of an emergency communication network following the *System of Systems* paradigm.

OMG DDS [18] is a Publish/Subscribe middleware that operates upon a Global Data Space; in perspective, in a security and service isolation context, each Global Data Space represents a single administrative domain.

Central concept in OMG DDS is that of Topic (Fig. 2): this is a data structure, described in terms of an unique name – in the Global Data Space – and a type, but accompanied by QoS attributes, such as, among others, Latency_Budget, Ownership, Liveliness, Reliability, Durability, etc.

Topics are typically objects, and their instances are univocally identified by the middleware in a single data space by means of a key, taken from one of the data type elements: that way, right information can only and suitably flow from intended publishers to interested subscribers at exact moments, synchronous or asynchronous, and on the basis of a stated QoS policy. In fact, matching between publishing and subscribing participants in a data domain is carried out at all data structure components, name, type, and QoS policy.

It has been recognized that OMG DDS offers means to cope with emergency response system requirements (Flexibility and Adaptivity, Computational and Energy Constraints, Service Availability Multiple Levels, Intra-Network Preemptive Services, Compatibility, and Consistency with existing standards) as well as with first responder user requirements (Mobility Support, Limited End-to-End Latency, Time Jitter Limitation, Throughput Maximization, Local Data Distributed – and Reliable – Storage, Centralized DB Access).

Specifically, OMG DDS employment in emergency situation appears to guarantee critical behaviors, such as:

- Universal Access to Information
- Continuous Adaptation to Changes

- Standard QoS Policies and Mechanisms
- Efficient and Scalable Data Distribution
- Predictable Resources Usage

Nevertheless, Pub/Sub middlewares in general, and OMG DDS as well, do not appear easily applicable to provide support to delay-sensitive interactive communications, namely, voice calls or VTCs, which yet represent a must-have requirement in every civil protection communication system design.

Our research moved to analyze how to, if possible, emulate some well-known protocols and standards with OMG DDS structure, and to assess if such an emulation is able to reduce information overhead exchange among network nodes.

5 Experiment Implementation Details

In Fig. 3, a possible architectural implementation, *compliant to MESA specification*, is devised for communication services within the use case described in Sect. 2. Databases are supposed to be distributed and accessed, timely and reliably, only by authorized operators.

As intuitively could be seen, OMG DDS features give a correct answer to these requirements, but for voice calls. Specific scheme of solution for such a shortfall is the main novel contribute of this paper and it will be deepened hereafter.

Voice and video calls on Internet and, for extension, in an all-IP network, are usually set up and torn down by means of Session Initiation Protocol (SIP) [19].

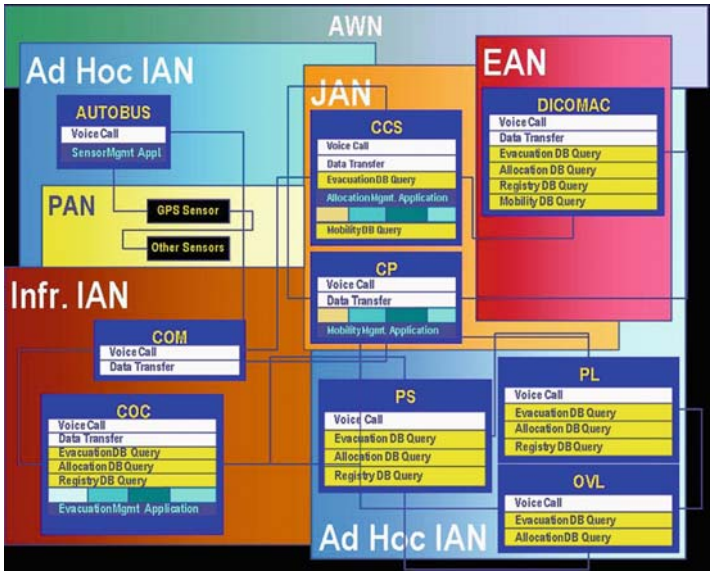


Fig. 3 Flooding use case MESA architecture correspondence

SIP is a protocol that can be classified as HTTP-like, meaning that it implements a Request/Response transaction model, where transactions occur between an User Agent Client and an User Agent Server. Other network elements – Proxy Servers, Session Border Controllers, specialized UASs as Redirect and Registrar Servers – cooperate through messages exchange to let UACs and UASs create, modify and terminate two-party (unicast) or multi-party (multicast) sessions consisting of one or several media streams.

Main SIP primitives or methods are:

(a) *Registration.* Entails sending a REGISTER request to a special type of UAS known as a Registrar. A Registrar acts as the front-end to a domain location service: this location service is then typically consulted by a Proxy Server.
(b) *Invitation.* When an UAC desires to initiate a session (for example, audio or video), it formulates an INVITE request, possibly forwarded by Proxies, to one or more servers (UASs), which will frequently need to query users about whether to accept the invitation.

Overall SIP behavior is that of a protocol strongly overheaded, especially when different and heterogeneous network segments are interconnected, even if dedicated to provide a sole service. Furthermore, because of Emergency Communication Networks are normally composed of wireless and wired parts connected through buffering interfaces and due to its Client/Server nature, SIP architecture is easily prone to: *bottleneck* and *single point of failure* disadvantages.

In our experiment implementation, both SIP functionalities have been emulated by OMG DDS constructs and developed applications; it turned out, too, that such an implementation is not affected by above reminded pitfalls, by virtue of the Pub/Sub scheme and the distributed communication platform.

5.1 Designed OMG DDS Topics

Our first objective was that of designing OMG DDS Topics able to emulate efficiently the two reminded SIP functionalities with only changes necessary to adapt to the middleware context: as we will show, this design led to a different, but simple and elegant, point of view in *Invitation* primitive.

In Table 1 the relevant IDL Module listing, containing REGISTER, INVITE and SERVICE Topics, is shown.

- *Topic REGISTER*. This Topic should have been representing the information that a user could be considered *Registered* onto the system.
  (a) For sake of AAA opportunity, a couple of a name string and a long integer, biunivocally determined, must be expressed and initialized for each user.
  (b) It appears to be not necessary assigning particular QoS attributes, but the Reliability (here the reference is to the well known primitive of *Reliable Broadcast* in Distributed Systems Theory).

Table 1 IDL description of designed topics

```
module Void_DDS {
    const long MAX_NAME = 100;
    typed string<MAX_NAME> nameType;
    enum State {free, busy};

    struct Register {
      long userID;
      nameType name;
    };
    #pragma keylist Resister userID

    struct Invite {
      long userID;
      long index;
      State state;
    };
    #pragma keylist Invite userID

    struct Service {
      long userID;
      nameType username;
      State state;
      long index;
      State state;
    };
    #pragma keylist Service userID
};
```

- *Topic INVITE*. This topic, unlike SIP mechanism where the caller *pushes* its request to UASs, should have been reflecting the fact that in a Pub/Sub scheme each participant announces, i.e., publish, its availability to be called and callers *pull* that information in order to contact them.
  (a) INVITE Type is formed by the user identifier, as in REGISTER, by a pseudo-unique system index, containing both naming information and information about which port is accepting incoming calls, and by a state, meaning the user is free or busy in another call.
  (b) Here, QoS policy needs to consider data Durability too, and the Topic has been made Transient.
- *Topic SERVICE*. We needed this Topic to be designed to give the system means to keep track of transactions, i.e., settings up and tearings down single calls.
  (a) Obviously, SERVICE is a system data structure that collects all the fields from the other two Topics.
  (b) Its associated QoS mechanism relies on the more specified, thus constraining, policy imposed on those Topics, namely INVITE.

5.2 Developed Applications

Our solution was developed upon one of the OMG DDS implementations available today, PrismTech OpenSpliceDDS. We used also a designing tool from the same implementation suite. In Fig. 4 is reported the overall diagram for the developed informative domain.

It contains several elements:

(a) *Topics.* Besides the above three Topics, there is another one, derived from filtering SERVICE, necessary to manage asynchronous change – by means of OMG DDS *Listener* constructs – in users activity
(b) *Participants.* They represent the possible activities of a generic user and the activity of the system – note the latter could be distributed
(c) *Applications.* Among others, three application modules were written in Java from scratch

- *Agent* is the user module dedicated to interface with the main GUI accessed on the device by operators and to publish REGISTER and INVITE in the Global Data Space, e.g., in the administrative domain of Voice Call Service.
- *CallBoard* is the user module by which a caller subscribes another user availability to be called and eventually sets up and tears down the call; it manages also to publish changes in SERVICE Topics.

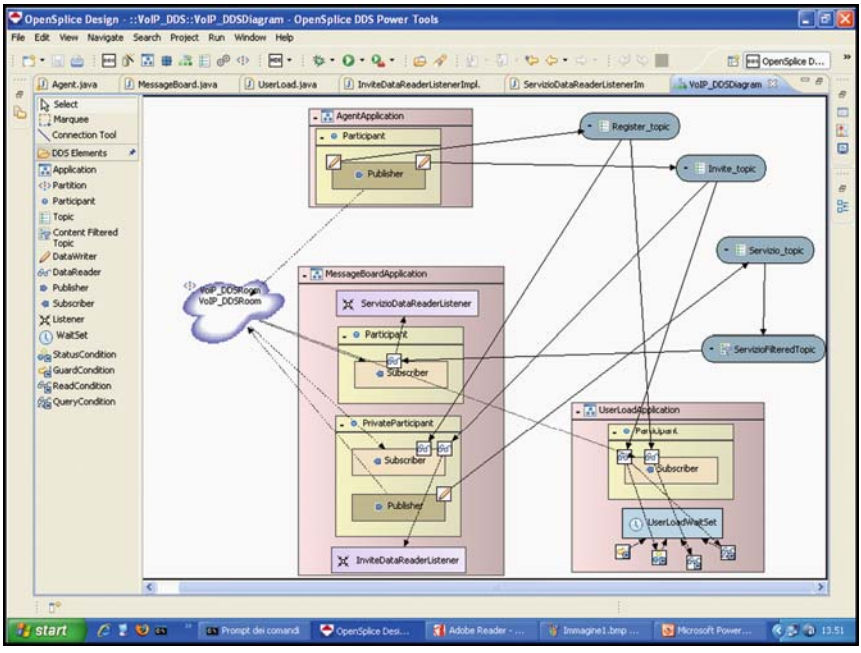


Fig. 4 Screenshot from the designing tool

Table 2 Experiment implementation distributed algorithm

```
systemLoad.start() ← subscribes INVITE and
                     REGISTER, enables WiatSets;
Agent.start();
Agent.publish(REGISTER) ← Systemload is running
                          and detects publishing;
Agent.publish(INVITE) ← Systemload is running
                        and detects publishing;
CallBoard.start() ← enables Listerners, SystemLoad
                    is running and detects
                    activity;
CallBoard.subscribe(INVITE) ←
            ← retrieves callable users and ports
            information, SystemLoad is running
            and detects subscribing;
CallBoard.select(User+Port);
    . . .
```

- *SystemLoad* is the system module that overviews synchronously – by means of OMG DDS *WaitSet* constructs – transactions among users and that reports, through subscriptions, changes in Global Data Space Topics.

Table 2 lists a simplified version of the distributed algorithm realized by the three modules; one should be thinking as *Agent* running on a user device and *CallBoard* on another one.

At the end of this algorithm, an operator is able to select another user to be called by means of a VoIP module, beyond the scope of this article.

6 Concluding Remarks and Future Directions

Employment of a standard Pub/Sub middleware facility has been proven in the context of emergency response communication networks, as an approach enabling *System of Systems* interoperability and integration, supporting services and classes of service usually required by civil protection operators and overtaking typical shortfalls in centralized service providing architectures, like bottleneck limitations and single point of failure weaknesses.

A novel extension has been provided, through an experiment implementation, to make available interactive delay-sensitive communication services, such as voice calls: it turned out as a change of paradigm, from a caller request push in a conventional Client/Server environment to a loosely coupled interaction between caller and be-called users, with evidence of a strong reduction in the system overhead and an increase in manageability.

In future work, we intend to be faced with cross-layer optimized adaptation among OMG DDS interoperability protocol [20], operating at above transport layer,

and underlying layers, differentiating on the basis of the diverse technologies composing the *System of Systems* (different wireless standards, wired networks, optical networks, etc.).

Acknowledgments This work was supported in part by MIUR-FIRB INtegrated SYstems for EMErgency (INSYEME) under Grant RBIP063BPH. Experiment implementation has been realized upon PrismTech OpenSpliceDDS© and designed by means of PrismTech OpenSpliceDDS PowerTools©, both employable through Academic Licence Agreement.

References

1. Catarci T, de Leoni M, Marrella A, Mecella M, Salvatore B, Vetere G, Dustdar S, Juszczyk L, Manzoor A, Truong HL (2008) Pervasive software environments for supporting disaster responses. IEEE Internet Comput 12(1):26–37
2. Kormentzas G, Katsikas S, Anerousis N, Venieris I (2007) Special issue: emerging middleware for next generation networks. Computer Commun 30(3):497–498
3. Balachandran K, Budka KC, Chu TP, Doumi TL, Kang JH (2006) Mobile responder communication networks for public safety. IEEE Commun Mag 44(1):56–64
4. Fantacci R, Vanneschi M, Bertolli C, Mencagli G, Tarchi D (2009) Next generation grids and wireless communication networks: towards a novel integrated approach. Wirel Commun Mob Comput 9(4):445–467
5. Juszczyk L, Dustdar S (2008) A middleware for service-oriented communication in mobile disaster response environments. Proceedings of the 6th international workshop on middleware for pervasive and ad-hoc computing, pp 37–42, Dec 2008
6. Chiti F, Fantacci R, Maccari L, Marabissi D, Tarchi D (2008) A broadband wireless communications system for emergency management. IEEE Wirel Commun Mag 15(3):8–14
7. Project MESA (2008) Statement of Requirements (SoR). Technical Specification 70.001 v3.3.1, Mar 2008
8. Project MESA (2007) System and network architecture. ETSI Technical Report 102 653 v3.1.1, Aug 2007
9. Project MESA (2005) System overview. Technical Report 70.012 v3.1.1, Dec 2005
10. Drugan OV, Plagemann T, Munthe-Kaas E (2006) Resource aware middleware services over MANETs. INFOCOM 2006. Proceedings of the 25th IEEE international conference on computer communications, pp 1–2, Apr 2006
11. Portmann M, Pirzada AA (2008) Wireless mesh networks for public safety and crisis management applications. IEEE Internet Comput 12(1):18–25
12. Soldatos J, Pandis I, Stamatis K, Polymenakos L, Crowley J (2007) Agent based middleware infrastructure for autonomous context-aware ubiquitous computing services. Computer Commun 30(3):497–498
13. Jiang X, Chen NY, Hong JI, Wang K, Takayama L, Landay JA (2004) Siren: context-aware computing for firefighting. Proceedings of the second international conference on pervasive computing. Springer, pp 87–105, Apr 2004
14. Alves S, Koldehofe B, Miranda H, Taiani F (2009) Design of a backup network for catastrophe scenarios. To appear in the First International Workshop on Advanced Topics in Mobile Computing for Emergency Management: Communication and Computing Platforms (MCEM 2009), Jun 2009
15. Eugster PTh, Felber PA, Guerraoui R, Kermarrec A (2003) The many faces of publish/subscribe. ACM Comput Surv (CSUR) 35(2):114–131
16. Vanneschi M, Archetti F, Ciciani B, Giordano S, Tisato F (2008) FIRB Project In.Sy.Eme (Integrated Systems for Emergency). Work Package 3: Mobile Grid. Starrylink, Nov 2008

17. Listanti M, Archetti F, Ronci F (2008) FIRB Project In.Sy.Eme (Integrated Systems for Emergency). Work Package 1: Reference Scenario. Starrylink, Nov 2008
18. OMG (2007) Data distribution service for real-time systems. Version 1.2, Jan 2007
19. IETF (2002) RFC 3261 – SIP: Session Initiation Protocol. Jun 2002
20. OMG (2009) The real-time publish-subscribe wire protocol DDS interoperability wire protocol specification. Version 2.1, Jan 2009

Part III
Localization and Applications

Localization Issues in a ZigBee Based Internet of Things Scenario

Ugo Biader Ceipidor, Massimiliano Dibitonto, Luca D'Ascenzo, and Carlo Maria Medaglia

1 Introduction

The expression *Internet of Things* [1,2] is wider than a single concept or technology. It is rather a new paradigm that involves a wide set of technologies, applications, and visions. Also, complete agreement on the definition is missing, as it changes with relation to the point of view. It can focus on the virtual identity of the smart objects and their capabilities to interact intelligently with other objects, humans, and environments or on the seamless integration between different kinds of objects and networks toward a service-oriented architecture of the future Internet. The evolution of computing and networking technologies is drawing a new scenario where the devices that compose the network no longer have homogeneous characteristics, functionalities, and communication means but, by interacting together, they are able to perform a wide array of tasks. This new generic type of device, also called "smart objects," can be identified in any device able to process information, interact with the surrounding environment, and with other devices. Smart objects' context awareness is a key enabler in this scenario: the ability of objects to acquire information on the surrounding environment process and manage collected data in an intelligent manner will lead the way to multiple brand new applications. The context may refer to a wide variety of physical parameters and more complex phenomena, such as temperature, humidity, presence, position, speed, or remote events.

This work focuses on localization as a smart object service. Localization may be absolute (identified on a global scale) or relative (identified in the frame of a given environment). Knowledge of the object's position, especially when combined with other information collected through sensors and shared through the connection with other smart objects, allows to develop systems capable of responding to

M. Dibitonto (✉)
Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy
e-mail: massimiliano.dibitonto@diee.unica.it

U.B. Ceipidor, M. Dibitonto, L. D'Ascenzo, and C.M. Medaglia
Centro per le Applicazioni della Televisione e delle Tecniche di Istruzione a Distanza (CATTID), University "Sapienza", Rome, Italy
e-mail: ugo.biader@uniroma1.it; carlomaria.medaglia@uniroma1.it; dascenzo@cattid.uniroma1.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_15,

changes in environmental conditions by applying rules or adaptive algorithms. In a service-based Internet of Things architecture, localization could be viewed as one of the automatically discoverable services provided by a peripheral network.

2 ZigBee WSN

Sensor networks represent a new stage in the development of infrastructure able to process external stimuli in order to describe, with increasing accuracy, the world around us. The advancement of technology has brought the first simple point-to-point structures to become real networks that connect different kind of devices, through paths which may vary depending on the workload and the status of the nodes of the network. The data collected by smart objects scattered into the environment can be transmitted to a central unit, with more hardware resources, that has the task of producing the set of commands that will be sent and executed from various actuators. This approach allows to integrate in the decision process the data coming from other possibly interrelated networks. Alternatively, these environmental data can be sent to subsets of devices, which through a process of aggregation, have expressed an interest in that particular type of information. In this way, it creates a true distributed intelligence that allows the individual device to make a choice dependent solely on the information received and not dictated by an external decision-making process. In this context, the ZigBee [3,4] is a protocol for communications wireless, based on IEEE 802.15.4 [5], infrastructure that provides a reliable and robust exchange of information between devices equipped with any kind of sensors. Thanks to these characteristics the ZigBee WSN are integrated in the Internet of Things scenario, becoming an enabling factor.

3 Localization via a ZigBee WSN

Network infrastructures created by a ZigBee WSN can be also used to determine the position of nodes [6–8]. Among the various techniques for estimating the position of a node, measuring the power of the signal received (RSSI) is one of the most suitable, thanks to its implementation simplicity and to the limited hardware resources required [9,10]. In particular in this work, studies and tests have been conducted using the kit CC2431DK of Texas Instruments that implements a trilateration algorithm in its Location Engine. The theoretical analysis of the system showed that the performance of the location algorithm [11, 12] is influenced by several factors such as the shape and dimensions of environment, objects, number of people present in the room, position of the device to locate (Blind Node), the location of reference nodes and their density in addition to the typical parameters of the routine. The formula that links the key variables for the calculation of the position is as follows

$$\mathrm{RSSI} = -(10\, n \log_{10} d + A),$$

where the parameter A is the absolute value of average power in dBm received at a reference distance of 1 m and the parameter n describes the decrease of the power varying with the distance d between transmitter node and receiver node. The parameter A can be measured with the desired accuracy before installing the nodes through a series of measures that considers the imperfect isotropic antennas. The parameter n cannot be calculated in advance and must be pre-set according to the experience of those who create the network.

3.1 Field Test Campaign

A field test campaign was performed with the objective of verifying the performance or the original Texas Instruments' localization algorithm. Field trials have been run in two different environments with different characteristics, used as "models" for application in real sites. The two environments have been chosen inside the CATTID building (Sapienza Campus). The first is the RFID Lab room, which has been considered as the worst-case scenario because of the presence of several wireless 802.11 networks, infrastructures for localization through the same router, some RFID readers and NFC devices. Interference on the same band used by the nodes is active at all times and with varying intensity. The second area considered is the "study room" available to students. That room can be a model for an office in which metal objects and computers create multiple paths and destructive interference. The Location Engine requires at least three references to estimate the position, but tests were initially performed by placing four Reference Nodes along the perimeter of the rooms, 2 m above the floor, trying to create a regular area and placing the Blind Node in a central position in the so created area.

As a preliminary operation, we calculated the value of the A parameter measuring the power received at a distance of 1 m through the software pre-installed on a node of the kit. The power measurement was done on all nodes, changing angular orientation in steps of 45°. After that process we chose the mean value of all measurements, corresponding to −40 dBm. Since it is related to a physic characteristic of the antenna, the A parameter could be considered constant in all the subsequent tests.

The first test was performed with the estimated value for A, and varying the parameter n over all the 32 possible values. The coordinates estimated were compared with the known position of the Blind Node. Results obtained show that, for low values of the n parameter, the error is so high that the node is located outside the area of interest. The error tends to decrease in the neighborhood of $n = 19$, and then rise again for higher values (Figs. 1 and 2).

The test was then repeated by increasing progressively from 4 to 6 the number of references, in order to estimate the importance of the number of inputs to the Location Engine (Figs. 3 and 4).

Results showed that increasing the number of Reference Nodes does not improve localization accuracy although, in few cases, it can rise the localization error.

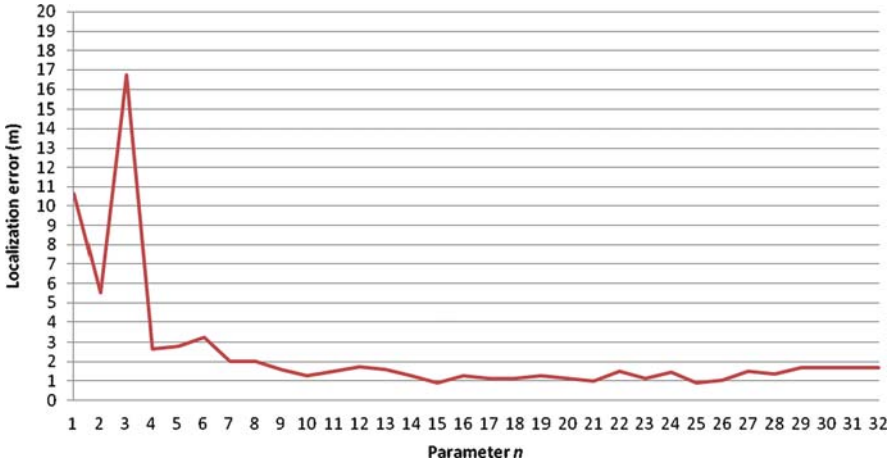


Fig. 1 Localization error varying *n* parameter with four Reference Nodes inside the RFID Lab

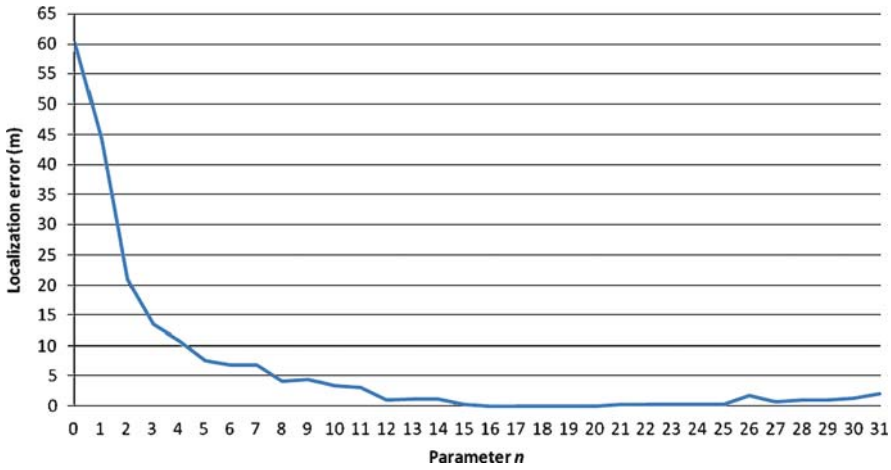


Fig. 2 Localization error varyng *n* parameter with four Reference Nodes inside the "study room"

3.2 Optimization of the Localization Algorithm

After the test campaign, that has shown the importance of the *n* parameter to achieve better localization results, the focus of research turned to find an automatic procedure for identifying the value that optimizes precision.

The existing routines have been integrated with an automatic procedure to estimate the value to be given to the exponent that describes the decay of power with distance. In this way, the attenuation model can rely on data that represent more accurately the characteristics of the signal propagation in that particular environment, since it is not inferred by who created the network, but from an

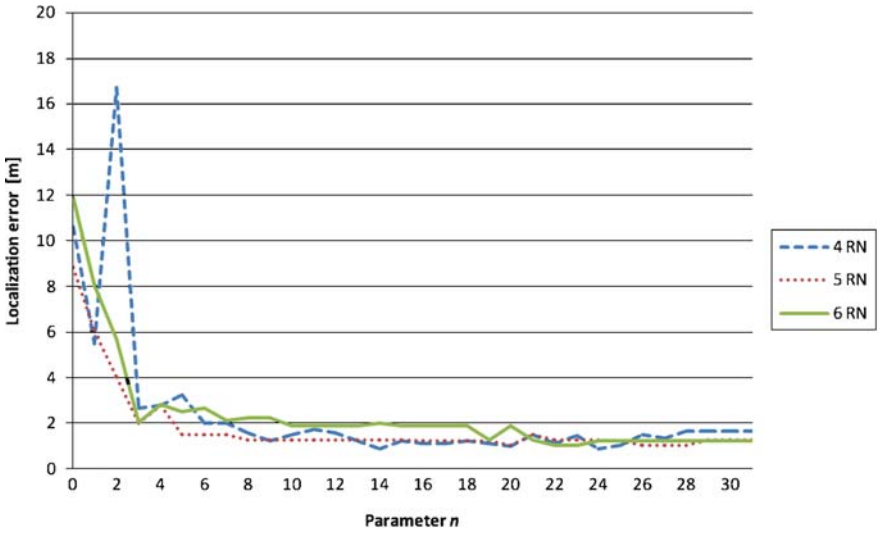


Fig. 3 Localization Error varying *n* parameter and number of Reference Nodes in the RFID Lab room

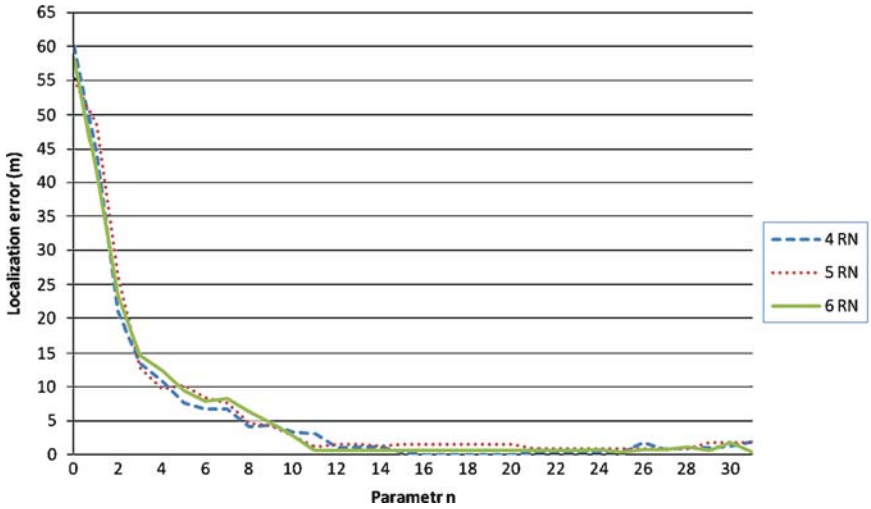


Fig. 4 Localization Error varying *n* Parameter and number of Reference Nodes in the "study room"

unbiased field testing. The test results also showed that, in the same environment, adjacent rooms had different values for the *n* parameter. This behavior led us to organize reference nodes into groups that match a room area and that share the same *n* parameter. The original algorithm requires an initial phase of configuration to set the coordinates of the Reference Nodes. In this phase, some additional steps have

been introduced to allow a network to provide the best configuration to a mobile Blind Node operating within it. The process designed and implemented includes the following steps:

1. Setting the coordinates for each Reference Node
2. Division of the references in groups
3. Estimation of the parameter n via a Blind Test

The first point is identical to the procedure followed to setup the software supplied with the kit CC2431DK. The second point aims at dividing the references into groups that identify different rooms. Through the functionality provided by the ZigBee protocol is possible to group a variable number of network addresses in a single 16-bit address: Using this feature to represent a room, you can ensure that the Reference Nodes can provide to the Blind Node the best n parameter value for that environment. In the third step, a Blind Node is placed inside the room, receives its own coordinates, and starts the auto-configuration procedure. The Blind Node repeatedly performs the location procedure, iterating on all possible values of the n parameter, and calculates the position error respect to the coordinates given above. The value of the n parameter that brings better accuracy is sent to the closest reference that transmits it to all nodes that are part of the group. The whole procedure takes about 10 s. Once the configuration phase of the network is ended, each room has its own identification number and any reference knows the value of the n parameter that better represents the propagation characteristics of the environment in a given time.

During the normal operation of the network, the Blind Node holds in a variable the ID of the room where it was located the last time. If it enters in a new environment, the identity of which does not coincide with the one stored in the memory, the Blind Node requires the n parameter to be used via an appropriate message. The reference that receives the request responds with the Blind Node Reference Node Configuration Response cluster that contains the value of the n parameter.

To avoid obsolescence of the n parameter due to changes in the environment (for instance a group of people enters the room), the Blind Node could be left in the initial known position, repeating the estimation procedure at a given interval of time, or when asked from an input coming from another device in the network which is able to detect changes that could affect location accuracy.

The described procedure was performed in the same environments described in the previous paragraph. The n parameter calculated was in 80% of cases corresponding to the one that showed the best location accuracy during the previous test campaign (Figs. 5 and 6).

4 Related Works

Localization is an important field of study and many works have been done on localization for ZigBee motes. Many of these works use RSSI-based localization algorithm for its simplicity and for the small amount of hardware resources required.

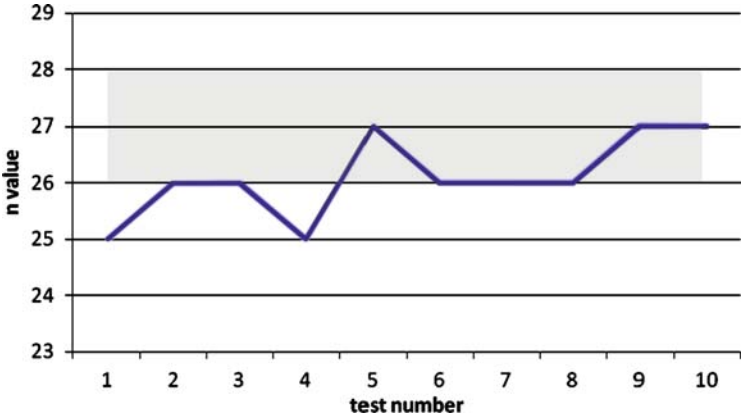


Fig. 5 Value of *n* parameter calculated from the automated procedure for the RFID Lab room. The gray area indicates the values of *n* that gave the best locating accuracy in the field test

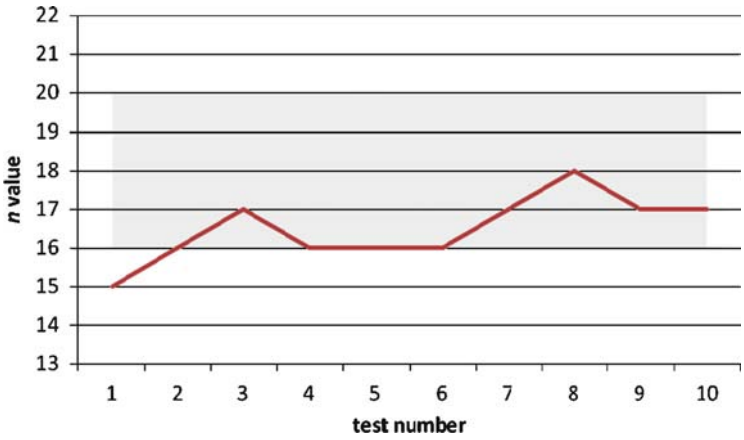


Fig. 6 Value of *n* parameter calculated from the automated procedure for the "study room." The gray area indicates the values of *n* that gave the best locating accuracy in the field test

In [13], the authors investigate a cooperative algorithm that uses the signal received both from the reference nodes and from the unknown nodes (the equivalent of blind nodes in this work). This algorithm infers the position of the blind nodes using a two-step RSSI-based algorithm:

- An initial region, where the node is expected to be, is calculated as the intersection of restricted distances from beacon (reference nodes in this work)
- In a second step, an iterative procedure is applied to refine the location measuring signal strength between the so located nodes

However, this approach needs that a number of blind nodes are in direct communication range in order to provide the refinement of the initial – coarse – region.

Basing the refinement method on the unknown nodes does not necessarily provide an independent method and, moreover, propagates the initial uncertainty on position. Last but not least, this scheme is likely to be time consuming and thus is not suited for moving objects as they vary their configuration between the two steps.

Another approach is the Adaptive Weighted Centroid Location [14]. It uses both RSSI and the Link Quality Indicator (LQI), an index provided from the IEEE 802.15.4 standard that represents the characterization of the strength and/or the quality of a received packet and that should be an integer ranging from 0 to 255. Reference nodes send their position to the blind nodes that uses LQI to evaluate their position and a weight to ensure a most precise localization. In this method, LQI values are reduced by a pre-calculated ratio which is determined through experimental tests in order to mitigate errors. Even if this algorithm offers better performances than previous Weighted Centroid [15] localization, it still does not provide a good accuracy and needs to be configured for ever application scenario.

It is widely recognized that Ultra Wide Band technology is technically the best foreseeable solution for localization [16] in the mid-term. Ultra Wide Band devices have, by definition [17], relative bandwidths larger than 20% or absolute bandwidths of more than 500 MHz and use pulses of very short duration broadly spread in the frequency domain. As described in [18], thanks to the high time resolution, UWB offers centimeter precision using time-based location estimation schemes.

UWB was conceived for short range, high data rate applications. Low rate WPANs have tighter power consumption requirements and thus are better suited for creating an autonomous and reliable network infrastructure for the peripheral part of the Internet of Things. A view of the current scenario, including power consumption and bitrate can be found in [19].

5 Conclusion

According to the obtained results, the localization algorithm can be improved if properly configured, shaping as precisely as possible the environment in which the signals propagate. The proposed procedure for automatically calculating the parameter that describes the power decay according to the distance in a given environment, improves location accuracy. The organization of the references nodes in groups representing different environments and the ability to assign to each of them a specific value makes the system more flexible and, therefore, it allows achieving greater accuracy than the original procedure. The achieved precision, even if lower than that of other localization techniques such as Ultra Wide Band, should be considered enough for providing localization as an added value service in a WPAN-based IoT scenario.

References

1. EC-EPoSS (2008) The Internet of Things in 2020. Joint Report of the European Commission and the European Technology Platform on Smart Systems Integration
2. ITU (2005) The Internet of Things. International Telecommunication Union (ITU), Internet Report Services, Geneva. November 2005
3. ZigBee Alliance (2009) http://www.zigbee.org. Accessed 10 July 2009
4. ZigBee Alliance (2009) ZigBee specifications. Accessed 10 July 2009
5. IEEE 802.15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low Rate Wireless Personal Area Networks (WPANs), IEEE, IEEE Standard 802.15.4–2006, 2006
6. Norris M (2006) Location monitoring with low-cost ZigBee devices. Embedded Control, Europe
7. Sugano M, Kawazoe T, Ohta Y, Murata M (2006) Indoor localization system using RSSI measurement of wireless sensor network based on ZigBee standard. In: IASTED international conference on wireless sensor networks, pp 1–6
8. Park W, Yoon M (2006) The implementation of indoor location system to control ZigBee Home Network. In: SICEICASE international joint conference, pp 2158–2161, 18–21
9. Cho H, Kang M, Park J, Kim H (2007) Performance analysis of location estimation algorithm in ZigBee Networks using received signal strength. 21st Int Conf Adv Inf Netw Appl Workshops 2:302–306
10. Hatami A, Pahlavan K (2005) A comparative performance evaluation of RSS-based positioning algorithms used in WLAN networks, Wireless Communications and Networking Conference, 2005 IEEE, Vol. 4, pp 2331–2337, 13–17, March 2005
11. Aamodt K (2008) CC2431 location engine – Application note AN042, Chipcon products for Texas Instruments
12. Siri Namtvedt, RSSI Interpretation and Timing, Texas Instruments Design Note DN505, SWRA114B
13. Chen W, Meng X (2006) A cooperative localization scheme for ZigBee-based wireless sensor networks. ICON '06. 14th IEEE Int Conf Networks 2:1–5
14. Behnke R, Timmermann D (2008) AWCL: Adaptive Weighted Centroid Localization as an efficient improvement of coarse grained localization. In Proceedings of the 5th Workshop on Positioning, Navigation and Communication, pp 243–250, 27 March 2008
15. Blumenthal J, Grossmann R, Golatowski F, Timmermann D (2007) Weighted centroid localization in ZigBee-based sensor networks. In: IEEE international symposium on intelligent signal processing, WISP, Madrid
16. Patwari N, Hero III A O, Ash J, Moses R L, Kyperountas S, Correal N (2005) Locating the nodes. IEEE Signal Process Mag 22(4):54–69
17. Siriwongpairat WP, Liu KJR (2007) Ultra-wideband communications systems, 1st edn. John Wiley and Sons, New Jersey
18. Gezici S, Zhi Tian, Giannakis GB, Kobayashi H, Molisch AF, Poor HV, Sahinoglu Z (2005) Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks, Signal Process Mag IEEE 22(4):70–84
19. Jin-Shyan L, Yu-Wei S, Chung-Chou S (2007) A comparative study of wireless protocols: Bluetooth, UWB, ZigBee, and Wi-Fi. In: Industrial electronics society, 2007. IECON 2007. 33rd Annual conference of the IEEE, pp 46–51

Low-Complexity Audio Signal Processing for Localization in Indoor Scenarios

Marco Martalò and Gianluigi Ferrari

1 Introduction

During the last years, wireless sensor networks (WSNs) have received significant attention from the research community, as one of the emerging technologies for the new millennium. A WSN is composed by many (e.g., hundreds) devices with limited processing and communication capabilities. Therefore, energy saving is one of the major issues and information processing has to be performed with low complexity. Possible applications for WSNs are surveillance, environmental monitoring, flow control, etc, and it may be possible to work in indoor scenarios [1]. The application of interest in this paper is the localization of a person or an object in indoor scenarios [2], but our approach is also suitable for outdoor scenarios. We will refer to the person (or object) to be localized as the *entity*. In these scenarios, it is realistic to assume that the nodes are *not* equipped with global positioning system (GPS) devices and, therefore, other techniques are needed to perform efficient localization.

Most of the works about localization in sensor networks are based on the assumption that the entities are equipped with devices which radio-communicate with some reference nodes (denoted as *anchors*). The positions of the anchors are supposed to be known and the position of the entity of interest is inferred by "combining" the information available at each anchor (e.g., by triangulation). In the literature, several techniques, based on different methods, have been proposed to obtain a sufficiently low estimation error [3, 4]. The computational complexity of these algorithms is a crucial issue for WSN-based applications. In [5], the authors propose a sub-optimal hierarchical algorithm, which solves the localization problem without resorting to the optimum maximum likelihood (ML) technique, whose computational complexity becomes too high to be of any practical interest. In [6], an adaptive approach to localization problems, obtained by solving a sequence of very small optimization subproblems, is considered.

M. Martalò (✉) and G. Ferrari
WASN Lab, Department of Information Engineering, University of Parma, Parma, Italy
e-mail: martalo@tlc.unipr.it; gianluigi.ferrari@unipr.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_16,

The problem of locating a source of speech has been widely studied in the field of sound source localization (SSL) using multiple input multiple output (MIMO) signal processing [7]. In particular, several analytical frameworks have been proposed for the estimation of the time difference of arrival (TDOA). Most of the methods are based on measuring the crosscorrelation (in time or frequency domains) between the output at different receivers [8, 9]. An interesting approach is that proposed in [10], where the authors derive a unified ML framework for sound source localization and beamforming for distributed meeting applications, taking into account both reverberation and environmental noise.

In this chapter, we will assume that the anchors are equipped with microphones (which have, typically, a low cost) and use the information collected by them to localize, through collaborative SSL-based signal processing, the entity of interest. This is reasonable in scenarios where the entity to be localized may not be equipped with these devices, e.g., when the entity is an enemy to be located in a battlefield. In such cases, it is thus necessary to use other methods to localize the entity, e.g., by employing other types of sensors, such as accelerometers, microphones, etc. Although SSL techniques are well established (especially for distributed meeting applications), they are mainly based on the computation of crosscorrelation and the use of ML estimators, which are computationally onerous. *In this chapter*, we derive SSL techniques which employ very limited computational complexity, trying to obtain the minimum penalty in terms of position estimation error. In particular, we present results based on a novel localization algorithm which, by considering the powers of the audio signals received at the microphones, determines the position of the entity. We first deal with one-dimensional scenarios, i.e., scenarios where the audio source moves along a straight line, deriving both *centralized* and *decentralized* localization algorithms, based on the solution of simple systems of equations. Then, we extend our approach to consider more realistic two-dimensional scenarios, where the anchor nodes are placed at the corners of a square grid.

2 One-Dimensional Scenarios

2.1 Statement of the Problem

Suppose that the entity to be localized is moving on a straight line (x-axis) and there are N anchors (microphones), denoted as $[m_0, \ldots, m_{N-1}]$, equally spaced at positions $[x_0, \ldots, x_N - 1]$, where

$$x_j = x_0 + D \cdot j \qquad j = 1, \ldots, N-1$$

and D is the constant distance between two consecutive anchors. Without loss of generality, suppose that $x_0 = 0$. This approach can be easily extended to scenarios with nonequally spaced sensors.

The audio power received at the ith node ($i = 0, 1, \ldots, N-1$) can be expressed as

$$P_i = k \frac{P_s}{(d_i)^\beta} \tag{1}$$

where k is a suitable parameter which depends on the audio propagation characteristics (dimension: [cm$^\beta$]), β is the pathloss decay exponent (adimensional), d_i is the distance of the audio source from m_i (dimension: [cm]), and P_s is the maximum emitted power by the sound source (dimension: [W]). Obviously, expression (1) can be applied provided that $d_i > d_{\text{crit}}$, where d_{crit} is a critical distance beyond which equation (1) holds. Although a more detailed statistical description of the model may be needed, results on acoustic emission of human head confirm this type of model [11]. Note that β depends on the type of audio source. If, for instance, the sound is emitted by an object (e.g., a motor), it is reasonable to assume that the sound may approximately have the same propagation characteristics in all directions (positive and negative directions on the x-axis). If, instead, the sound is emitted by a person, the sound will decay slowly in front of the face, whereas it will decay faster on the opposite direction. For the ease of simplicity, we will assume that $\beta = \beta_{\text{forward}} = 2$ in the "forward" direction, whereas $\beta = \beta_{\text{backward}} = 4$ in the "backward" direction. We remark that in an homogenous scenario the propagation exponent β should be the same in all directions. For the purpose of analysis and without leading the generality of our framework, we suppose that β depends on the propagation direction. A different propagation modeling would simply require to change a few equations. However, the value of β changes from person to person and a more accurate description may be needed [11]. As will be shown later, an accurate characterization of k is not crucial in our analytical framework, since the same value of k is considered for all sensors.

The problem consists in locating the audio source on the basis of the N audio power received at the anchors. Our goal is to derive an efficient cooperative processing algorithm, with low computational complexity, for the localization of the entity. In Sect. 2.2, we will derive a centralized algorithm in the considered one-dimensional scenarios, whereas in Sect. 2.3, distributed localization algorithms will be proposed.

2.2 Centralized Localization Algorithm

In a scenario with *omnidirectional* audio source emission (e.g., a motor), the parameter β is the same in all directions. Therefore, by identifying the two nodes which receive the highest audio powers, one can determine the position of the entity. For instance, consider the scenario depicted in Fig. 1, where the source is between m_{i-1} and m_i. The powers P_{i-1} and P_i received at m_{i-1} and m_i, respectively, will be the highest ones. The distance x^*_{crit} of the audio source from m_{i-1}, i.e., $x^*_{\text{crit}} = x^* - (i-1)D$, can be directly obtained, once the powers at the $(i-1)$th and ith sensors are collected.

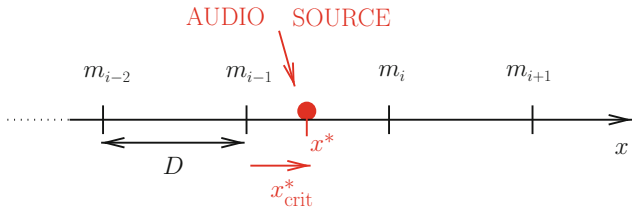


Fig. 1 Reference one-dimensional scenario

In the presence of *directive* audio sources (e.g., a human speaker), the above approach cannot be applied, since the value of β changes according to the direction in which the sound is emitted. For example, if the entity is between m_{i-1} and m_i and speaks "forward" (i.e., in the positive direction of the x-axis), even if $P_i > P_{i-1}$, it might happen that the entity is closer to m_{i-1} than to m_i. Vice-versa, if the entity is speaking "backward" and $P_i > P_{i-1}$, then for sure the entity is closer to m_i. It might even happen (depending on the value of D and the values of β_{forward} and β_{backward}) that if the entity is on the left of m_{i-1} (i.e., between m_{i-2} and m_{i-1}) and speaks forward, then the powers perceived at m_{i-1}, m_i, and m_{i+1} are the highest ones.

As can be understood from the illustrative examples in the previous paragraph, a generalized approach needs to be considered when the "direction" of the sound (forward or backward) has also to be determined. In this case, one needs to consider at least three nodes. Suppose that the highest received powers are those perceived at the anchors m_{i-1}, m_i, and m_{i+1} (the order is not relevant). Then, only one of the following exclusive situations can happen:

1. The entity speaks forward and is between m_{i-2} and m_{i-1}
2. The entity speaks forward and is between m_{i-1} and m_i
3. The entity speaks backward and is between m_{i-1} and m_i
4. The entity speaks forward and is between m_i and m_{i+1}
5. The entity speaks backward and is between m_i and m_{i+1}
6. The entity speaks backward and is between m_{i+1} and m_{i+2}

Each of the above conditions is associated with a specific system. For example, in the first case, assuming, as mentioned in Sect. 1, that $\beta_{\text{forward}} = 2$, the following system admits a unique solution $x^*_{\text{crit}} < 0$:

$$\begin{cases} P_{i-1} = k \dfrac{P_{\text{s}}}{(x_{\text{crit}})^2} \\ P_i = k \dfrac{P_{\text{s}}}{(D - x_{\text{crit}})^2} \\ P_{i+1} = k \dfrac{P_{\text{s}}}{(2D - x_{\text{crit}})^2}. \end{cases}$$

In the second case, assuming $\beta_{\text{backward}} = 4$, the following system admits a unique solution $x^*_{\text{crit}} > 0$:

$$\begin{cases} P_{i-1} = k\dfrac{P_s}{(x_{crit})^4} \\ P_i = k\dfrac{P_s}{(D - x_{crit})^2} \\ P_{i+1} = k\dfrac{P_s}{(2D - x_{crit})^2}. \end{cases}$$

In general, there will be only one system (out of the six possible ones) which will admit an acceptable solution ($-D < x^*_{crit} < 2D$). Therefore, the audio source location and direction can be univocally determined by finding such a system.

2.3 Distributed Localization Algorithms

Although the analytical framework described above is very simple to be implemented, it requires a global network knowledge, since the three highest received powers (among all the N powers received at the anchors) are used to determine the audio source position and direction of emission. In practical networks, a centralized solution may not be feasible, since extra nodes with higher computational resources may be required. Therefore, it is of interest to derive distributed algorithms, where data are gathered and disseminated with the smallest possible number of interanchor communications. In the literature, several distributed algorithms have been proposed, based especially on the use of machine learning techniques [12, 13]. In the following, we derive two possible distributed strategies. The common feature of these strategies is that only the three nodes with the highest received powers are involved in the possible systems described at the end of Sect. 2.2. In other words, the location of the entity is determined from the data perceived by three anchors and, then, disseminated to all other nodes – this might be of interest for tracking operations.

The first proposed protocol can be described as follows:

- A packet with the information about the three nodes with the highest powers is created at node m_0 and propagated along the x-axis. At the first transmission act, node m_0 only stores its received power.
- If an anchor receiving the packet has a measured power higher than any of the three stored in the packet, it modifies the packet by discarding the lowest power and introducing its own.
- If a node has a received power lower than the three powers collected in the packet, data gathering stops, since it is not possible (according to the propagation model (1)) to find forward an anchor with higher received power.
- Once the nodes with the highest received powers are identified, the one-dimensional localization algorithm described in Sect. 2.2 is carried out at these nodes.
- Finally, the estimated position and direction are disseminated through the network to all other nodes.

While it is possible to show that $O(N/2)$ interanchor communications are needed, on average, during the data gathering phase, one should note that in the worst case scenario the number of transmissions is $O(N)$. Therefore, one may design a more efficient gathering algorithm which starts from the middle of the anchor sequence, instead of one edge, of the network. This protocol can be described as follows.

- A packet with the information about the three nodes with the highest powers is created at node $m_{\lfloor N/2 \rfloor}$ and it is propagated along the two directions of the x-axis. As in the previous case, anchor $m_{\lfloor N/2 \rfloor}$ stores only its power.
- For each direction, if a node receiving the packet (with already three stored values) has a higher measured power, it modifies the packet by discarding the lowest power and introducing its own; the gathering phase stops when the three highest are collected (in each direction).
- The three nodes with the overall highest received powers are determined according to the information in the packets collected above.
- Once the nodes with the highest received powers are identified, the one-dimensional localization algorithm described in Sect. 2.2 is carried out at these nodes.
- Finally, the estimated position and direction are disseminated through the network to all other nodes.

In Fig. 2, the average number of interanchor communications $\overline{n}_{\text{tx}}$ is shown, as a function of the number of nodes, for the two distributed protocols described above. The interanchor distance is set to $D = 50\,\text{cm}$. To this regard an ad hoc simulator, written in Matlab [14], has been created. The average is computed by simulating different (independent) positions and directions, computing the number of communication steps for each run and, finally, averaging. As expected, $\overline{n}_{\text{tx}}$ is an increasing function of the number of anchors, since, on average, one may need more steps before reaching the nodes with the highest received powers. However, one can observe that the second protocol is more efficient, since it requires a smaller number of communication steps before completing the localization process. This is due to

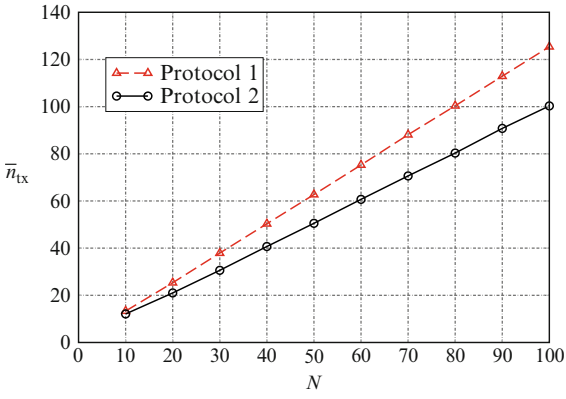


Fig. 2 Average number of communications $\overline{n}_{\text{tx}}$, as a function of the number of nodes, for the two distributed protocols described above. The interanchor distance is set to $D = 50\,\text{cm}$

the fact that the first protocol is "unbalanced," i.e., it may happen that the audio source is close to m_{N-1}. The second protocol, instead, is more "balanced," since it never happens that the audio source is N steps away from the anchor ($m_{\lfloor N/2 \rfloor}$) which initializes the data gathering phase.

3 Two-Dimensional Scenarios

3.1 Statement of the Problem

Suppose that the entity to be localized is moving on a square area of side D (e.g., a room), and there are four microphones, denoted as $[m_1, m_2, m_3, m_4]$, equally spaced at the corners of the square area. Without loss of generality, suppose that the origin of the axes is at the center of the square area. Therefore, the anchors are placed at $(\pm D/2, \pm D/2)$.

The audio power received at the ith anchor ($i = 1, 2, 3, 4$) can be expressed as in (1), where d_i is now the euclidean distance of the audio source, located at (x^*, y^*), from m_i, i.e.,

$$d_i = \sqrt{|x^* - x_i|^2 + |y^* - y_i|^2}.$$

As in the one-dimensional case, in this case as well β depends on the type of audio source and sound emission. For ease of simplicity, we assume that the person can speak only along one of the axes. In this case, we assume that the decay factor in a frontal region of span angle θ is equal to $\beta = \beta_{\text{forward}} = 2$, otherwise it is $\beta = \beta_{\text{backward}} = 4$. In a scenario with *omnidirectional* audio source emission (e.g., a motor), the parameter β is the same in all directions, i.e., $\theta = 2\pi$. Therefore, by identifying the audio power distribution among the nodes, one can determine the position of the entity by solving (1) for each anchor node. A pictorial description of the scenario is given in Fig. 3.

The problem consists in locating the audio source on the basis of the four audio powers received at the anchors and the knowledge of the sound emission characteristics, i.e., θ.

3.2 Centralized Localization Algorithm

Without loss of generality, we focus on the case when the entity speaks toward the side delimited by m_1 and m_4, i.e., the "eastern side" of the area. However, similar considerations can be carried out for the other three sides.

In order to determine the position of the entity, the value of β in (1), at all four anchors, has to be known. One should note that, if the entity is close to the center of

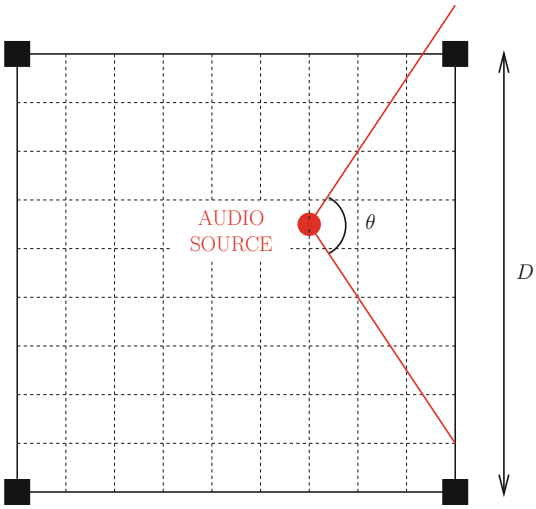


Fig. 3 Reference two-dimensional scenario

the side, none of the microphones will observe a decay factor equal to 2, since none of the microphones is spanned by the forward emission lobe of the entity. On the other hand, if the entity is close to the other side but still in the middle, m_1 and m_4 will observe a value of β equal to 2, whereas m_2 and m_3 will observe a value of β equal to 4. Moreover, only m_1 (m_4, respectively) will observe $\beta = 2$ if the entity is in the upper (bottom, respectively) part of the area. Therefore, one has to simply compute the equations of the straight lines that divide these four sectors. After a few geometrical considerations, denoting as β_i the decay factor at anchor i, it is possible to verify that

$$\begin{cases} \beta_i = 4 \quad \forall i & \text{if } x > x_{\text{crit}} \text{ and } |y| < m|x| + q \\ \beta_1 = \beta_4 = 2,\ \beta_2 = \beta_3 = 4 & \text{if } x \leq x_{\text{crit}},\ y < -(mx + q), \text{ and } y > mx + q \\ \beta_1 = 2,\ \beta_2 = \beta_3 = \beta_4 = 4 & \text{if } y > 0,\ x < (y - q)/m, \text{ and } x > -(y + q)/m \\ \beta_4 = 1,\ \beta_1 = \beta_2 = \beta_3 = 4 & \text{otherwise} \end{cases}$$

where $x_{\text{crit}} = (D/2)\tan((\pi - \theta)/2)$, $m = \tan(\theta/2)$, and $q = -x_{\text{crit}} \cdot m$.

If one applies similar considerations to the other three sides, it is possible to identify a set of 16 systems (four for each possible direction). As one can see, the computational complexity has increased rapidly. In fact, we have increased only one dimension of the problem, increasing the number of systems to be solved from 6 to 16 systems. However, the complexity remains acceptable, since solving 16 systems has still lower complexity than that of an ML-based algorithm.

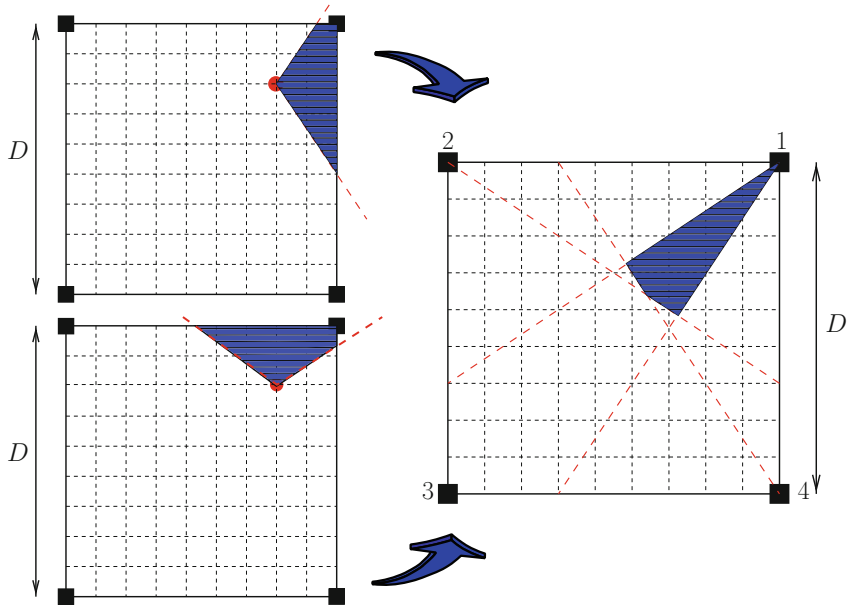


Fig. 4 Ambiguity zone for the anchor m_1 in a two-dimensional scenario

4 Results and Discussion

As previously anticipated, we have performed simulations using Matlab, in order to determine the effectiveness of the proposed approaches. In particular, we have compute the average errors in estimating the position and direction, denoted as ε_{p} and ε_{d}, respectively, by averaging over 1,000 independent trials the differences between the true position and direction and their estimates. During each trial, different positions and directions of the entity are randomly generated. The distance between consecutive anchors is set, in all cases, to $D = 50\,\mathrm{cm}$.

For one-dimensional scenarios, we found that $\varepsilon_{\mathrm{p}}^{(1-\mathrm{dim})}{=}\varepsilon_{\mathrm{d}}^{(1-\mathrm{dim})} = 0$, thus confirming the uniqueness of the solution of the 6 systems. In two-dimensional scenarios, we still found that $\varepsilon_{\mathrm{p}}^{(2-\mathrm{dim})} = 0$, but $\varepsilon_{\mathrm{d}}^{(2-\mathrm{dim})} \simeq 0.1$. In other words, the position is still correctly determined in all cases, but in 10% of the cases, the estimated direction is erroneous. A more detailed analysis has shown that this is due to the fact that there exists *two* possible systems with the same values of β_i and d_i ($i = 1, 2, 3, 4$). In this case, the position can be correctly estimated, but the direction is ambiguous. An example of the ambiguity zone for the anchor m_1 is shown in Fig. 4. In this region, there could be uncertainty between the emission directions toward the eastern or northern side. A proper strategy to solve this ambiguity still remains an open problem.

Acknowledgments We would like to thank Sandro Mattiacci, Claudio Malavenda, Luca Di Donato, and Paolo Proietti (Elsag Datamat S.p.A, Rome, Italy) for useful discussions on localization issues and audio signal processing. This work has been supported by a SPINNER 2013 fellowship.

5 Concluding Remarks

In this chapter, we have proposed a novel approach to perform low-complexity localization based on audio signal processing. A set of "anchors," which perceives the sound intensity (through audio sensors) emitted by an "entity" and collaborate together, estimate (a) the position of the entity and (b) the direction of sound emission. We have derived a framework for both one and two-dimensional scenarios, also showing possible distributed approaches in the one-dimensional case. Since ideal sound propagation conditions (i.e., no noise) have been assumed, the future work will be devoted to the derivation of proper techniques to counter-act the presence of acquisition and communication noises.

References

1. Akyildiz I, Su W, Sankarasubramaniam Y, Cayirci E (2002) A survey on sensor networks. IEEE Commun Mag 40(8):102–114
2. Bachrach J, Taylor C (2005) Localization in sensor networks. In: Stojmenović I (ed) Handbook of sensor networks: algorithms and architectures. Wiley, New York
3. Dricot J-M, Bontempi G, Doncker PD (2010) Static and dynamic localization techniques for wireless sensor networks. In: Ferrari G (ed) Sensor networks: where theory meets practice, Springer, pp 249–281
4. Savarese C, Rabaey J-M, Beutel J (2001) Locationing in distributed ad-hoc wireless networks, In: Proceedings of the IEEE international conference acoustics, speech, and signal processing (ICASSP), vol 4. Salt Lake City, UT, May 2001, pp 2037–2040
5. Dardari D, Conti A (2004) A sub-optimal hierarchical maximum likelihood algorithm for collaborative localization in ad-hoc networks. In: Proceedings of the IEEE sensor and ad hoc communications and networks (SECON), Santa Clara, CA, October 2004, pp 425–429
6. Carter M-W, Jin H-H, Saunders M-A, Ye J (2006) SpaseLoc: an adaptive subproblem algorithm for scalable wireless sensor network localization. SIAM J Optim 17(4):1102–1108
7. Huang Y, Benesty J, Chen J (2006) Acoustic MIMO signal processing. Springer, Heidelberg
8. Brandstein M-S, Adcock J-E, Silverman H-F (1997) A closed-form location estimator for use with room environment microphone arrays. IEEE Trans Acoust Speech Signal Process 5(1):45–50
9. Brandstein M-S, Adcock J-E, Silverman H-F (1995) A practical time-delay estimator for localizing speech sources with a microphone array. Comput Speech Lang 9(2):153–169
10. Zhang C, Florencio D, Ba D-E, Zhang Z (2008) Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. IEEE Trans Multimed 10(30):238–248
11. Dunn H-K, Farnsworth D-W (1939) Exploration of pressure field around the human head during speech. J Acoust Soc Am 10(1):184–199
12. Roos T, Myllymaki P, Tirri H, Misikangas P, Sievanen J (2002) A probabilistic approach to wlan user location estimation. Int J Wirel Inf Networks 9(3):155–164
13. Cristianini N, Taylor J-S (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
14. Matlab Website, http://www.mathworks.com

Integrated GPS-Denied Localization, Tracking, and Personal Identification

Stefano Tennina, Luigi Pomante, Fabio Graziosi, Marco Di Renzo, Roberto Alesii, and Fortunato Santucci

1 Introduction

Future wireless communication systems will have to enable and provide a multiplicity of services for a tremendously increasing number of mobile users. In this context, it is very important that wireless technologies might provide these services to the final users in a secure and robust way, guaranteeing their privacy regardless of the specific and requested service. In such a context, biometric technologies are expected to play a fundamental role in delivering wireless services in an exponentially increasing secure way. Moreover, future services are expected to be more and more specialized to the user's personal needs, as well as to the user's current location in space at a given time [1–5]. In such a context, the availability of accurate estimates of users' location will be an essential requirement for future wireless devices to guarantee context-awareness.

Moving from the above considerations, this research work offers an integrated solution for localization, tracking, and personal identification for radio-navigation systems deployed in realistic GPS-denied environments, such as indoor buildings. The main component of the proposed integrated solution is an embedded wireless biometric badge (Fig. 1), i.e., a "system-on-badge" system, which performs four main tasks: (a) localizing and tracking people by using a fully distributed positioning technique; (b) scanning, storing, and verifying people's fingerprints; (c) checking if the user is the badge's owner based on fingerprint matching; and (d) sending related outcomes wirelessly to the rest of the system (e.g., to a gateway), without the need to transmit biometric data over the wireless medium, thua achieving security while at the same time preserving users' privacy.

S. Tennina (✉), L. Pomante, F. Graziosi, F. Santucci, and R. Alesii
Department of Electrical and Information Engineering and Center of Excellence in Research DEWS University of L'Aquila, 67040 Poggio di Roio, L'Aquila (AQ), Italy
e-mail: stefano.tennina@univaq.it; luigi.pomante@univaq.it; fabio.graziosi@univaq.it; roberto.alesii@univaq.it;

M.D. Renzo
French National Center for Scientific Research (CNRS) Laboratory of Signals and Systems (LSS) Ecole Supérieure d'Electricité (SUPELEC) 3 rue Joliot-Curie, 91192 Gif-sur-Yvette (Paris), France
e-mail: marco.direnzo@lss.supelec.fr

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_17,

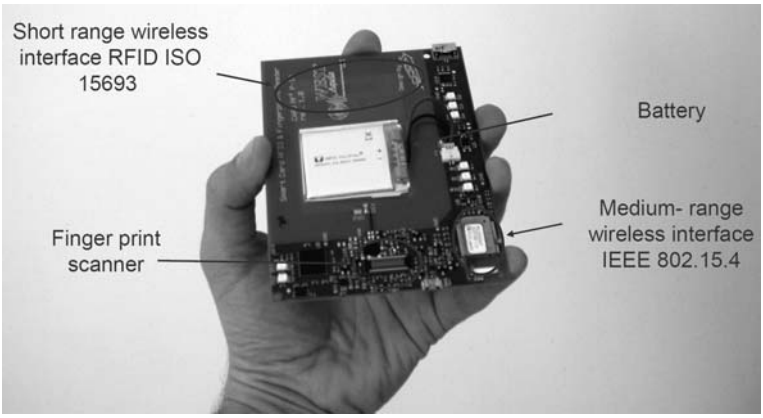


Fig. 1 Biometric Badge (developed by WEST Aquila s.r.l., a SME spin-off of the University of L'Aquila)

The remainder of this paper is as follows: Sect. 2 gives some insights on the biometric badge device; Section 3 presents some details about the novel positioning algorithm implemented as a software routine on the biometric badge's microprocessor, while in Sect. 4, a performance evaluation of the positioning algorithm is given. Finally in Sect. 5, the intended demo, which aims at showing the badge functionalities, is described.

2 Biometric Badge

The embedded biometric badge (Fig. 1) is a "system-on-badge," which performs several tasks: it scans and verifies the fingerprint of people; check if the user is the badge's owner on the basis of fingerprint matching; sends related outcomes wirelessly to the rest of the system, without the need to transmit biometric data of the user over the wireless medium (so, in a secure way from the point of view of transmitting critical data of the user).

Such features allow the badge to support a full range of applications (Fig. 2), mainly due to the embedded fingerprint reader that enables both the public and domestic use of such a technology in a secure and safe way.

The badge is currently equipped with:

- A Texas Instruments' SoC CC2430, which embeds an INTEL 8051 micro controller and a radio transceiver CC2420 based on the IEEE 802.15.4 standard. It is used for wireless medium-range communications, as well as localization operations
- A fingerprint sensor reader with its embedded companion chip provided by UPEK. This chip is the key element for handling biometric data: it makes it

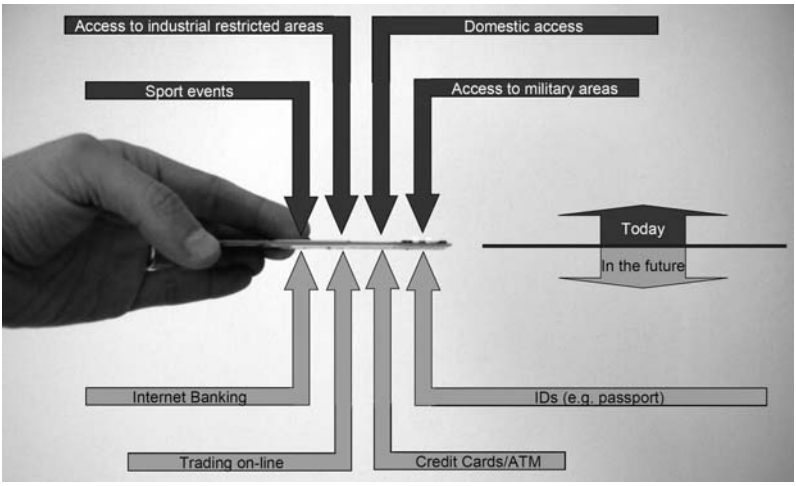


Fig. 2 Possible applications of our biometric badge

possible to authenticate people on the basis of fingerprint information, as well as to store data in a memory protected from external physical attacks. Moreover, only this chip and the gateway know how to decode the messages they send to each other

- An RFID tag based on the ISO15693 standard and its companion chip provided by Montalbano Technology, which allow the microcontroller to access data stored in the tag
- A rechargeable battery, its driver to monitor the charge status, and a user interface with eight LEDs and a push-button

3 Positioning Algorithm

As far as localization and tracking are considered, the badge exploits a novel distributed localization algorithm, which is called *enhanced steepest descent* (ESD) [6, 7], and represents an improved version of the well-known classical steepest descent (SD) algorithm. In the following section, a brief overview of the ESD algorithm and a comparison with the SD algorithm are presented. The interested reader can refer to [7] for a more detailed description, as well as a performance evaluation and comparison.

3.1 Notation

First, let us define the following notation, which will be used to describe the algorithm (a) bold symbols are used to denote vectors and matrices; (b) $(\cdot)^T$ denotes

transpose operation; (c) $\nabla(\cdot)$ is the gradient; (d) $\|\cdot\|$ is the Euclidean distance; (e) $\angle(\cdot,\cdot)$ is the phase angle between two vectors; (f) $\hat{\mathbf{u}}_j = \left[\hat{u}_{j,x}, \hat{u}_{j,y}, \hat{u}_{j,z}\right]^{\mathrm{T}}$ denotes the estimated position of the blind/unknown node U_j; (g) $\mathbf{u}_j = \left[u_{j,x}, u_{j,y}, u_{j,z}\right]^{\mathrm{T}}$ is the trial solution of the optimization algorithm for the unknown node U_j; (h) $\bar{\mathbf{u}}_i = [x_i, y_i, z_i]^{\mathrm{T}}$ with $i = 1, .., N$ are the positions of the anchor/reference nodes A_i; (i) $\hat{d}_{j,i}$ denotes the estimated (via ranging measurements) distance between reference node A_i and the unknown node U_j.

Both SD and ESD algorithms belong to the family of multilateration methods. In particular, in such a method, the position of an unknown node U_j is obtained by minimizing the error cost function $F(\cdot)$ defined in (1) as follows:

$$F\left(\mathbf{u}_j\right) = \sum_{i=1}^{N} \left(\hat{d}_{j,i} - \left\|\mathbf{u}_j - \bar{\mathbf{u}}_i\right\|\right)^2. \tag{1}$$

In general, the minimization of an error cost function can be done using a variety of numerical optimization techniques, each one having its own advantages and disadvantages in terms of accuracy, robustness, speed, complexity, and storage requirements [8]. Since optimization methods are iterative by nature, we will denote by the index k the kth iteration of the algorithm, and with $F\left(\mathbf{u}_j(k)\right)$ and $\mathbf{u}_j(k)$ the error cost function and the estimated position at the kth iteration, respectively.

3.2 Classical Steepest Descent

The SD is an iterative line search method which allows to find the (local) minimum of the cost function in (1) at step $k+1$ as follows [8, pp. 22, Sect. 2.2]:

$$\mathbf{u}_1(k+1) = \mathbf{u}_1(k) + \alpha_k \cdot \mathbf{p}(k), \tag{2}$$

where α_k is a step length factor, which can be chosen as described in [8, pp. 36, Chap. 3] and $\mathbf{p}(k) = -\nabla F\left(\mathbf{u}_1(k)\right)$ is the search direction of the algorithm.

In particular, when the optimization problem is linear, some expressions to compute the optimal step length to improve the convergence speed of the algorithm exist in the literature. On the other hand, when the optimization problem is nonlinear, as considered in this contribution, a fixed and small step value is generally preferred in order to reduce the oscillatory effect when the algorithm approaches the solution. In such a case, we have $\alpha_k = 0.5\mu$, where μ is the learning speed [8].

3.3 Enhanced Steepest Descent

The SD method provides, in general, a good accuracy in estimating the final solution. However, it may require a large number of iterations, which may result in a too

slow convergence speed, especially for mobile ad hoc wireless networks. In order to improve such convergence speed, we propose in this contribution an enhanced version of it, which we call enhanced steepest descent (ESD).

The basic idea behind the ESD algorithm is to continuously adjust the step length value α_k as a function of the current and previous search directions $\mathbf{p}(k)$ and $\mathbf{p}(k-1)$, respectively. In particular, α_k is adjusted as follows:

$$\begin{aligned} \alpha_k &= \alpha_{k-1} + \gamma \text{ if} \quad && \theta_k < \theta_{\min} \\ \alpha_k &= \alpha_{k-1}/\delta \text{ if} \quad && \theta_k > \theta_{\max} \\ \alpha_k &= \alpha_{k-1} \quad && \text{otherwise,} \end{aligned} \tag{3}$$

where $\theta_k = \angle(\mathbf{p}(k), \mathbf{p}(k-1))$, $0 < \gamma < 1$ is a linear increment factor, $\delta > 1$ is a multiplicative decrement factor, and $\theta_{\min}$ and $\theta_{\max}$ are two angular threshold values that control the step length update.

By using the four degrees of freedom γ, δ, $\theta_{\min}$, and $\theta_{\max}$, we can simultaneously control the convergence rate of the algorithm and the oscillatory phenomenon when approaching the final solution in a simple way, and without appreciably increasing the complexity of the algorithm when compared to the classical SD method. Basically, the main advantage of the ESD algorithm is the adaptive optimization of the step length factor α_k at run time, which allows to dynamically either accelerate or decelerate the convergence speed of the algorithm as a function of the actual value of the function to be optimized. In the next section, we will show the performance of this algorithm on a real implementation.

4 Proof-of-Concept via Experimental Testbed

In order to assess both implementation issues and performance of the proposed ESD algorithm via experiments, we have implemented a testbed platform by using Texas Instruments/Chipcon CC2431 (see [9]) sensor nodes as anchor/ reference nodes and our badge as blind/unknown node.

4.1 Ranging Model

This sensor node platform uses an RSS-based ranging method, and requires a (known) RSS-to-distance calibration curve to estimate the distance between pairs of nodes from RSS measurement [10], as follows:

$$d = 10^{[\frac{RSS-A}{10 \cdot n}]}, \tag{4}$$

where d denotes the transmitter-to-receiver distance, n is the propagation path-loss exponent, A represents the RSS value measured by a receiver that is located 1 m away from the transmitter (i.e., reference distance), and RSS is the actual measured value.

In order to estimate this calibration curve, we use the standard procedure described by [10], which consists in deploying a grid of nodes in the area of interest and extracting the desired parameters by postprocessing the gathered data. Accordingly, a 6 m × 10 m grid of sensor nodes has been deployed in a conference room of our Faculty in L'Aquila [11]. The devices located in the ground floor are receiver nodes, while transmitter nodes are deployed at the edge of the measurement area, thus yielding a minimum and maximum transmitter-to-receiver distance of 0.5 m and 11.7 m, respectively. Moreover, the transmitters can be located at different heights with respect to the ground floor (from 5 cm to 1.2 m). To estimate the calibration curve, the transmitters broadcast packets in a time-scheduled fashion such that collisions are avoided, and the receivers collect RSS values for each received packet, and then send a report to the host PC.

The RSS-to-distance reference curve in (4) is obtained via a least-square best linear fitting from several collected RSS values (every receiver node measures RSS values during a 5 min acquisition window, resulting in approximately 2,000 RSS values). The obtained result is shown in Fig. 3 along with real measurements. Note that, in Fig. 3 (a) the RSS values are represented as absolute values in arbitrary units, as provided by the receiver nodes, (b) the distance d in the horizontal axis is normalized to the reference distance of $d_0 = 1$ m, and (c) the fitting parameters are $A = 59.66$ and $n = 1.84$. Note that a path-loss exponent smaller than free space propagation is obtained (i.e., $n < 2$), which is probably due to the fact that the receiver nodes are located very close to ground floor, which provides a strong constructive reflected propagation path in addition to the direct one.

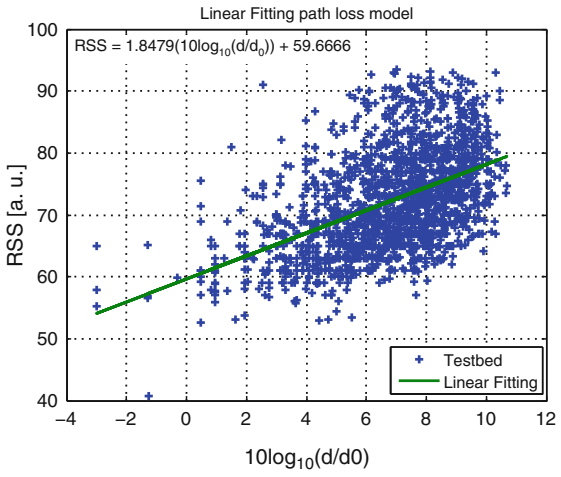


Fig. 3 RSS-to-distance ranging model in an empty conference room

4.2 System Setup

In order to try to overcome the issues related to the off-line RSS-to-distance ranging model calibration, i.e., assessing the performance degradation assuming a fixed propagation model in dynamically varying environment, we have deployed the testbed in the conference room during a half-day meeting. This testbed was composed of nine anchor nodes placed on the perimeter of the conference room, which dynamically estimates the propagation parameters with the solution already presented in [12], and a blind node placed in the middle of the room's area.

The event was characterized by four main phases, which well describe the dynamic nature of the event and, as a consequence, the dynamic nature of the propagation environment to be analyzed. The following is a brief description of each phase:

1. The first phase, which took place before the starting of the ceremony, is characterized by a progressive increase of the number of people inside the room, some of them very close and in motion around the blind node to be localized.
2. The second phase, which took place during the development of the ceremony, is characterized by several people (staying either seated or stand) inside the room, and some people coming in and going out of the room.
3. The third phase, which took place at the end of the ceremony, is characterized by the vast majority of people staying stand and leaving the conference room.
4. The fourth phase corresponds to the scenario with no people in the room, thus giving a virtually static indoor scenario with almost fixed propagation characteristics.

The data collected during this measurement campaign have been used as an input to the ESD algorithm implemented over the badge's microcontroller unit via an off-line emulation procedure.

4.3 Results CC2431

In order to understand the improvement of dynamic updating the channel-dependent parameters, we can look at Table 1. The following conclusions can be drawn. (a) For a fixed phase, the performance improves significantly when A and n are updated during the progress of the conference. (b) The improvement is more remarkable during phase two, which is a very dynamic phase and where the dynamic adaptation is more important. (c) The continuous training is also beneficial in phases one and three, but the improvement is less evident due to the short duration of these two phases. (d) When a fine estimation of the propagation parameters is available, using the ESD algorithm to refine the estimated position is beneficial to improve the accuracy. (e) The reason why the ESD does not improve the performance in the first case study is due to the fact that the ESD needs the RSSI-to-distance curve to refine the position. Since this curve is not updated continuously in the first case study, the algorithm

Table 1 Average positioning error [meters] over the observation time. The value shown into the parentheses represents the improvement that can be obtained refining the search with the ESD algorithm

Phase	Fixed Params	Dynamic Params
1	2.69 (3.70)	2.20 (1.90)
2	3.04 (5.48)	1.22 (1.01)
3	2.77 (3.36)	2.72 (1.90)
4	3.04 (1.70)	2.11 (1.25)

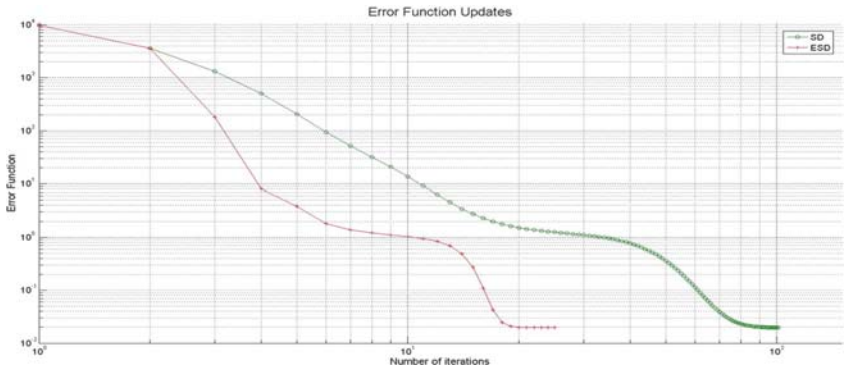


Fig. 4 ESD significantly improves the convergence speed of SD by reaching similar accuracy levels with a considerably less number of iterations

may diverge from the actual solution. This conclusion is also confirmed by the fact that in an almost static scenario (phase 4), the ESD improves the overall accuracy also without updating the channel-dependent parameters. It is worth to note that the values in Table 1 confirm results obtained in [12].

Finally, Fig. 4 obtained from a subset of data of our measurement campaign, compares the performance of the ESD with a MATLAB-based simulation of the SD algorithm assuming the same input data. It is evident that the accuracy of the final estimation of the ESD algorithm is similar to that obtained with an SD one, but ESD outperform SD in terms of the convergence speed.

5 Demo Description

The demo focuses on presenting the capabilities of the wireless biometric badge integrating localization and tracking functionalities along with an automatic personal identification mechanism, for, e.g., controlling the access to restricted areas with a high level of security. The wireless biometric badge is the result of the activities conducted by several professors, researchers, and students from the Center of Excellence in Research DEWS [11] at the University of L'Aquila in Italy, as well as its R&D Spin-Off, WEST Aquila S.r.l [13].

Fig. 5 Successfully authentication procedure with the biometric badge

The aim of the demo is to show two main functionalities of the wireless biometric badge. On the one hand, we show that every badge-provided user can be localized in a simple way by resorting to the aforementioned location algorithm, which runs in the badge's microcontroller unit [9]. Localization is performed in a distributed and decentralized fashion by receiving data from some fixed anchor/reference nodes, and by estimating the related distances based on received signal strength indication (RSSI) measurements [12, 14]. On the other hand, we show that when a user approaches an area with restricted access, the system can activate the authentication procedure of the user, thus either allowing (see Fig. 5) or denying him to have access to that area. This latter authentication procedure takes place via on-card fingerprint matching, thus avoiding any wireless transmission of sensible users' data.

References

1. Goldsmith AJ, Wicker SB (2002) Design challenges for energy-constrained ad hoc wireless networks. IEEE Commun Mag 9(8):8–27
2. Bachrach J, Taylor C (2005) Handbook of sensor networks – algorithms and architectures – localization in sensor networks. In: Wiley series. Wiley, New York
3. Mauve M, Widmer J (2001) A survey on position-based routing in mobile ad hoc networks. IEEE Netw 9(6):30–39
4. Bulusu N, Heidemann J, Estrin D (2000) GPS-less low-cost outdoor localization for very small devices. IEEE Wireless Commun 7(10):28–34
5. Slijepcevic S, Megerian S, Potkonjak M (2003) Characterization of location error in wireless sensor networks – analysis and applications. In: International workshop on information processing in sensor networks, pp 593–608
6. Tennina S, Di Renzo M, Graziosi F, Santucci F (2008) Distributed and cooperative localization algorithms for WSNs in GPS-less environments, The Italian Institute of Navigation (I.I.N.). In: Integration of navigation with communication and remote sensing applications, vol 18. pp 870–876

7. Tennina S, Di Renzo M, Graziosi F, Santucci F (2009) ESD – A novel optimization algorithm for positioning estimation of WSNs in GPS-denied environments – from simulation to experimentation, Int J Sensor Netw 6(3/4):131–156
8. Nocedal J, Wright S (2006) Numerical optimization, 2nd edn. Springer, New York
9. TI's website (2008) http://focus.ti.com/docs/prod/folders/print/cc2431.html
10. Aamodt K (2008) CC2431 location engine – Application note AN042, Chipcon products for Texas Instruments, pp 20
11. Centre of Excellence DEWS http://www.diei.univaq.it/dews
12. Tennina S, Di Renzo M, Graziosi F, Santucci F (2008) Locating ZigBee nodes using the TI's CC2431 location engine – a testbed platform and new solutions for positioning estimation of WSNs in dynamic indoor environments. In: Proceedings of the first ACM international workshop on mobile entity localization and tracking in GPS-less environments, San Francisco, CA, pp 37–42
13. WEST Aquila S.r.l. http://www.westaquila.com
14. Tennina S, Di Renzo M (2008) ESD – A novel optimization algorithm for positioning estimation in wireless sensor networks – analysis and experimental validation via a testbed platform. In: Proceedings of the 17th international conference on computer communications and networks, ICCCN, St. Thomas, US Virgin Islands, pp 1–7

Design and Implementation of Smartphone Applications for Speaker Count and Gender Recognition

Alessio Agneessens, Igor Bisio, Fabio Lavagetto, and Mario Marchese

1 Introduction

By combining the functions of mobile phones and PDAs, smartphones give mobile network providers the opportunity to come up with more advanced and innovative services. Among these are the so-called context-aware services, highly customizable services tailored to the user's preferences and needs and relying on the real-time knowledge of the user's surroundings, without requiring complex configuration on the user's part.

In order to provide context-aware services, a description of the smartphone's environment must be obtained by acquiring and combining context data from different sources, both external (e.g., cell IDs, GPS coordinates, nearby Bluetooth devices) and internal (e.g., idle/active status, battery power, accelerometer measurements). The number of active speakers in the smartphone's surroundings can be useful context information. Speaker count is applicable to numerous speech-processing problems (e.g., cochannel interference reduction, speaker identification, speech recognition) but it does not yield a simple solution. Several speaker count algorithms have been designed, both for closed- and open-set applications. Closed-set speaker count implies the classification of data belonging to speakers whose identity is known, while in the open-set scenario, there is no available a priori knowledge on the speakers.

Audio-based gender recognition also has many possible applications (e.g., for selecting gender-dependent models to aid automatic speech recognition, in content-based multimedia indexing systems and interactive voice response systems) and numerous methods have been proposed, involving a wide variety of features and classifiers.

Although in many cases promising results have been obtained, for both speaker count and gender recognition, available methods are not specifically designed for mobile device implementation, and thus their computational requirements do not

A. Agneessens, I. Bisio (✉), F. Lavagetto, and M. Marchese
Department of Communications, Computer and System Science,
University of Genoa, Genoa, Italy
e-mail: igor@dist.unige.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_18,

take into account smartphone processing power and context-aware application time requirements. The chapter is organized as follows. The proposed speaker count and Gender Recognition algorithms are described in Sect. 2. The algorithms' experimental results are presented in Sect. 3. The features of the Symbian smartphone applications implementing the proposed methods are described in Sect. 4. Conclusions are listed in Sect. 5.

2 The Proposed Methods: Pitch Definition and GMM Classification

The fundamental frequency of a periodic signal is defined as the reciprocal of its period. Audio signals such as speech exhibit a relative periodicity, and in this case the fundamental frequency, also referred to as pitch, is usually the lowest frequency component which relates "well" to most of the other components [2].

For voiced speech, pitch is usually defined as the rate of vibration of the vocal folds [1], so for this reason, it can be considered a reasonably distinctive feature of an individual. Estimating the pitch of an audio sample could therefore help classify it as either 1P or 2P: the basic idea is that if audio sample pitch estimates are all close in value, the sample is likely to be one-speaker, and it is likely to be two-speaker if different pitch values are detected.

Many pitch estimation algorithms have been proposed in past years, involving both time- and frequency-domain analysis [2]. Many developed methods are context-specific, but pitch estimators designed for a particular application depend on the domain of the data and are less accurate when applied to a different domain.

In this study, a method based on the signal's autocorrelation was chosen because of its good applicability to speech and ease of implementation Since pitch is linked to the signal's periodicity and the autocorrelation shows how well the signal correlates with itself at a range of different delays, given a speech sample with sufficient periodicity, its autocorrelation will present its highest values at very short delays and at delays corresponding to multiples of pitch periods [3]. To estimate the pitch of an audio frame, the frame's autocorrelation is first computed in the delay interval corresponding to a possible pitch range $[p_1, p_2]$. The peak of this portion of autocorrelation is then detected and the pitch is estimated as the reciprocal of the delay corresponding to the autocorrelation peak.

A Gaussian Mixture Model (GMM) is a PDF given by the sum of weighted Gaussian PDFs. Each class is represented by its own GMM, whose parameters (the components' mean vectors, covariance matrices and weights) are computed based on a training set of that class's feature vectors. GMM training is carried out with the Expectation–Maximization (EM) method, which is used with training sets originated from probability distribution mixtures when no information is available on either the assignment of the data points to the mixture components, or the parameters of any component of the mixture. Once GMMs have been trained for all classes, a given unknown feature vector can be classified by evaluating the various GMMs and selecting the class with the greatest GMM value for that feature vector.

2.1 Speaker Count

The classification of an audio sample consists of several steps. The signal is first divided into L-sample abutted frames, and a pitch estimate is computed for each frame by applying the method described above. Sets of D consecutive frames are grouped together in blocks, in order to allow the computation of a pitch Probability Distribution Function (PDF) for each block. Consecutive blocks are overlapped by V frames (i.e., the last V frames of a block are the first V of the following one) in order to take into account the possibility of a signal section representing fully voiced speech falling across consecutive, nonoverlapping blocks, and therefore its contributions to the classification process being divided between the two blocks. Each block is classified based on the dispersion of the PDF of its pitch estimates. The block PDFs span a frequency interval corresponding to admissible pitch values, and such interval is divided into a certain number of smaller frequency bins. In order to compute the PDF of a block, the pitch is estimated for each of the block's frames and the Fourier transform of each frame is computed as well. The contribution of each frame to the block's PDF is given by the frame's power (obtained from the frame's Fourier transform) at the frequency corresponding to its pitch estimate. The value of the PDF bin containing the frame's pitch estimate is therefore increased with such power. After all pitch estimates have been accounted for, the PDF is normalized by the sum of all bin values.

This method led to more distinct PDFs and more accurate features, thus significantly improving the method's performances when compared with computing PDFs by simply executing a "histogram count" (which increases by 1 the value of a PDF bin for every pitch estimate falling into such bin). From each block's PDF, a feature vector is extracted and subsequently used by a GMM classifier to classify the block as either 1P or 2P. Once all individual blocks have been classified, the audio sample is classified through a "majority vote" decision (the chosen class is the one which received the most blocks). Different feature vectors were evaluated by combining different individual features representing the block PDF dispersion, i.e., the PDF maximum (*Max*), the PDF standard deviation (*StdDev*), the PDF roll-off (i.e., the frequency interval containing a given percentage of the PDF's area, *ROff*) and the absolute value of the difference between the PDF mean and the pitch corresponding to the maximum PDF value (*Abs*). Each evaluated feature combination was used to train a separate GMM classifier, and the feature vector leading to the best test set classification (see Sect. 3) was selected as the ultimate feature vector.

2.2 Gender Recognition

In addition to the speaker count algorithm, a method for single-speaker gender recognition was designed as well. During the course of the study, it was observed that satisfying results could be obtained by using a single-feature threshold classifier, without resorting to GMMs and individual frame Fourier transforms as in speaker count.

The chosen feature is the mean of the blocks' "histogram count" PDF. In fact, pitch values for male speakers are on average lower compared to female speakers, since pitch can be defined as the vibration rate of the vocal folds during speech and male vocal folds are greater in length and thickness compared to female ones.

"Histogram count" PDFs are employed because it was observed that the derived feature was sufficiently accurate, so it was decided not to use the weighted PDFs (which require the Fourier transform of individual frames) as in the speaker count method in order to significantly reduce the time required by the smartphone application to classify unknown samples. The gender recognition of an audio sample is identical to the speaker count problem as far as the subdivision of the audio sample in frames and blocks and the frame-based pitch estimation. Gender recognition differs from speaker count in individual block classification, which no longer involves Fourier transforms of all frames and GMM classification of blocks, but simply requires the computation of the mean of the blocks' "histogram count" PDF (see Sect. 2.1). Individual blocks are classified as "Male" (M) or "Female" (F) by comparing their PDF mean with a fixed threshold computed based on a training set.

3 Experiments and Results

Training and testing of the classifiers were carried out using a database of audio samples acquired with a smartphone audio-recording application, thus allowing the development of the proposed methods based on data consistent with the normal execution of the smartphone applications. Five different situations were considered: 1 male speaker (1M), 1 female speaker (1F), 2 male speakers (2M), 2 female speakers (2F) and 2 mixed speakers (2MF). The database was acquired using a 22 kHz sampling frequency and 16 bits per sample, and all recordings are 4.5 s long. All audio samples refer to different speakers in order to evaluate classifier performances using data deriving from speakers that did not influence classifier training (open-set application). A total of 50 recordings was acquired, 10 for each situation.

For the speaker count classifier, half of the recordings for each of the five situations was used for GMM training, the other half for testing. For gender recognition, half of the single-speaker situations were used to compute the decision threshold, the other half for performance evaluation.

During the experiments, the frame size L was set to 2,048 samples, large enough to allow reliable pitch estimation, but not too large to produce (for an audio signal of fixed length) an inadequate amount of blocks to use for the majority rule classification. The block size D was set to 20 frames, which was considered a large enough number of frames per block to compute the block PDFs with, while not large enough to produce (for an audio signal of fixed length) an insufficient number of blocks. The block overlap V was set to ten frames, a trade-off between having many, heavily-overlapped blocks (which implies consecutive blocks bearing redundant information and added computational load) and having few, slightly-overlapped blocks (with the risk of having signal sections representing fully voiced

speech fall across consecutive blocks). As for PDF computing, pitch values in the range [50 Hz, 500 Hz] were considered as suggested in [3] and individual bins represent intervals of approximately 10 Hz.

3.1 Speaker Count Results

Different feature vectors were evaluated, and comparison of test set classification accuracies was used to select the most discriminating one. The feature vector ultimately used for GMM classification is [*Max*, *StdDev*]. It performs slightly better than the other possible vectors. In Fig. 1, the confusion matrixes with the classification results of test sample blocks (a) and whole test samples (b) are shown. Whole test sample classification accuracy is higher than block classification accuracy because the "majority vote" decision employed for whole sample recognition may neutralize the effect of misclassified blocks.

To enhance the speaker count classifiers performance, an additional set of experiments addressing the speaker count problem ignoring the situations that most of all led to classification errors, i.e., 2M and 2F, was carried out. In fact, samples belonging to these two situations can be misclassified as 1M and 1F respectively, since same-gender speakers could have pitch estimates close enough in value to lead to 2P PDFs similar to 1P PDFs of the same gender. Therefore, a new GMM classifier was designed in order to distinguish not two classes (1P and 2P) but three classes: 1M, 1F, and 2MF. The [*Mean*, *Max*] vector has been chosen. The only classification errors involve exclusively class 2MF, i.e., test sample blocks belonging to classes 1M and 1F were never mistaken one for another. In Fig. 2, the confusion matrixes with the classification results of test sample blocks (a) and whole test samples (b) are shown.

Figure 3 displays the classification results shown in Fig. 2b mapped to the two-class (1P and 2P) speaker count, for a better comparison with the results shown in Fig. 1b. As can be seen, ignoring 2M and 2F samples did indeed lead to better speaker count performances.

a

Actual \ Predicted	1P	2P
1P	56	44
2P	47	53

Correct (average): 55

b

Actual \ Predicted	1P	2P
1P	60	40
2P	60	40

Correct (average): 60

Fig. 1 Classification results of test sample blocks (**a**) and whole test samples (**b**)

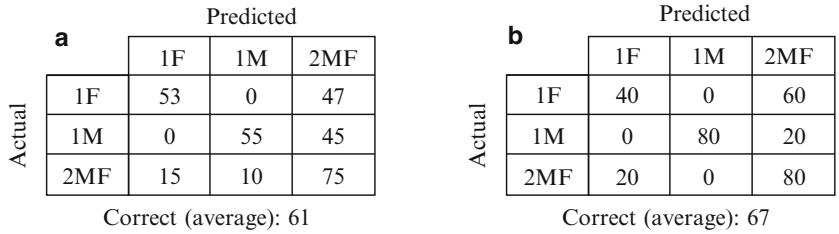
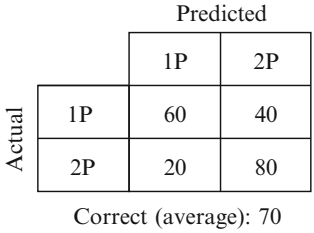


Fig. 2 Classification results (percent) of test sample blocks (**a**) and whole test samples (**b**)

Fig. 3 Classification results of whole test samples mapped to the two-class speaker count



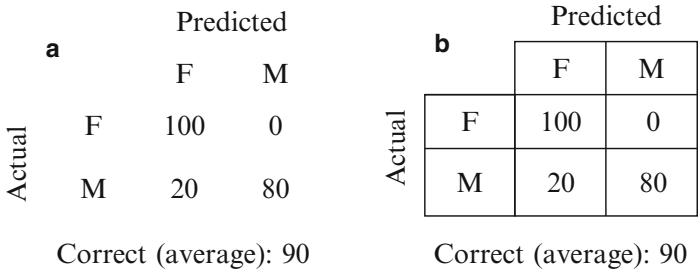


Fig. 4 Gender recognition results for test sample blocks (**a**) and whole test samples (**b**)

3.2 Gender Recognition Results

Figure 4a displays the test sample block classification results (in percentage) for the gender recognition problem. The chosen threshold on the mean of the "histogram count" pitch PDF was the one that led to the best training sample block classification results. The confusion matrix with the classification results of whole test samples is shown in Fig. 4b. As can be seen, while both classes are well-recognized, the totality of female speaker blocks and whole samples were correctly classified.

4 Smartphone Implementation

Two separate Symbian OS applications implementing the speaker count and gender recognition methods have been designed as part of this study. Various measures have been taken to reduce the applications' computational loads. Since most of the computational time was initially spent on frame autocorrelation computing, it was

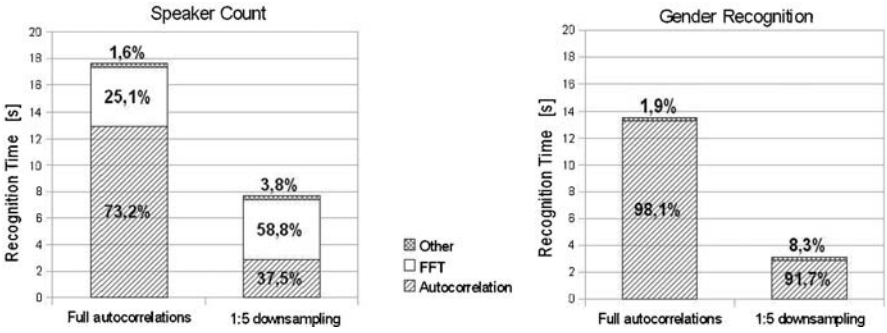


Fig. 5 Recognition time for speaker count and gender recognition applications, using full autocorrelations and autocorrelations downsampled by a factor of 5

decided not to compute the full autocorrelations, but with a downsampling factor of 5, thus greatly reducing the amount of autocorrelation samples to be computed. In order to still correctly determine the maximum of the full autocorrelation, thus preventing errors in subsequent pitch estimation, a maximum "fine search" method was designed. Firstly, it determines the maximum of the downsampled autocorrelation, and then computes the full autocorrelation samples adjacent to the downsampled autocorrelation maximum (and ignored during the initial downsampled autocorrelation computing) and compares them with the downsampled autocorrelation maximum, thus finding the maximum of the full autocorrelation.

Using full frame autocorrelations, the speaker count and gender recognition smartphone applications required approximately 17.6 and 13.5 [s], respectively, to classify samples of 4.5 [s]. With the autocorrelation maximum "fine search" method, it was possible to bring the recognition time down to approximately 7.7 and 3.1 [s], as shown in Fig. 5. The difference between the two applications is due to the Fourier transform computation for all frames required for speaker count.

5 Conclusions

A speaker count method for context-aware smartphone applications designed to recognize single-speaker samples from two-speaker samples and to operate in an open-set scenario is presented in this paper. A method for single speaker gender recognition is also described. Both proposed methods have been designed and implemented as Symbian OS smartphone applications by considering the small computational load is a fundamental requirement for mobile device implementation.

References

1. de Cheveignè A, Kawahar H (2002) A fundamental frequency estimator for speech and music. J Acoust Soc Am 111(4):1917–1930
2. Gerhard D (2003) Pitch extraction and fundamental frequency: history and current techniques. Ph.D. Dissertation, Department of Computer Science, University of Regina
3. Web site: http://www.phon.ucl.ac.uk/courses/spsci/matlab/lect10.html. (2009) University College of London, Department of Phonetics and Linguistics

Video Coding with Motion Estimation at the Decoder

Claudia Tonoli, Pierangelo Migliorati, and Riccardo Leonardi

1 Introduction

Compression efficiency is a fundamental requirement in all video coding systems. It becomes even more important in the case of small wireless devices because they are often subject to tighter power and bandwith constraint. In predictive video coding schemes, the key feature to achieve compression efficiency is motion estimation. The basic idea of this approach is to exploit the temporal redundancy across frames, estimated using the motion information. Usually motion estimation is performed at the encoder side, and then the motion field is transmitted to the decoder, together with the compressed prediction error. The decoding of each block of a frame consists in simply extracting the predictor, which is identified thanks to the motion vector, from the reference frame and adding the prediction residue. Both the motion vector and the residue are computed by the encoder. The decoding process, that is applied blockwise, cannot prescind from their complete transmission.

Despite that such systems are based on a blockwise decoding, it is not true that each block of a frame is independent from its neighbors. On the contrary, the structure of natural images generally imposes a strong spatial correlation among adjacent blocks, but such spatial correlation is not completely exploited in traditional system. A coding technique, able to reduce the redundancy between the already decoded part of the frame and the motion information, would improve the compression efficiency, possibly completely discarding the motion information from the transmitted bitstream.

Thanks to arithmetic coding and prediction techniques, motion information is nowadays compressed very efficiently. Nevertheless, especially for low bit rate video coding, the motion field still has a nonnegligible impact on the overall bit-rate. The idea of skipping the transmission of the motion information and reestimate it at the decoder has recently attracted an increasing interest. For example, in [1] an algorithm for motion derivation at the decoder side for the H.264/AVC codec is presented. This algorithm is based on a template similar to those used in texture coding.

C. Tonoli (✉), P. Migliorati, and R. Leonardi
Department of Electronics for Automation, Signals and Communication Lab, University of Brescia, Brescia, Italy
e-mail: claudia.tonoli@ing.unibs.it; pierangelo.migliorati@ing.unibs.it; riccardo.leonardi@ing.unibs.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_19,

In this chapter, we propose a method for motion estimation at the decoder. The proposed approach relies on the knowledge of the prediction residue, transmitted by the encoder, and it is based on Least Square Error prediction. Preliminary simulation results seem to be very promising.

The chapter is structured as follows. In Sect. 2, a brief description of the use of motion compensation in predictive video coding schemes is given. The proposed algorithm is described in detail in Sect. 3, and simulation results are presented and discussed in Sect. 4. Concluding remarks are given in Sect. 5.

2 Motion Compensation in Traditional Video Coding Schemes

Predictive video coding is based on motion estimation at the encoder and motion compensation at the decoder. In this section, the basic ideas of predictive video coding are briefly introduced. The highlighted details will be useful in the following. For a more complete description of this topics, we refer the reader to [2–4].

In predictive coding, the suitable predictor for each block is determined at the encoder, usually performing a block based motion estimation. The prediction residue, i.e., the difference between the current block and its predictor, is computed and encoded. The information sent to the decoder includes the residue, together with the motion field.

The decoder reconstructs each frame operating in a strictly blockwise mode, since each block is reconstructed independently from its neighbors. The motion vector associated to the current block is used as an index for the set of possible predictors. Once the correct predictor has been identified, the prediction residue is decoded and added to the prediction values.

This method obviously requires that one motion vector is transmitted for each block. Due to the efficiency of modern entropy coding techniques, the transmission of the motion field is in general not very expensive in terms of bit-rate. Especially in high bit-rate coding, where DCT coefficients quantization is very fine, motion information represents a small part of the overall transmitted rate. Nevertheless, in low bit-rate coding the amount of rate assigned to the signal coefficient is lower, so the motion rate becomes much more important.

3 The Proposed Algorithm for Motion Estimation at the Decoder

In this section, the proposed method of motion estimation at the decoder side is presented.

First of all, the fundamental ideas are introduced, focusing on the definition of the side information. The algorithm is then outlined, sketching a structure that can be applied with different spatial coherence evaluation parameters. Finally, the LSE based parameter is introduced, and its computation is described in some detail.

3.1 Decoder Side Information

The term "Side Information" can be found very frequently in recent papers on video coding, and it often acquires different meanings, depending on which specific field we are looking at. To a very general extent, it refers to pieces of information that are not exactly the values of the coded signal, but a somewhat higher level correlated information, which is indeed crucial for the proper decoding of the signal. In particular, the centrality of the concept of side information and the way the side information is dealt with is one of the distinguishing elements of the Distributed Video Coding (DVC) paradigm (see, for example, [5, 6]). In this paradigm, each frame is encoded independently from the others. Due to such assumption of independent frame coding, the motion estimation is not performed at the decoder, and the motion field has to be inferred at the decoder side, based on the side information.

Borrowing the concept of side information from Distributed Video Coding, in this work, we assume that the side information for the current frame corresponds to the previously decoded frames. In more detail, since the encoder motion compensation algorithm is known, when the decoder begins to decode a frame block, it already has some knowledge about that block, in terms of correlated information. It is known for sure that the motion compensated predictor for that block belongs to the block matching reference frame, and, more precisely, to the search window. In fact, the motion vector in motion compensation behaves exactly as an index for the set of predictors corresponding to the search window. Since the reference frame has already been decoded, the set of candidate predictors for the current block is completely known. Equivalently, it is possible to say that the final reconstructed block will be one of these candidate predictors corrected with the received prediction residue for that block.

Moreover, we introduce an a priori hypothesis that, despite its generality, turns out to be true in the great majority of cases. We assume that the signal to be coded is characterized by "spatial continuity," i.e., edges preserve their continuity across the block boundaries. This means that, given the neighborhood of a block, it is possible to infer that the more suitable predictor in a candidate set will be the one that matches at best the neighborhood edges. See Fig. 1 for an example of a well matched and a bad matched predictor, respectively. If we assume that block decoding is performed in raster scan order, the causal neighbors of the current block have already been decoded. Therefore, it is possible to use the information carried out by the position of their edges to try to match the candidate predictors.

These remarks about the side information role will be the basis for the selection of a predictor for motion compensation in the absence of the motion vector, and for the consequent motion estimation at the decoder.

3.2 Outline of the Predictor Selection Algorithm

Let us consider a predictive coding scheme based on motion compensation, as described in Sect. 2. Our aim is to avoid the transmission of motion vectors,

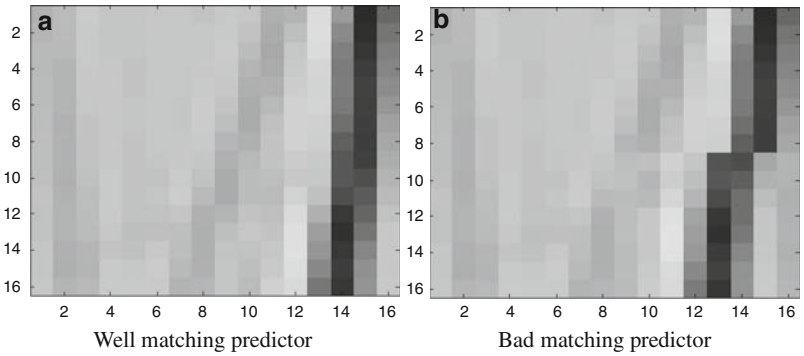


Fig. 1 Spatial continuity at block edges. (**a**) Well matching predictor. (**b**) Bad matching predictor

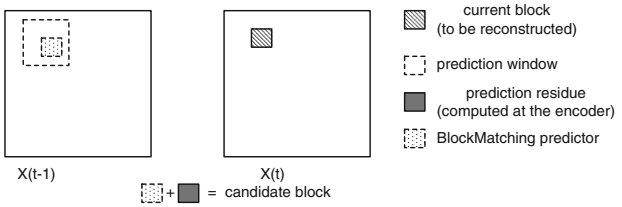
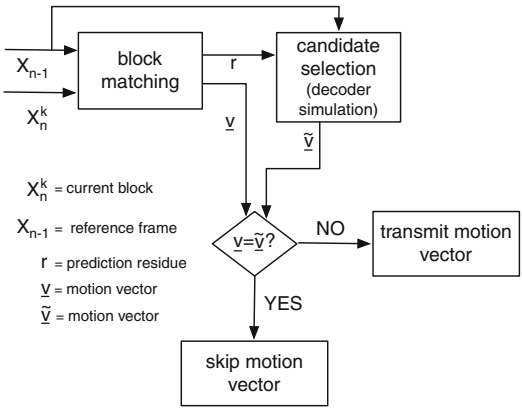


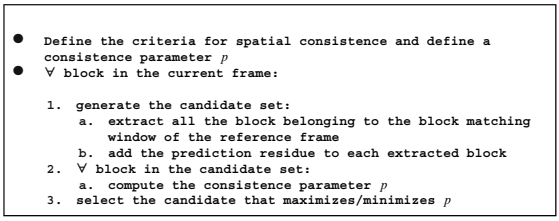
Fig. 2 Candidate set generation

nevertheless achieving a reconstruction quality close to that obtainable in case of transmission of the whole motion field. In order to try to do that, we apply the principles about side information described in Sect. 3.1.

For each block to be reconstructed, the set of candidate predictors is generated, as depicted in Fig. 2. A ranking of the candidates is then performed, in order to find out which candidates fits at best the coherence conditions, as described in the following lines. The already decoded causal neighborhood of the current block is considered. The macroblock composed by the current block and its three causal neighbors is constructed, replacing the current block with the tested candidate. A parameter p measuring the matching of the candidate block with the neighborhood is computed, in order to select the predictor that guarantees the best matching with the side information. Obviously, the matching parameter plays a crucial role in the algorithm performance, since it has to capture the matching of each candidate predictor and to select the most suitable one. In this chapter, we present a method based on Least Squared Error prediction. Such method will be described in more detail in Sect. 3.3. The reason why LSE prediction has been chosen to highlight the spatial coherence is that such technique is based itself on the exploitation of the correlation among adjacent pixels. The presented algorithm relies on the principle that a block correlated to the given neighborhood should be well predictable from the neighborhood, while a less correlated block should produce a greater prediction error (Fig. 1).

Since we want to control at the encoder side the quality of the reconstructed signal, as it usually happens in predictive coding schemes, we apply our method





- Define the criteria for spatial consistence and define a consistence parameter *p*
- ∀ block in the current frame:

1. generate the candidate set:
   a. extract all the block belonging to the block matching window of the reference frame
   b. add the prediction residue to each extracted block
2. ∀ block in the candidate set:
   a. compute the consistence parameter *p*
3. select the candidate that maximizes/minimizes *p*

Fig. 3 Scheme of the proposed system

Fig. 4 Algorithm structure

first at the encoder. In detail, for each block, the encoder simulates the operations of the decoder, and, based on the quality of the reconstructed block, decides whether the motion vector for that block is omissible or not. In our implementation, a simple threshold on the quality of the reconstructed block has been applied. A more precise rate-distortion analysis, like the one performed for example in the H.264/AVC encoder, could lead to a performance improvement because the effect of a motion vector skip on the overall reconstructed PSNR could be estimated more precisely.

3.3 Candidate Selection Based on Least Square Error Prediction

In the framework described in Sect. 3.2, in the absence of the motion information, the only criterion for the decoder to select one block among the candidates is the good match with the intra side information, i.e., the neighborhood.

The decoder-based motion compensation algorithm has been implemented according to the steps described in Fig. 4. As stated in step 2, each candidate needs to be tested in order to produce the parameter p, i.e., a "measure" of the correlation of that block with the known neighbors, and to get a ranking of the candidates. The steps to be performed to obtain such ranking are listed below.

1. Test each candidate block in the following way:
   (a) Prediction of the upper left quadrant; for each pixel, the correlation matrix is reestimated, based on the neighbors and on the true value of the past pixels in the block (i.e., the predicted pixels in the past are not considered in the estimation)
   (b) Compute the MSE on the upper left quadrant between the predicted block and the true candidate block
2. Choose the candidate that results to be more predictible, i.e., such that the MSE computed at the step 1 is smaller than that obtained for any other candidate (p = MSE).

The LSE prediction has been implemented as described in [7], where it is used to perform still image compression: each pixel is predicted, based on its causal neighborhood, and the prediction residue is encoded and transmitted. The prediction is shown to be orientation adaptive.

The LSE prediction computation is now briefly reported. Further details are given in [7]. For each pixel, a training window is set, as depicted in Fig. 5. According to the training values, the prediction coefficients are adaptively computed. The correlation matrix is estimated as described in the following.

$$c_i = [x_{i-1}, x_{i-2}, \ldots, x_{i-L}] \tag{1}$$

where x_{i-j} is the jth causal neighbor of x_i, for $i = 1, 2, \ldots, L$.

A $W \times L$ matrix, whose rows are the c_i vectors, is created:

$$C = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_W \end{bmatrix} \tag{2}$$

The covariance matrix is then computed as:

$$R_{xx} = C^T C; \tag{3}$$

while the covariance vector r_x is computed as

$$r_x = C^T y \tag{4}$$

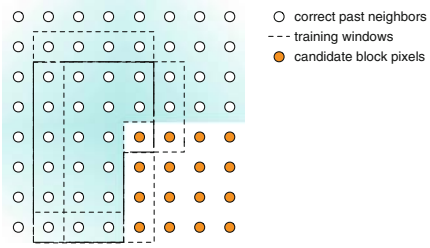


Fig. 5 Sliding training window

where

$$y = [x_{n-1}, \ldots, x_{n-L}]^T \tag{5}$$

According to the theory of least squared error prediction, the coefficient vector a is computed as

$$a = \left(C^T C\right)^{-1} \left(C^T y\right) \tag{6}$$

The main drawback of this algorithm is its huge computational complexity, due to the frequent matrix inversion operations that are needed. In the literature, several techniques have been presented to reduce the complexity [8, 9]. In our implementation, the edge based technique presented in [7] has been used. It can be seen that the complexity can be reduced with a performance impairment of about 1% more wrong block.

4 Experimental Results

In this section, the performance of the proposed algorithm is presented and discussed.

In order to evaluate the proposed method, the percentage of correctly reconstructed motion vectors has been computed. As a groundtruth reference, the lossless case is considered. On the original CIF format sequence, the block matching is performed, on blocks of size 16×16, as a means to compute the motion field and the prediction residue. Since the work presented in this paper is aimed at exploring a new field, many optimizations have not been introduced yet: no multiple reference is considered for the block matching, and the reference for each frame is the previous frame.

When the prediction residue has been computed, the motion estimation method is applied and the percentage of correctly estimated motion vectors is computed for each frame. It is worth to remark that no error propagation is taken into account in the presented results. It is always assumed that the encoder controls the decoding process and, when a block cannot be correctly estimated in the absence of motion vector, the motion information is transmitted. So the neighbors of a block are always correct, either because their motion vector has been estimated correctly or because motion has been transmitted.

In Table 1, the results in terms of percentage of correct blocks is reported for the first two frames of four test sequences, namely Foreman, Mobile, Highway,

Table 1 Percentage of correctly predicted blocks in the lossless case

Sequence	Correctly estimated motion vectors (%)
Foreman	75.93
Mobile	34.67
Highway	84.92
Harbour	41.22

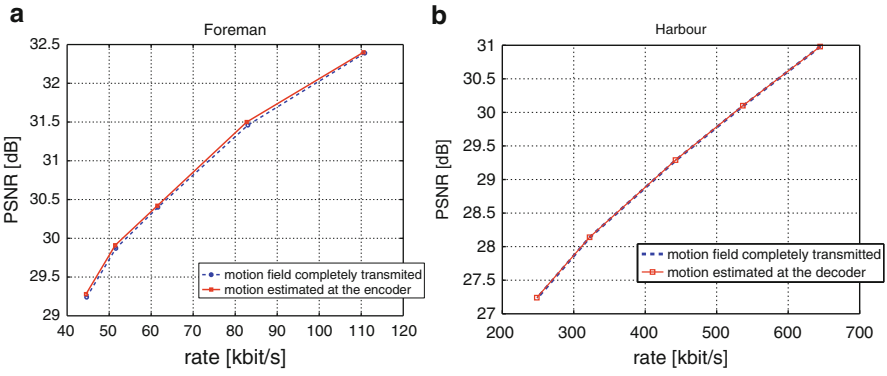
and Harbour. It can be noticed that the performance is strongly dependent on the sequence content. For Highway and Foreman, very good results can be achieved, whereas for Mobile and Harbour, the algorithm is less effective.

In order to give an idea of how the presented algorithm could perform in a realistic scenario, it has been applied to lossy coding. In particular, low bit rate coding has been considered because in this case, skipping the motion information could be particularly advantageous.

In more detail, the rate and PSNR values for the case of transmission of the whole motion field have been obtained using a simplified H.264 codec. The block size has been set to 16 and the considered prediction mode is P, i.e., mono-directional prediction, with a single reference picture. No deblocking has been performed on the reconstructed frame. An important remark has to be given about the rate estimation. The coding efficiency in modern predictive codec, such as H.264 codec, depends heavily on how arithmetic coding is performed. Since our method has not been really implemented in H.264 yet, it is impossible to measure exactly the rate savings. In order to produce a reliable estimate, the bits devoted to the motion transmission for each block have been computed, and, for the correctly predicted block, the result has been subtracted from the overall bit-rate. A signalling overhead has also been taken into account.

The estimated performance of the considered method is reported, for the first ten inter-frames (i.e., frames from 2 to 11, since the first frame is intra-coded) of the Foreman and Harbour sequence in CIF format, at 15 fps, in Fig. 6a and b.

In order to help a more precise performance assessment, the percentage of skipped motion vector is also reported, in Table 2. In the case of lossy coding, it can happen that the selected motion vector is not the correct one, but the selected candidate is not too different from the correct block. In this case, it can be seen that the proposed method can lead to a slight performance improvement.



PSNR curves for the Foreman sequence

PSNR curves for the Harbour sequence.

Fig. 6 PSNR curves for the test sequences Foreman and Harbour. (**a**) PSNR curves for the Foreman sequence. (**b**) PSNR curves for the Harbour sequence

Table 2 Percentage of correctly predicted blocks for the lossy compression of the Foreman sequence

PSNR	Percentage of skipped motion vectors
31.51	21.03
30.42	15.0
29.91	14.69
29.28	17.50

5 Conclusions

Compression efficiency is particularly important in small, low power devices. Side information at the decoder side, i.e., correlated information about the signal that has to be decoded, can be exploited to improve compression efficiency in predictive video coding. Starting from the side information, the encoder can infer important knowledge that helps in decoding the signal. As an example of how side information at the decoder side can be exploited in video coding, in this chapter, we have proposed a method that partially avoids the transmission of motion vectors in predictive video coding schemes based on motion compensation. Simulation results show that the proposed approach can lead to bit-rate savings.

References

1. Kamp S, Evertz M, Wien M (2008) Decoder side motion vector derivation for inter frame video coding. In: Proceedings of 15th IEEE international conference on image processing 2008 (ICIP 2008), 12–15 Oct 2008, pp 1120–1123
2. Wiegand T, Sullivan G, Bjontegaard, G (2003) Overview of the H.264/AVC video coding standard. IEEE Trans Circuits Syst Video Technol 13(7):560–576
3. Wiegand T, Sullivan GJ, and Luthra A (2003) Draft ITU-T recommendation and final draft international standard of joint video specification, Joint Video Team Doc. JVT-G050r1, June 2003
4. Kappagantula S, Rao KR (1985) Motion compensated interframe image prediction. IEEE Trans Commun 33(9):1011–1015
5. Aaron A, Zhang R, Girod B (2002) Wyner-Ziv coding for motion video. In: Proceedings of 36th Asilomar conference on signal, systems and computers (ACSSC '02), Pacific Grove, CA, vol 1, pp 240–244
6. Puri R, Ramchandran K (2003) PRISM: a new reversed multimedia coding paradigm. In: Proceedings of the IEEE internatinal conference on image processing, Barcelona, Spain, Sept 2003
7. Li X, Orchard MT (2001) Edge directed prediction for lossless compression of natural images. IEEE Trans Image Process 10(6):813–817
8. Li X, Orchard MT (2002) Novel sequential error-concealment techniques using orientation adaptive interpolation. IEEE Trans Circuits Syst Video Technol 12(10):857–864
9. Wu X, Barthel K (1998) Piecewise 2D autoregression for predictive image coding. In: Proceedings of the IEEE international conference on image processing, Chicago, IL, Oct 1998, vol 3, pp 901–904

Inter-Vehicle Communication QoS Management for Disaster Recovery

P. Orefice, L. Paura, and A. Scarpiello

1 Introduction

The chapter deals with the QoS management in a relief communication network that enables interoperability among rescuers: different vehicles, including aerial ones, are operating in a disaster recovery scenario, and a multiplicity of services with rigorous requirements is required. Further, private and/or public communication infrastructures at the crisis site are compromised or completely out of order and, often, especially during the first beginning phases, the anarchy reigns over every rescue action due to the panic of the people. In such a scenario, the first help and the first actions can be more and more effective if detailed and organized information on the state of the sites involved in the event can be acquired in a short time and effectively distributed among the rescuers. At this end, an aerial vehicle flying over the area affected by disaster, can provide real-time transmission of data toward a Mobile Ground Station (MGS) that is located near this area and interconnected with several terrestrial communication networks. In this context, different access wireless technologies have to interoperate seamlessly to guarantee that each rescuer can perceive a given satisfactory quality of service. Such a challenging goal is typical also of the Next Generation Network (NGN) paradigm but in the considered scenario, it must be reached in a very hostile environment characterized by both time availability and preexistent infrastructures shortage.

As regards the public safety and the disaster response, an international partnership between ETSI and TIA, referred to as the MESA Project [1, 2], has defined globally applicable technical specifications for digital mobile broadband technology. In [3], it is presented a system architecture for the interoperability and integration among Private Mobile Radio (PMR) systems (TETRA), public communication

P. Orefice (✉), L. Paura, and A. Scarpiello
Laboratorio Nazionale di Comunicazioni Multimediali, CNIT, via Cinthia,
Monte S. Angelo, 80126 Napoli, Italy
e-mail: paolo.orefice@cnit.it; amedeo.scarpiello@cnit.it

L. Paura
Dipartimento di Ingegneria Biomedica, Elettronica e delle Telecomunicazioni,
Università di Napoli Federico II, via Claudio 1, 80125 Napoli, Italy
e-mail: paura@unina.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_20,

networks (GSM/GPRS/UMTS), and broadband wireless technology (WiFi and WiMax), all operating in a Public Safety and Disaster Recovery scenario. In particular, in order to optimize the QoS management, the authors propose the use of appropriate mapping strategies among service classes supported by the different wireless systems involved in the integrated network architecture. This solution, however, requires the implementation of $n!/[(n\text{-}2)!\cdot 2!]$ mapping tables, where n is the number of different wireless access technologies. Moreover, the *entry* of a new access technology requires a rather complex procedure of upgrading since the implementation of other mapping tables among services class of the new technology and the preexisting ones must be performed. Finally, in [4], it has been proposed an interesting mobile ad-hoc satellite wireless mesh networking approach designed for an emergency scenario, in which the full mobility of rescue teams at the disaster site represents one of the major requirements for an emergency communication system. The combination of ad-hoc mobility together with IPv6 mobility mechanisms gives seamless mobility in the disaster site to rescue units. However, due to the satellite link, this solution is both very sensitive against weather conditions and very expensive for bandwidth usage.

In our proposal, the QoS provisioning in heterogeneous Relief Network (RN) is accomplished by resorting to both a specific DiffServ-based procedure and appropriate mapping strategies among DiffServ and the service classes supported by each wireless access technology. Such an approach allows one to benefit from the specific QoS management capabilities of each technology and to guarantee, at the same time, the scalability property against the number of access technologies involved in the whole network. To realize the proposed QoS Management Architecture (QMA), two steps are required:

1. Implementation of the DiffServ scheme for which we resort to usage and customization of Open-source Linux-based packages for the Advanced Routing and Traffic Control [5, 6]
2. Developing of software modules to implement mapping tables which are integrated with DiffServ scheme in order to accomplish a fully QMA.

It is worthwhile to underline that we used Commercial Off-the-Shelf (COTS) components to assure a low-cost solution for our QMA.

The chapter is organized as follows. In Sect. 2, the envisaged disaster recovery scenario is described and a RN architecture is presented. Section 3 is focused on the QoS-provisioning proposed architecture and its implementation. In Sect. 4, performance results are reported. In Sect. 5, conclusions are drawn.

2 Disaster Recovery Scenario and Relief System Architecture

Since a disaster occurs, relief vehicles move toward the crisis site and reaching the most critical areas of the disaster. It is a very useful support to get real-time information about the crisis site by means not only the satellite network but also by a

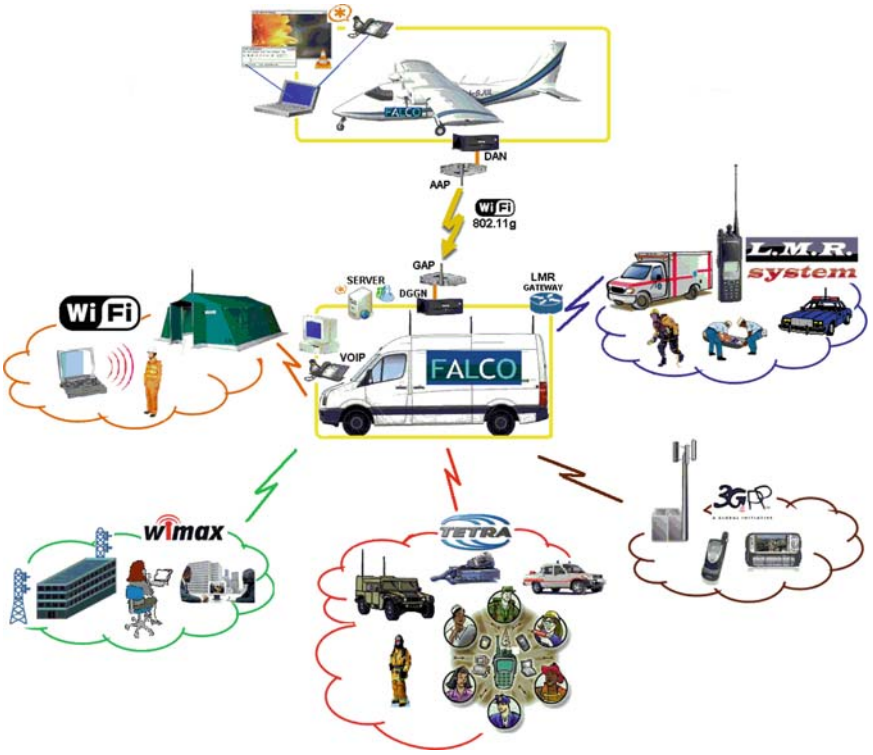


Fig. 1 Disaster recovery scenario and relief system architecture

properly equipped aircraft with camera and sensors. It flies over the area affected by disaster and transmits real-time video, images, and data to a mobile ground station, a vehicle located near the struck area. This station is a kind of headquarters that provides support both to local rescuers and remote public safety agencies, thanks to several types of wireless connections. Figure 1 depicts the envisaged disaster recovery scenario and the system architecture specifically designed for this scenario.

The air-MGS link is based on IEEE 802.11 g wireless protocol and uses rugged Air and Ground Access Point (AAP and GAP). On aircraft, there is a laptop that runs a real-time video streaming server, Video LAN Client (VLC), and a VoIP softphone to allow on board operator to talk with the rescue units operating on ground in the injured area. In addition, some sensors, such as photogrammetric and infrared cameras, are installed to transmit images of struck area to FTP server located on MGS. These images are stored in FTP server and available for download by relief units located in far sites. On MGS, there is a rack server to provide VoIP, FTP, chat services (respectively Asterisk, FileZilla, FreeCS) to all rescue units involved in crisis management that require them. To provide conversational voice service to rescue units, it is utilized a LMR Gateway that links to existing LMR systems making critical adaptation of LMR audio and signalling to IP.

Finally, to accomplish QoS features according to our scheme, properly configured Linux Boxes have been adopted, named DiffServ Air Node (DAN) on aircraft and DiffServ Gateway Ground Node (DGGN) on MGS. More specifically, DAN has two LAN interfaces, one links to laptop, the other one links to AAP LAN interface. Similarly, DGGN, links from one side to GAP, through a LAN interface, and from the other side to MGS LAN that provides network connection to all IP devices. In addition, DGGN has specific interfaces to interconnect different access wireless technologies (Wi-Fi, WiMax, TETRA, 3GPP/HSDPA).

3 QoS-Provisioning Proposed Approach

Envisaged relief services such as VoIP, real-time streaming video, bulk and small data transfer via FTP, Telnet and remote control require different QoS treatments. In particular, VoIP is delay time, jitter and packet loss sensitive; real-time video streaming, instead, consumes a large amount of bandwidth but it is jitter tolerant and less sensitive than VoIP against the delay and packet loss. Finally, other envisaged relief services are based on TCP transport protocol, and therefore are more tolerant to packet loss, delay and jitter. In the literature, several methods have been proposed to provide quality of services control on IP networks. Such solutions are not, however, fully suitable whenever they must be adopted in a disaster recovery context. In particular, the DiffServ strategy, though it is a promising approach for its scalability and aggregation capabilities of traffic flows, does not fit well in this scenario of rescue because it doesn't allow to exploit at best the native QoS capabilities and the characteristics of each wireless technology involved in RN. So, to achieve a certain degree of quality for these different relief services, prioritized packet scheduling over efficient DiffServ nodes and opportune QoS mapping strategies are required.

3.1 The Proposed QoS Management Architecture

The main goal of the proposed QMA is to provide an end-to-end QoS solution robust regardless the wireless technology that is being used to access the RN. Such a high level of support and transparency implies that strong and reliable integration methods have to be developed. The main design guidelines and merits of this architecture are outlined as follows:

- the architecture must be modular, so that it is able to *separate* each specific communication technology, facilitating in this way the future integration of new wireless technologies into the RN;
- the designed architecture must be enough flexible to allow the interworking of several access wireless technologies;
- the proposed QoS scheme must exploit the well-known coarse-grained DiffServ strategy because it requires no signalling overhead, which is especially critical on

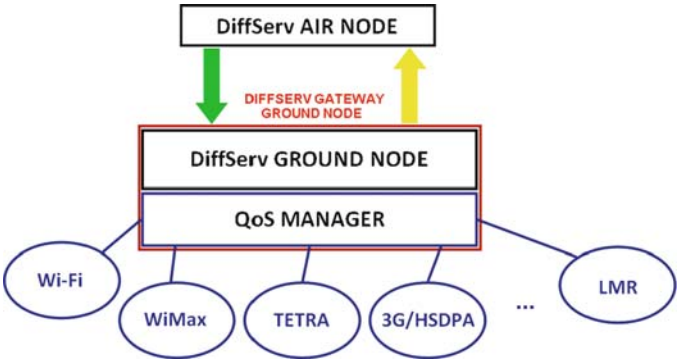


Fig. 2 QoS management architecture

the air-ground link and no signalling delay for path establishment, which makes it more efficient for short-lived flows. Morever, it is scalable because minimal state information is required at boundary DiffServ nodes (Air and Ground). Finally, it optimizes the QoS management via an adequate mapping between DiffServ classes and the specific classes of service supported by each wireless technology used in the RN; in this way, the QoS capabilities as well as the characteristics of every wireless technology are exploited at best;

- the QoS scheme also supports the QoS in existing and common Public Safety LMR systems by means of a LMR Gateway specifically designed.

Figure 2 shows the proposed architecture.

3.1.1 DiffServ Architecture in a Relief Network

The analysis of all application requirements has turned to be fundamental to design a fair mapping among envisaged relief services and DiffServ Classes. DiffServ nodes map the packet's DSCP to a Per-Hop Behavior (PHB) [5], a forwarding procedure that a node performs on a packet. Both DiffServ Air and Ground Node should support the DSCP-to-PHB mapping. The PHB states how to treat the traffic belonging to an aggregate of flows at a node. In the chapter, two commonly used PHB, Expedited Forwarding (EF) and Assured Forwarding (AF) are focused. The EF aims to provide a service characterized by low delay, low jitter, low loss and assured bandwidth. In our QoS scheme, this PHB has been chosen for conversational voice service in order to provide the highest level of aggregate quality of service that is crucial in envisaged disaster recovery scenario. Thus, the arrival rate of EF packets must not exceed the service rate at the node interface, so as to satisfy the features of EF PHB. The AF defines four independent forwarding classes for packet delivery. Within each AF class, there are three kinds of drop precedence with each packet to determine the importance of the packet. For example, AF12 means high priority and middle drop precedence. AF11 and AF12 have the same priority. If the

queue is full, packets marked with AF12 will be dropped first than AF11. In case of congestion, packets with high drop precedence are more likely to be discarded. In our QoS scheme, AF12 PHB has been chosen for the real-time streaming video application, a very interesting relief service that allow to know immediately the state of the sites damaged by disaster. AF11 PHB has been selected for telnet/SSH and remote data control service to adjust fundamental parameters of the applications from the MGS. Note that in the envisaged disaster recovery scenario, packets generated by such services have higher priority than the streaming video ones, since the control of the RN components on board is crucial. For the instant messaging and small data transfer/retrieval service, AF21 and AF22 PHB have been chosen, respectively. Finally, the packets of bulk data transfer/retrieval service can be forwarded in best effort mode since such services are not crucial operation for the first help and the first actions carried out after a catastrophic event. Table 1 summarizes the mapping between envisaged relief services and DSCPs, used in our QoS scheme.

According to DiffServ architecture, Fig. 3 shows the packet treatment procedure in the DiffServ Nodes. When the packets enter into interface input, a classifier first differentiates the types of traffic. The generic flow of traffic is identified by destination IP address and port number of the incoming packets. In our system, the EF

Table 1 Mapping between envisaged relief services and DSCPs

	DSCP	
Relief service	Name	Value
Conversational voice	EF	101110
Telnet/SSH, remote control	AF11	001010
Real-time video	AF12	001100
Instant messaging	AF21	010010
Small data transfer/retrieval	AF22	010100
Bulk data transfer/retrieval	AF23	010110

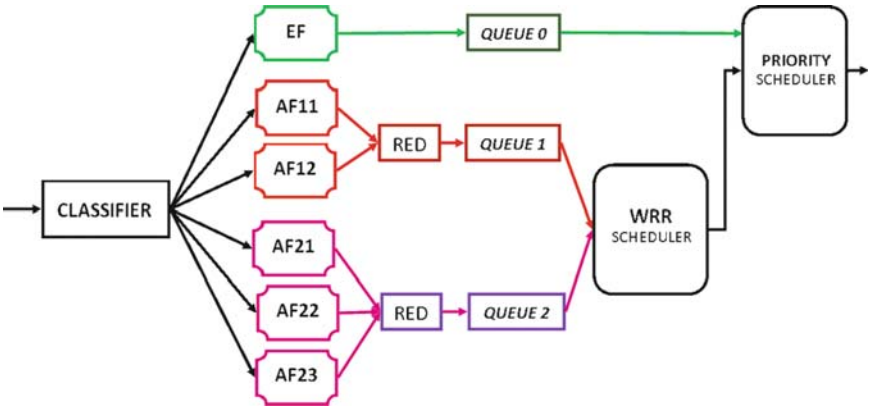


Fig. 3 Packet treatment in DiffServ nodes

packets are sent to a queue, Queue 0, with strict priority to ensure their deliveries. The AF1-class packets are fed into a common queue, Queue 1, according to a strategy named RED (Random Early Detection) [7], which discards the packets with an increasing probability as the FIFO queue buffer fill up. This queuing strategy has been adopted because it allows a fair treatment of aggregates of TCP flows [7]. In particular, when the aggregate is constituted by several TCP flows, the RED algorithm is able to equally distribute the bandwidth among the single flows, even in presence of different rates of sources. This result is achieved by limiting the most aggressive flows by dropping their packets with a higher probability. This drop activates the flow control mechanisms of TCP giving rise to a source rate decreasing. RED strategy has been implemented also for AF2-class packets, which are fed into a common queue, Queue 2. For each queue filled with RED algorithm, the occupancy is evaluated by an exponential moving average.

The computed occupancy value, say *Avr*, is compared with two thresholds, say *MinThres* and *MaxThres*. When a new packet arrives and $Avr < MinThres$ the packet is accepted. If $MinThres < Avr < MaxThres$ the packet is discarded with a probability p_a, function of avg. If $Avr > MaxThres$ the packet is discarded.

The strict priority scheduler accepts the inputs from Queue 0 and Weighted Round Robin scheduler (WRR). The EF traffic has the highest priority in the priority scheduler. In here, a kind of WRR scheduler called Hierarchical Token Bucket (HTB) is utilized. HTBs help in controlling the use of the outbound bandwidth on a given link. Since the maximum achievable throughput of the 802.11 g air-ground wireless link is approximately 27 Mbps [8], HTB queuing discipline for link sharing has been implemented to assure that the AF1 traffic aggregate occupies the max bandwidth of 2 Mbps and the AF2 traffic aggregate occupies the max bandwidth of 24 Mbps.

3.1.2 QoS Manager

The relief communications in a disaster recovery scenario occur in a highly heterogeneous environment where rescuers, employing different access technologies, can effectively communicate with each other. In this perspective, the data flows have to be adapted to the change of surrounding conditions. The ultimate goal is to render the RN seamless, namely, the proposed solution must assure the full interoperability of all the available technologies and guarantee, at the same time, that each rescuer perceives a given quality of service which depends on the access technology that he utilizes. More specifically, we propose to resort to mapping strategies among DiffServ and the service classes supported by the different systems involved in the integrated RN to benefit of specific QoS management of each wireless access technology by preserving their native QoS capabilities. Figure 4 depicts the functions that the QoS manager has to provide.

Specifically, the incoming aggregate flow from DiffServ ground node is processed by the *Class Selector,* which extracts the EF packets related to voice service, AF21 packets related to instant messaging service and the packets related to all

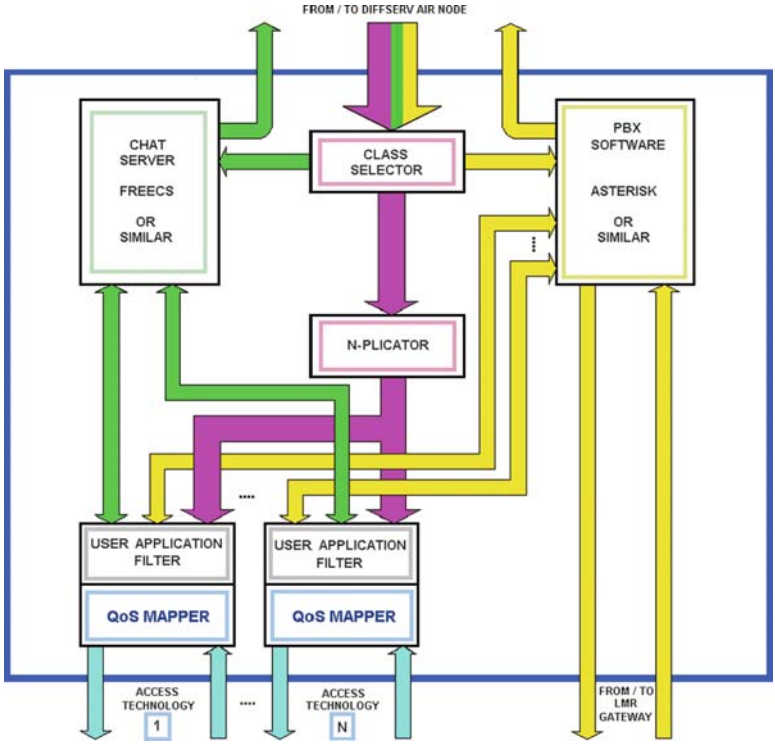


Fig. 4 QoS manager

remaining relief services. The first ones are sent to a VoIP server, Asterisk-based or similar, which allows the intercommunication of different rescuers, including on board operator. The second ones are sent to a chat server, like FreeCS, which allows every rescuer to chat with each other, regardless its location. The remaining packets, related to all others relief services, are sent to input of the *N-plicator Module* (NM) that provides N copies of them at output. Then each copy is sent to *QoS Management Module* (QMM) input implemented for each of the N wireless technology used in integrated RN. QMM consists of two modules: *User Application Filter* (UAF) and *QoS Mapper* (QM). The former selects the applications that are really utilized by each specific rescue team, consistent with the used access technology. The latter deals with the mapping between DiffServ and service classes supported by the specific wireless technology. In particular, for each wireless technology used in the RN, it has been identified the QoS class that best fits the DiffServ EF class. In the same way, we proceeded to define the mapping of DiffServ AF classes. Table 2 summarizes the results of such an analysis: for any DiffServ class has been reported the corresponding QoS class of each specific technology used in the RN. Note that, unlike [3] where the addition of a new access technology to the existing n ones requires the software implementation of other n new mapping tables, in our QMA proposal

Table 2 Mappings table

DiffServ	3G/HSDPA	WiFi (802.11e)	WiMAX	TETRA
EF (Expedit Forwarding class)	Conversational class (real-time conversational)	Highest priority Traffic Class (TC7)	UGS (Unsolicited Grant Services)	Teleservices and circuit switched data transfer services
AF1 (Assured Forwarding AF11, AF12)	Streaming class (streaming real-time)	Middle-high priority Traffic Classes (TC6, TC5)	rtPS (real-time Polling Service)	Connection oriented, packet switched real-time services
AF2 (Assured Forwarding AF21, AF22, AF23)	Interactive class (Interactive best effort)	Middle-low priority Traffic Classes (TC4, TC3, TC2)	nrtPS (non-real-time Polling Service)	Connection oriented, packet switched non real-time data services
BE (Best Effort class)	Background (background best effort)	Lowest priority Traffic class (TC1)	BE (Best Effort)	Connectionless packet switched data services

the inclusion of a new access technology only requires to introduce a further column in the aforementioned mapping table, namely, only one mapping module has to be implemented among DiffServ and service classes, which are supported by the added access technology.

3.1.3 QoS Software Implementation on Linux DiffServ Nodes

The QoS management strategy has been implemented on Linux platform for both DAN e DGGN. QoS modules are already present in Linux kernels of version 2.4 and later ones. It has been needed to add the modules for DiffServ and implement new ones to support mapping with other wireless communication technologies involved in disaster recovery scenario. Finally, Linux kernel has been recompiled on DiffServ Nodes. The implementation on a DiffServ Node provides a full set of *Traffic Conditioning Modules* (TCM) that include a marker, a classifier, a scheduler, service handlers for EF and AF and several queuing disciplines such as token bucket filters, FIFO and RED queues. All traffic conditioners have been implemented as kernel modules that can be activated by the *tc* command, which is part of the *iproute* package [9]. The TCM, used in our implementation, are outlined as follows. The *Service Handler* is the marking module of the implementation. It compares all incoming packets to the flows held in its table and writes the according DSCP into the IP header. Since this module has no metering functionality, the dropping probabilities of AF packets are set by the *Precedence Handler* module

(PHM). The *Dsclsfr* module is a combination of a Behavior Aggregate (BA) classifier and a scheduler. The classification procedure is executed when enqueuing a packet and forward the packets according to their DSCPs to one of seven traffic conditioners. Those conditioners are intended to handle the two AF classes and EF traffic considered. The scheduling performed by the de-queue function is a combination of priority scheduling used only by EF and WRR fair queuing implemented for AFs. The weights of the algorithm are configurable and can be specified via the command line. The PHM is a *color-aware two-rate three color marker* [10]. The AF-PHB defines four independent service classes, each operating at three levels of dropping probability. In our scheme, we considered only two AF classes. Traffic below the negotiated bandwidth limit has the lowest probability of being dropped ("is marked green"). A packet is marked "yellow" (to a higher dropping probability), if it does not exceed a certain exceed bandwidth. All other traffic is marked "red." The PHM specifies the color-part of the AF-DSCP, while preserving the color of already marked incoming packets. As shown in the QMA, to assure the mapping provisioning between DiffServ and different wireless communications technologies exploited in RN, some novel module and script have been added in the DGGN based on Linux PC. In particular, for each access technology, a specific QMM has been implemented. The basic structure of each module is common to all; it is composed of two parts, which allow both the achievement of QoS requirements (expected by each RN user) and the QoS mapping.

4 Performance Results

The effectiveness of the proposed QoS management solution has been investigated through a testing campaign. More specifically, it has been set up as a testbed based on the Wideband Radio Channel Emulator made by Elektrobit, named PropsimC2. Figure 5 depicts the testbed architecture.

Radio Frequency (RF) input of the PropsimC2 links to AAP RF output by a low-loss RF cable. AAP links to DAN by the first FastEthernet (FE) interface. The second FE interface of DAN links to server that runs VLC, FTP applications. RF output of the PropsimC2 links to GAP RF input by a low-loss RF cable. GAP links to DGGN by the FE interface. Finally, a laptop links to DGGN by 802.11 g interface configured in ad-hoc mode. The testing sessions have been performed by sending video TCP streaming, audio UDP streaming, and FTP bulk data on different ports in order to verify the different treatment operated on different classes of traffic. Through the HTB *qdisc* implemented in the our COTS Linux boxes (DAN

Fig. 5 Testbed architecture

and DGGN, respectively), two classes are created and the maximum rate, which each class can consume, has been set to assure that the AF1 traffic aggregate occupies the max bandwidth of 2 Mbps and the AF2 traffic aggregate occupies the max bandwidth of 24 Mbps. For the EF traffic, instead, has been set the highest priority in the priority scheduler in order to preserve the conversational voice flow in each condition. Tests have shown that the differentiated treatment of traffic works correctly; in particular, the conversational voice flow is guaranteed in each condition. The bit-rate of FTP bulk data download decreases when the channel conditions become poorer. In these conditions, FTP packets (belonging to AF23 class) are dropped, while the video steaming continues to be of good quality. If the channel conditions get worse further, it has been verified that the perceived video quality gets worse (frames with some blocks appear) while the perceived audio quality is still good. Finally, the carried out tests confirm that the QoS mapping implemented among DiffServ and 802.11 service classes works correctly.

Performance analysis are currently under study and, before the FALCO project deadline [11] which is at the end of November 2009, results in terms of delay, jitter, packet loss and throughput will be available.

5 Conclusions

This chapter deals with the inter-vehicle communication QoS management in a very hostile scenario, which is a disaster recovery. To assure that each rescuer can perceive a given satisfactory quality of service QoS regardless the used wireless technology for the access to the network a modular architecture that resorts to a DiffServ scheme is proposed. The proposed QMA not only provides seamless QoS support over different wireless technologies for the access network but exhibits a scalability property against the entry of any new access technology since the new entry can be managed just adding a new specific QMM without requiring the upgrading of the ones already present in the system. A testbed has been designed and performance evaluations are currently on the way and will be available for the project deadline.

References

1. MESA Project http://www.projectmesa.org
2. Boukalov A (2004) Cross standard system for future public safety and emergency communications. Vehicular technology conference
3. Durantini A, Petracca M, Vatalaro F, Civardi A, Ananasso F (2008) Integration of broadband wireless technologies and PMR systems for professional communications. Fourth international conference on networking and services
4. Iapichino G, Bonnet C, del Rio Herrero O, Baudoin C, Buret I (2008) A mobile ad-hoc satellite and wireless mesh networking approach for public safety communications. 10th international workshop on signal processing for space communications

5. Nichols K et al (1998) Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers, RFC 2474. http://www.faqs.org/rfcs/rfc2474.html
6. Linux Advanced Routing & Traffic Control. http://lartc.org/
7. Floyd S, Jacobson V. (1993) Random early detection gateways for congestion avoidance. IEEE/ACM Trans Netw, Agosto 1(4):397–413
8. White paper (2003) Maximizing your 802.11g investment. Proxim
9. Kuznetsov A iproute2 release 990824. ftp://ftp.sunet.se/pub/network/ip-routing/iproute2-2.2.4-now-ss990824.tar.gz
10. Heinanen J, Guerin R. (1999) A two rate three color marker. RFC 2698, September 1999
11. FALCO project. http://www.falco.cnit.it

Part IV
RFID and Sensor Networks Technologies

Beyond the ID in RFID

Christian Floerkemeier, Rahul Bhattacharyya, and Sanjay Sarma

1 Introduction

Wireless sensors are increasingly deployed in a number of novel application domains. One such emerging application domain is the perishable and pharmaceutical supply chain which transports large amount of sensitive goods. The net worth of temperature sensitive goods required to be transported is estimated to exceed $41 billion in the pharmaceutical supply chain [1]. Similarly, in the perishable goods supply chain, wastage due to overheating and excessive shock is estimated to be $35 billion and in some supply chains it can be as high as 33% of the transported freight [2]. Wireless sensors that monitor temperature, humidity, shock or vibration due to rough handling, and concentration of organic gases, such as ethylene, can thus play an important role in maintaining a good's state and reduce wastage. Other application domains for wireless sensing include structural health monitoring, where strain, displacement, and acceleration measurements can be used as indicators of a structure's health, and industrial monitoring where the health of machinery can be monitored.

In this chapter, we analyze the role that passive RFID-based wireless sensors can play in such monitoring applications. We begin by analyzing the requirements of an ideal wireless sensor device and discuss to what extent battery-powered wireless sensors meet these requirements. We then analyze the strengths and weaknesses of passive RFID-based sensors. Our analysis focuses, in particular, on an emerging concept of RFID tag antenna-based sensing that has the potential to result in ultra low cost RFID tags.

C. Floerkemeier (✉), R. Bhattacharyya, and S. Sarma
Auto ID Labs, Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: floerkem@mit.edu

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_21,

2 Requirements of Wireless Sensing

The requirements from wireless sensing units vary according to the specifics of the application being considered. Nevertheless, there is a superset of parameters that play an important role in many applications.

- *Spatial sampling frequency*: The ideal spatial sampling frequency is highly application dependent. In cold (supply) chain monitoring, one would ideally temperature monitor every logistical unit individually. However, due to cost constraints, temperature monitoring is often restricted to selected locations such as trucks and warehouses. The pallet of frozen goods accidentally left outside can thus not be detected.
- *Temporal sampling frequency*: Ideally, a particular physical quantity can be measured continuously. Due to limited power sources and on-board storage, this is often not feasible and the sampling frequency needs to be reduced. As a result, anomalies might be detected with a delay or missed all together.
- *Reporting frequency*: Captured data and anomalies detected should ideally be reported with a minimum delay. However, due to limited communication range and resulting lack of coverage, it might not always be possible to report anomalies in real-time. In many cases, the real-time reporting of anomalies is also not mandatory. A particular event might only be reported when the sensor node moves into the range of a interrogator or when an interrogator approaches the wireless sensor to read out the data.
- *Service lifetime*: The ideal lifetime of a sensor is very much application dependent. Building infrastructure monitoring requires wireless sensors that monitor the health of the structure over its entire lifetime – typical of the order of tens of years. Other applications only require sensors with service lifetimes ranging from months to several years.
- *Communication range*: Large communication range means that the number of interrogators or base stations that need to be deployed can be reduced. It also allows for real-time reporting because the wireless sensor node is more likely to be in the range of an base station, interrogator, or other wireless sensor node that can relay the information.
- *Overall cost*: Cost is an important factor to consider in wireless sensor deployments. This includes not only the cost of the wireless sensor nodes themselves, but also any supporting infrastructure such as base stations or interrogators for communication. The overall cost of a sensor deployment is influenced by a number of other factors. The spatial sampling frequency required and communication range influence, for example, the number of base station and wireless sensor devices needed.

3 Strengths and Weaknesses of Active, Battery Powered Wireless Sensors

We have seen from the previous section that there is a need for sensor development that addresses not only cost and service lifetime but also other aspects such as spatial and temporal sampling frequency. To avoid the additional complications associated with cumbersome lead wiring and to promote easy retrieval of data from the deployed sensors, wireless sensing systems have gained popularity and are currently being applied from fields as diverse as ecology [3] to aircraft health monitoring [4]. There are numerous architectures proposed for wired sensors – Motes [5] and BT Nodes [6] serve as good examples.

Battery powered wireless sensors can have significant communication range and due to the on-board power supply they can measure data almost continuously or at short time intervals. Due to the relatively large communication range, it is also easier to provide communication coverage for an application resulting in real-time reporting of the captured sensor data. There is no need to rely on the mobility of the wireless sensors or the mobile interrogators or base stations to gather the sensor data captured.

Battery powered wireless sensors are, however, also prone to some shortcomings. These sensors are typically associated with costs exceeding USD 20 due to the discrete components that make up an active sensor tag and this price tag can restrict their application and the spatial sampling frequency. One can envision using these sensors to track reusable assets such as storage containers; however, these sensors are not useful for the ubiquitous monitoring of each logistical unit passing through the supply chain. For instance, it would not make economic sense to deploy a battery powered temperature sensor costing USD 20 on cases of orange juice cartons passing through the supply chain. It is possible to deploy one such sensor in the freezer unit and infer that the goods are at the correct temperature based on the freezer temperature monitoring. However, it is assumed that the goods are inside the freezer and this cannot account for goods being accidently left outside the freezer unit. In supply chain monitoring applications, the deployment of battery powered wireless sensors is thus limited to specific supply chains such as those operated by the US Department of Defense to mobilize equipment and machinery or the monitoring of selected locations.

Battery powered wireless sensors can also measure continuously and convey real-time data to an acquisition system. However, the power requirements of this approach drive up the cost and reduce the battery life. In applications where the lifetime of the sensors needs to exceed tens of years, such as infrastructure monitoring, this becomes a limiting factor.

4 Passive Wireless Sensing

In an attempt to address the issue of low cost and long service life in wireless sensor node development, we examine the advantages of using passive sensing devices

that employ power scavenging mechanisms to conduct their functions. Passive devices tend to be cheaper, due to the lower complexity of the component electronics and have theoretically infinite lifetime as they are completely dependent on external power to function. Trade-offs include sacrificing communication range and reporting frequency; however, in the case of supply chain and infrastructure monitoring applications, this might in fact be acceptable in some scenarios where cost and lifecycle issues govern. Due to the low cost, passive wireless sensors can be placed on a much larger number of objects and/or locations leading to a more accurate spatial sampling. In applications such as supply chain monitoring, passive wireless sensors can also leverage the existing RFID reader infrastructure reducing the overall cost.

4.1 Passive RFID Front-end Combined with Sensor Electronics

Current research and commercial efforts focus on using a passive UHF RFID tag microchip as a front-end to integrated sensor electronics. Here, the tag microchip with integrated sensor electronics will be either custom designed or the system may be assembled from discrete components, as demonstrated by Smith et al. [7] in their WISP platform (cf. Fig. 1), where the sensor tag consists of an analog front-end, microprocessor, and a range of sensors. Being passive, this approach does address the sensor lifetime issue and it is also compatible with the UHF RFID infrastructure already deployed. It can also be augmented with a battery that allows for data collection when the sensor tag is not in range of an UHF RFID reader.

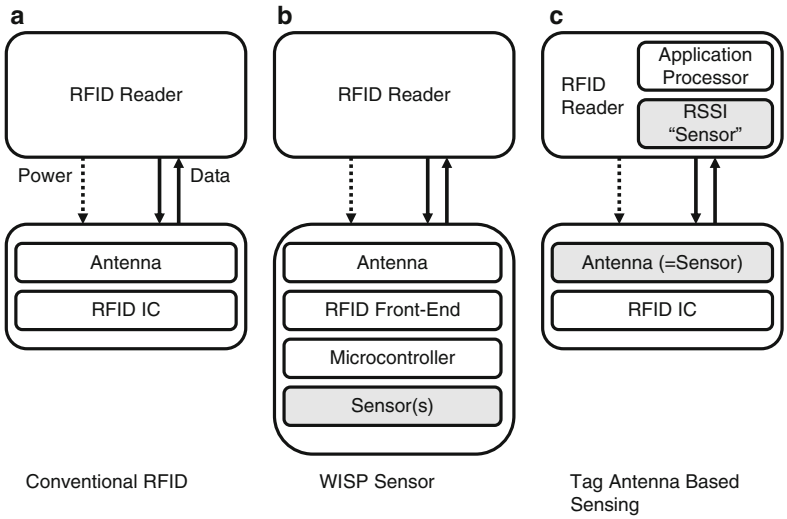


Fig. 1 Different approaches to passive RFID sensing: (**a**) conventional RFID tag, (**b**) WISP sensor, and (**c**) tag antenna based sensing approach

4.2 RFID Tag Antenna Based Sensing

We now present a methodology to utilize the passive UHF RFID tag antenna itself as a sensor (cf. Fig. 1). As we will demonstrate, research has shown that sensors can be constructed by calibrating a change in some physical phenomenon to changes in tag power characteristics. RFID tag antennas are often designed such that in the absence of metals or fluids the tag antenna impedance is a perfect conjugate match to the tag IC to maximize power transfer between antenna and microchip. Presence of metal or fluids in close proximity to the RFID tag antenna changes the tag antenna properties and this manifests in changes in the tag power characteristics and read range [10]. Thus, a sensor can be constructed by calibrating a change in some physical parameter – displacement, temperature, strain, moisture, etc. to a change in RFID tag antenna properties brought about by say a change in proximity of the tag to a metal plate or increase of fluid concentration in the environment of the tag antenna. The change can be detected by monitoring the received signal strength (RSSI) at the RFID reader (cf. Fig. 1).

We now present two case examples illustrating the tag antenna based sensing concept.

- *A tag-antenna displacement sensor*: Bhattacharyya et al. [8] describe an RFID tag antenna based displacement sensor. A metal plate is fixed to the bottom of a simply supported beam at a certain distance from an RFID tag (cf. Fig. 2). As the midpoint of the beam displaces under loading, the metal plate comes closer to the RFID tag modifying the tag antenna impedance and changing the tag power properties. Figure 3 demonstrates how tag backscatter power can be unequivocally be related to beam midpoint displacement. As we can observe from the graph, this sensor has a dynamic range of about 2.5 cm and an accuracy of about 2 mm.

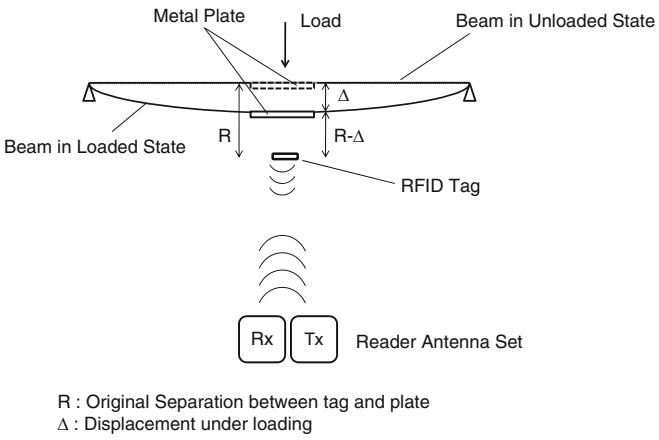


Fig. 2 Beam midspan displacement measurement [8]

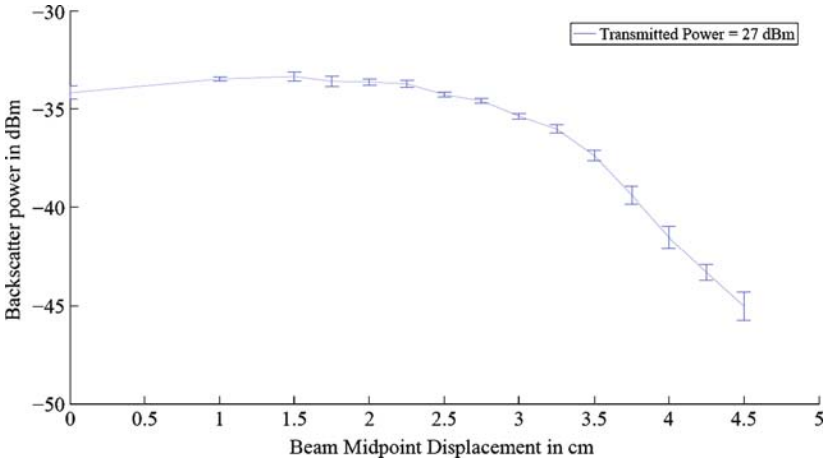


Fig. 3 Backscatter power vs. midspan displacement [8]

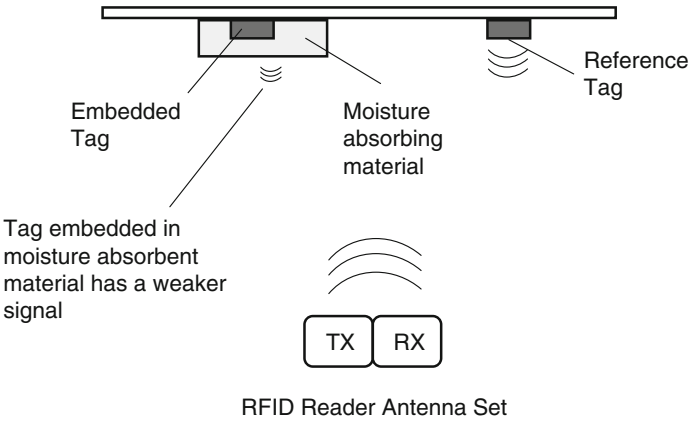


Fig. 4 RFID moisture sensor [9]

- *A tag-antenna moisture sensor*: Siden et al. [9] embed an RFID tag in layers of absorbent material such as blotting paper (cf. Fig. 4). The blotting paper absorbs the moisture from the ambient surroundings. As the amount of moisture absorbed increases, the severity of the detuning increases changing the tag power properties.

A sensor constructed using this paradigm utilizes a standard tag IC priced at a few cents and thus no additional costs are incurred for custom silicon fabrication. The tag antenna that is utilized may be a standard off the shelf variety priced at a few cents or else a custom-designed antenna. Custom-designed antennas are slightly more expensive; however, the cost of the entire sensor tag cannot be expected to be less than $1. For instance, a low-cost displacement sensor could be deployed on

every member of a truss bridge deck, thus ensuring that no anomalies are missed. Similarly, a moisture sensor costing less than a dollar could be attached to every individual carton of electronics components passing through the supply chain to make sure they were not subject to any moisture along the way. Passive RFID relies completely on the reader transmitted power for tag operations and in this sense has a theoretical infinite lifetime. This directly addresses the concern of sensor life in infrastructure monitoring. Finally, the tag-reader and reader communication protocols conform to the Gen 2 Protocol [11], which provides the additional benefit of interoperability.

As with every other kind of sensing approaches, tag antenna sensing has its limitations as well. As this is a passive technology, the read range for these types of sensors is restricted to a few meters. Presence of static metallic objects in the vicinity of the sensor can be accounted for during the calibration stage; however, in order to account for dynamic objects in the environment, it is necessary to consider measurements with respect to a reference tag whose antenna impedance is unaffected by the varying physical parameter much like the reference tag in Fig. 4, which is not wrapped in absorbent material. Relative measurements with respect to a reference tag also eliminate the aspect of reader-tag separation, provided the reader is not moved out of tag read range.

While the concept of tag antenna-based sensing is illustrated with two examples, it is possible to imagine extending this paradigm for other sensing applications such as touch, ambient light levels, and accelerations. Changes in properties of the RFID tag antenna may be temporary or permanent. Temporary changes may be acceptable for applications where instantaneous measurements are of importance. However, in many applications, it is important for the sensor to maintain state and convey important information such as the violation of a safety threshold level. For instance, the RFID displacement sensor measures the state of displacement of the member at the instant of measurement and changes in tag antenna properties are thus temporary. However, the RFID tag antenna may be designed so as to suffer a major permanent change when a threshold state is reached for the physical quantity being monitored ensuring that critical state information is recorded. For example, in the case of an RFID strain sensor, we can cause a permanent change in antenna impedance for a strain greater than 3,500 μ-strain, the failure strain level for concrete, brought about perhaps by the snapping of a part of the tag antenna element. Thus, in the future whenever the tag sensor is read, it is possible to note the occurance of the failure.

5 Conclusions and Outlook

Sensing and intelligent monitoring is no longer the monopoly of a few specific fields, but has started to permeate all areas of life. Each application domain brings its own unique set of monitoring challenges.

Battery powered wireless sensors are the most common commercial wireless sensors used today. However, limited battery life and higher costs limit their

deployment in some sensing applications. Passive RFID based wireless sensors and, in particular, the RFID tag antenna based sensing paradigm have several advantages including low cost, capacity for ubiquitous deployment, and theoretically infinite lifetime all of which are very desirable properties. This sensing paradigm cannot communicate real-time updates and relies on external communication infrastructure to query the sensor. Initial work in this area is demonstrated using a displacement and humidity sensor but tag antenna based sensing can be envisioned for the sensing of a variety of physical parameters such as temperature, strain, acceleration, touch, light, and radiation by appropriately modifying the tag antenna design and underlying dielectric medium. Furthermore, by inducing permanent physical changes, we can potentially convert these sensors into reliable alarm sensors. Thus, we believe that tag antenna based sensing introduces a sensing paradigm that trades off real-time updates and perhaps some accuracy for low cost, long lifespan, and standardized communication. By utilizing infrastructure monitoring and supply chain operations as case examples, we argue that these trade-offs might infact be acceptable and the gains from adopting this sensing approach may outweigh the shortcomings.

Alternative current research and commercial efforts focus on using a passive RFID front-end combined with sensor electronics. Here, the microchip may be custom designed, but this drives up the overall cost. Alternatively, the system may be assembled from discrete components, as demonstrated by Smith et al. [7] in their WISP platform; however, the assembly of discrete components this kind of sensor requires, drives up the cost of this sensor as well.

Finally, there has been some emphasis on the development of devices using printed and chipless RFID solutions. For instance, there has been research into the development of cost-effective printed RFID based temperature sensors [12]. This research has the potential to provide even lower cost alternatives to the tag antenna based sensing approach described in this chapter, however, the technology is still in its infancy and whether it will revolutionalize asset monitoring solutions remains to be seen.

References

1. Bishara RH (2006) Cold chain management – an essential component of the global pharmaceutical supply chain. Am Pharm Rev Jan/Feb pages 1–4
2. Delen D, Hardgrave BC, Sharda R (2008) The promise of RFID-based sensors in the perishables supply chain. RFID Research Center, University of Arkansas, Fayetteville, AR
3. Ho CK, Robinson A, Miller DR, Davis MJ (2005) Overview of sensors and needs for environmental monitoring. Sensors 5(1): 4–37
4. Bai H, Atiquzzaman M, Lilja D (2004) Wireless sensor network for aircraft health monitoring. In: Proceedings of the 1st international conference on broadband networks, San Jose, CA, 25–29 Oct 2004, pp 748–750
5. Kahn JM, Katz RH, Pister KSJ (1999) Next century challenges: mobile networking for "Smart Dust". In: MobiCom '99: proceedings of the 5th annual ACM/IEEE international conference on mobile computing and networking, ACM, New York, NY, pp 271–278

6. Beutel J, Kasten O, Ringwald M (2003) BTnodes – applications and architecture compared. In: Karl H, Wolisz A (eds.) TKN Technical Report TKN-03-012, 1. GI/ITG KuVS Fachgespräch Sensornetze, Technical University Berlin, Telecommunication Networks Group, Berlin, July 2003
7. Sample A, Yeager D, Powledge P, Mamishev A, Smith J (2008) Design of an RFID-based battery-free programmable sensing platform. IEEE Trans Instrum Meas 57(11):2608–2615
8. Bhattacharyya R, Florkemeier C, Sarma S (2009) Towards tag antenna based sensing – an RFID displacement sensor. In: International Conference on IEEE RFID, 27–28 Apr 2009, pp 95–102
9. Siden J, Zeng X, Unander T, Koptyug A, Nilsson HE (2007) Remote moisture sensing utilizing ordinary RFID tags. In: Sensors, 2007 IEEE, 28–31 Oct 2007, pp 308–311
10. Aroor SR, Deavours DD (2007) Evaluation of the state of passive UHF RFID: an experimental approach. IEEE Syst J 1(2):168–176
11. EPC Global Standards (2008) EPC global class 1 generation 2 UHF air interface protocol standard. http://www.epcglobalinc.org/standards/uhfc1g2/
12. The Fraunhofer Institute of Integrated Systems and Device Technology IISB. http://www.printedelectronicsworld.com/articles/. Printed Electronics World

Performance Characterization of Passive UHF RFID Tags

Leena Ukkonen and Lauri Sydänheimo

1 Introduction

Passive ultra-high frequency (UHF) radio frequency identification (RFID) systems have gained a large interest during the recent years. One of the key components of an RFID system is a tag, which consists of an antenna and an IC chip. The growing use of passive UHF RFID systems in various applications requires new tag designs whose performance should also be characterized reliably.

The operation abilities of passive UHF RFID systems depend mainly on two fundamental operational principles of passive UHF tags:

1. The capability of the tag for wireless energy collection from the reader, i.e., tag's energy harvesting. This depends on the impedance matching between the tag's antenna and the IC and the IC sensitivity.
2. The strength and clarity of desired backscattered signal from the tag, i.e., the radar cross-section (RCS) and differential radar cross-section (delta RCS) of the tag.

The energy harvesting information provides factors that define a minimum transmission power level for the reader unit to turn the tag on over a certain reading distance. All of this comes down to the impedance matching between the tag's antenna and the IC and the IC sensitivity, which is typically around −10 to −15 dBm. The goal of matching is to deliver as much power as possible from the tag antenna to the IC.

With proper tag antenna design, both energy harvesting and backscattering properties of the tag can be affected, and thereby system performance can be improved.

For some years ago, measuring the read range of the passive tags was the main performance characterization method [1, 2]. However, read range measurements are very difficult to standardize because the used reader equipment and the measurement environment have a large effect on the results. In addition, by concentrating only on the read range it is quite difficult to get a hold on the specific parameters

L. Ukkonen (✉) and L. Sydänheimo
Tampere University of Technology, Department of Electronics, Rauma Research Unit, Kalliokatu 2, 26100 Rauma, Finland
e-mail: leena.ukkonen@tut.fi

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_22,

of the tag design, i.e., what is the threshold power of the tag at a certain frequency and distance. Thereby, more sophisticated characterization methods have to be developed [3, 4]. The most realistic picture of the tag performance can be achieved by measuring the tag in dynamic mode, i.e., in operation with the IC [4].

This chapter discusses the performance characterization of passive UHF RFID tags. The key factors affecting the performance – impedance matching and RCS properties – are presented and discussed. The performance of tags is analyzed with two modeling and measurement examples.

2 Functioning of Passive UHF RFID Systems

Figure 1 presents the components of a passive UHF RFID system. Passive UHF RFID systems use electromagnetic waves in coupling and communication between reader unit and tag. The reader sends continuous wave (CW) signal to the tag to activate its microchip and then commands that are modulated to the signal. The tag responds with its identification code using backscattering of modulated electromagnetic wave. There is no internal source of energy in the tag's microchip, and thereby it gets all the energy needed for functioning from the electromagnetic wave emitted by the reader.

The communication between the reader and the tag is achieved by the tag switching its load impedance, which modulates the RCS of the tag. The RCS of a scattering target is the equivalent area of the target based on the target reradiating or scattering the incident power isotropically. The RCS of a target is not necessarily equivalent to its physical dimension. It can be described as a representation of how effectively a target can scatter the incident power. When the target is a loaded antenna, such as a tag antenna with the IC chip, the RCS can be altered by terminating the antenna with different load impedances. The modulating depth of the RCS affects the tag's read range: deeper RCS modulation results in longer read range [5, 6].

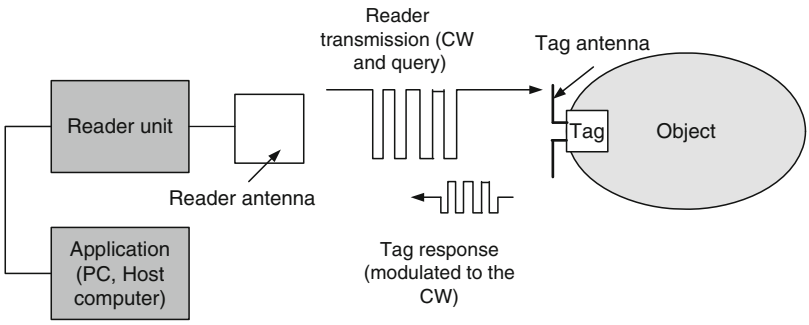


Fig. 1 The components of a passive UHF RFID system

2.1 Equivalent Tag Circuit

An RFID tag consists of an IC chip and an antenna. One of the important characteristics of the IC chip is its input impedance. A proper match between the tag antenna and the chip is important because it directly influences the RFID system's performance characteristics such as the read range of the tag [4–8].

Figure 2 presents the Thevenin equivalent circuit of a tag. The rms voltage V is induced when the incident wave from the reader reaches the tag. The complex chip impedance is given by $Z_L = R_L + jX_L$ and the tag antenna impedance is given by $Z_A = R_A + jX_A$. In the ideal complex conjugate matching $Z_A = Z_L^*$.

Matching of the tag's antenna and the IC can be characterized with power reflection coefficient (PRC) where s is the power wave reflection coefficient [8]:

$$|s|^2 = \left| \frac{Z_L - Z_A^*}{Z_L + Z_A} \right|^2 \tag{1}$$

The received power P_r by the reader can be expressed as [3, 4]

$$P_r = \frac{P_t G_t^2 \sigma \lambda^2}{(4\pi)^3 R^4} \tag{2}$$

where P_t is the transmitted power by the reader, G_t is the gain of the transmitter (reader) antenna, σ is the RCS of the tag, λ is the wavelength, and R is the distance between the tag and the reader antenna. Equation (2) shows that P_r is proportional to the RCS of the tag.

In the backscattering modulation process, the impedance of the IC chip is changed between two states, which can be, for example, matched and shorted state. By changing the input impedance, the RCS of the tag and the received power by the reader change. The difference between the radar cross-sections of the two modulation states is called differential (or delta) RCS ($\Delta\sigma$, ΔRCS) [5, 9]. It defines the quality of the amplitude shift keying (ASK) modulation of the tag: larger $\Delta\sigma$ allows the reader to detect the binary code in the backscattered signal more clearly.

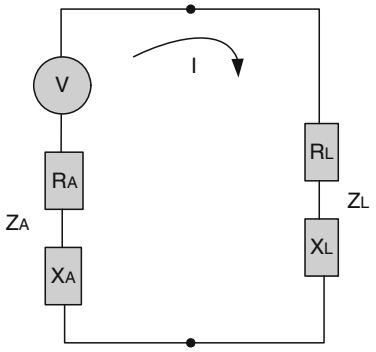


Fig. 2 The Thevenin equivalent circuit of a tag

Equation wise, differential RCS can be defined as [10]

$$\Delta\sigma = \frac{G_{\text{tag}}^2 \lambda^2}{16\pi} |s_1 - s_2|^2 \tag{3}$$

where G_{tag} is the gain of the tag's antenna and s_1 and s_2 are the power wave reflection coefficients for the two modulation states.

3 Threshold Power Level and Backscattered Signal Strength of a Tag

Very often, the limiting factor for tag read range is the delivery of power to the IC chip of the tag. The transmitted power by the reader is determined in the communication regulations, and thereby the power delivery to the IC chip at given distance is determined by the tag antenna design. The IC chip requires a certain amount of power to operate, and the tag antenna design and its matching structures determine how efficiently the power is delivered to the IC chip. In addition, the materials in the vicinity of the tag affect the properties of the antenna and the IC chip. The minimum transmission power required to activate the tag at a specified distance and specified frequency is called the tag's threshold power level.

However, the operating range and the reading reliability of the tag depend also on the backscattered signal strength, which is detected at the receiver of the reader. The minimum signal strength, which the reader can reliably detect and decode is usually presented as the sensitivity of the reader (a power level in dBm). Typically, the sensitivity of a commercial reader unit is around −70 to −90 dBm. A typical method for measuring the backscattered signal from the tag is using a signal source attached to a transmitting antenna to send a query command to the tag. The strength of the backscattered signal is then measured with a vector signal analyzer. The measurement can be carried out with different transmission frequencies and power levels to further tag characterization.

4 Overview of RFID Signal Measurements

Threshold power measurement is a rather straightforward operation. It includes finding the tag IC start power level at each frequency and getting a proper answer for ID query. Typically, the power level at RF port is known and the remaining options are measurement distance and antenna gain efficiency with cable loss.

Application environment can change the basic radiowave propagation properties. Typically, these are higher attenuation, multipath propagation, and changes in wavelength if the measurement is carried out through a dielectric medium. These can change the threshold power level of tag in one way or another and it must be considered with an application-specific tag design. Some threshold measurements

Table 1 Measurement results and the RCS of the tags

	10 mm	15 mm	20 mm
866 MHz			
Transmitted power	11.9	19.8	16.6
Received power	−43.3	−44.1	−42.6
RCS (dBsm)	−31.6	−40.3	−35.6
915 MHz			
Transmitted power	11.5	16.3	13.5
Received power	−43.7	−43.2	−42.1
RCS (dBsm)	−31.1	−35.4	−31.5
950 MHz			
Transmitted power	12.9	12.7	12.1
Received power	−42.4	−41.5	−40.9
RCS (dBsm)	−30.9	−29.8	−28.6
Resonance frequency	905 MHz	970 MHz	955 MHz
Transmitted power	10.2	10.7	11.3
Received power	−43.6	−42.3	−41.9
RCS (dBsm)	−29.8	−28.4	−28.7

can be carried out even with the commercial reader units with adjustable power level. However, in that case the results lack exact frequency information.

Backscattering properties of a tag have typically been measured in nondynamic mode, i.e., a tag design is measured with match, short and open circuit separately [3, 11]. However, even a good tag IC matching is not always optimum for all power levels and frequencies. Thus, backscattered, i.e., received, power level can vary quite a lot at different frequencies as shown in Table 1. The dynamic measurement delivers additional information over frequency range with sufficient operational power level. In addition to received power level, the other important factor is the level change between matched and nonmatched states. This element is related with delta RCS and quality of matching. Proper dynamic backscatter measurement requires vector signal analysis in operational mode.

5 Examples of Performance Characterization

In this section, two performance characterization examples are presented: an analysis of the effects of dipole antenna width [4] and an analysis of the impedance matching properties of a bow-tie tag antenna [12].

5.1 Dipoles of Various Widths

This characterization example deals with dipole-type tags of various widths. The length of the dipole was kept constant at 110 mm and the width was varied from 10 to 15 mm and 20 mm. The desired input impedance for each of the antennas was

$40 + j133\,\Omega$, the complex conjugate of the EPC Gen 2 RFID IC chip used with the dipole antennas in the measurements. The desired operating frequency for the dipole antenna was within the 900–960 MHz RFID band.

The effect of dipole width on the ΔRCS of the dipole tags was studied with finite element method (FEM) modeling. Figure 3 presents the modeling results. It can be seen that the ΔRCS increases with the dipole width, and thereby the dipole area.

Figure 4 presents the measured threshold power levels of the three tags. Changing the tag width affects the impedance matching of the antenna and the IC chip, and thereby the optimal functioning frequency shifts.

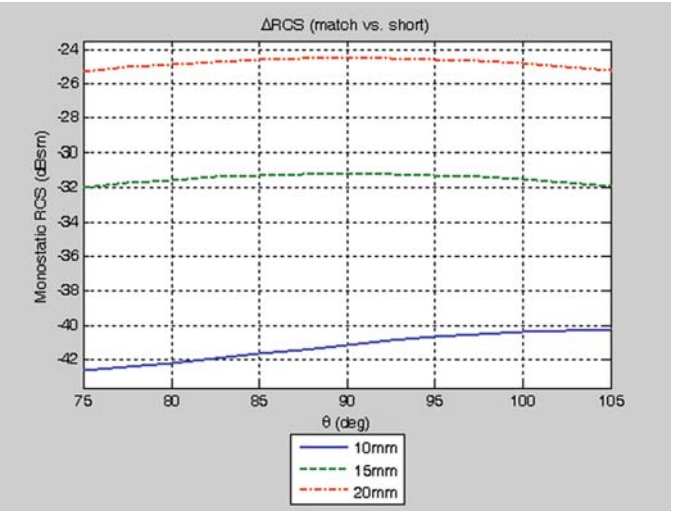


Fig. 3 Modeled ΔRCS of the dipole tags

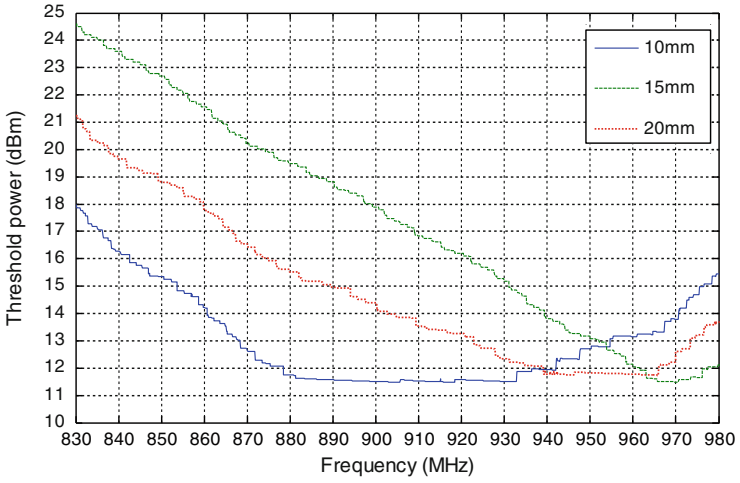


Fig. 4 Threshold power levels of the dipole tags

In addition, the strength of the backscattered signal from the tag was measured, and based on backscattered signal strength and threshold power level, the RCS of each tag was calculated at specified frequencies based on (2). Table 1 presents these results.

5.2 *Analysis of Impedance Matching Properties of a Bow-Tie Tag Antenna*

It is important to understand how good impedance matching is good enough. In this section, we present comparison results of four tags that have different level of matching properties. Four levels (best, fair, poor, and worst) are chosen to demonstrate either the change of the parameters of the tag in ideal environment or when the tag has been used on some other material, which has different electrical properties leading to a change in the tag performance.

The tag used in this chapter is a bow-tie type tag antenna with a T-match structure (Fig. 5). To obtain four different matched tag antennas, the dimensions of matching structure were adjusted during simulations to have the desired results. Simulations were carried out based on FEM. Length and width of the matching part was changed. In the case of complex-valued UHF RFID tag antenna and IC, we evaluate the conjugate matching by PRC in dB (PRC) [8].

According to the PRC, we chose four models from simulation results, which we named as best matching, fair matching, poor matching, and worst matching. The simulated PRC of four tags using Alien Higgs2 IC chip with impedance of $17 - j145\,\Omega$ is shown in Fig. 6. The best matching circuit and the antenna design for free space are shown in Fig. 5.

To analyze how the different matching levels affect the operation, we measured the minimum required transmitted power to activate the tag, i.e., threshold power of the tag (P_{th}) in dBm at measurement equipment transmitter port.

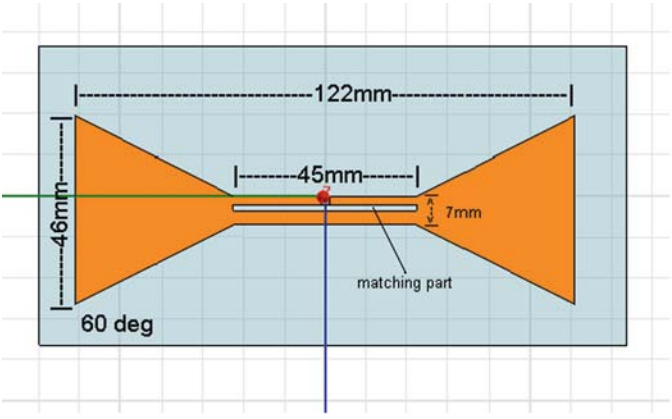


Fig. 5 Bow tie tag antenna geometry

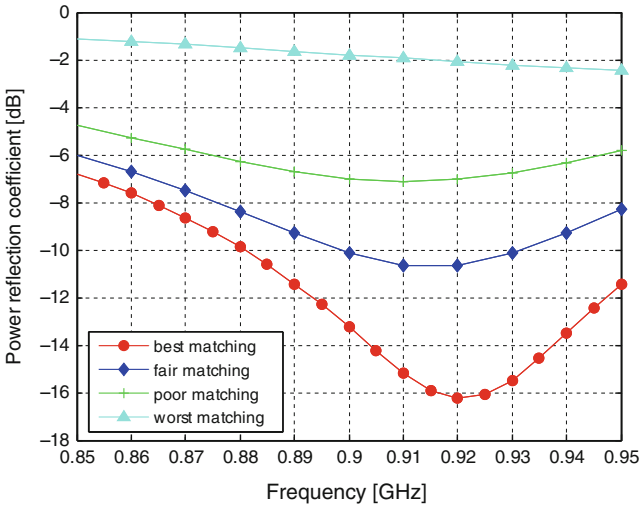


Fig. 6 Simulated PRC

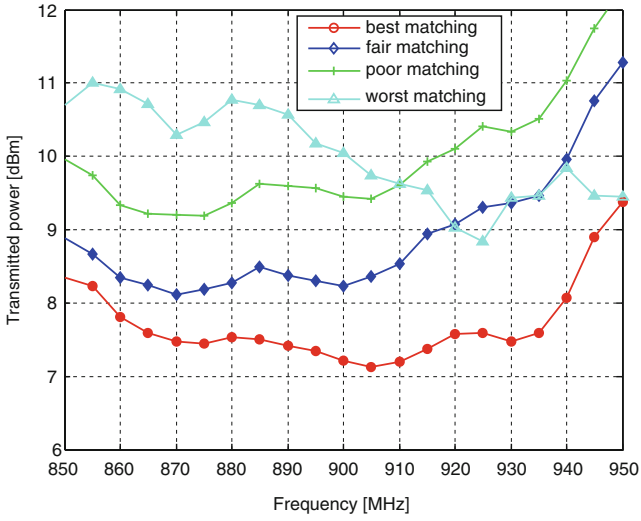


Fig. 7 Measured P_{th}

Measurements were carried out in an anechoic chamber where the bow-tie tag was placed at height of 1 m and at the distance of 1 m from the transmitter/receiver antennas. The position was kept invariable, only tag was changed. P_{th} was measured with Tagformance system [13] with linearly polarized transmitter and receiver antennas. Measurement results of bow-tie tag with different matching at 1 m are presented in Fig. 7. The PRC and P_{th} comparison at two frequencies is shown in Table 2.

Table 2 PRC and P_{th} at two frequencies

	Simulated PRC at 866 MHz (dB)	P_{th} (dBm) at 866 MHz	Simulated PRC at 915 MHz (dB)	P_{th} (dBm) at 915 MHz
Best matching	−8.2	7.6	−15.9	8.1
Fair matching	−7.2	9	−10.7	9.7
Poor matching	−5.6	10	−7.1	10.6
Worst matching	−1.3	11.9	−2.0	11.1

From the results we can see that at 915 MHz, the PRC of best matching can be as good as −16 dB. It means that only 2–3% of power will be reflected. This can be considered as good matching. Compared to good matching with poor matching structure, the PRC is half of that with best matching, and P_{th} is 2–3 dB higher. It means that with poor matching at the same distance 2–3 dB more transmitted power is required. The worst matching requires about 3–5 dB more transmitted power at the same reading distance. Power increase is feasible as long as reading distances are short enough to be within regulations.

6 Conclusions and Future Work

This chapter is concentrated on performance characterization of passive UHF RFID tags. We have presented and discussed threshold power and backscattered signal strength measurements and PRC calculation and analysis. Measuring the tags in operational (dynamic) mode is the most efficient way to characterize passive UHF RFID tag performance. In the future, we will concentrate on analyzing further the differential RCS of tags and the function of received modulation depth in the operational mode. In addition, contactless radiation pattern measurement methods for passive UHF RFID tags will be developed.

References

1. Choi W, Son HW, Shin C, Bae J, Choi G (2006) RFID tag antenna with a meandered dipole and inductively coupled coil. In: Proceedings of IEEE international symposium of antennas and propagation, pp 619–622
2. Sydänheimo L, Ukkonen L, Kivikoski M (2006) Effects of size and shape of metallic objects on performance of passive radio frequency identification. Int J Adv Manuf Technol 30: 897–905, Springer
3. Nikitin P, Rao KVS (2006) Theory and measurement of backscattering from RFID tags. IEEE Antennas Propag Mag 48(6):212–218
4. Sydänheimo L, Nummela J, Ukkonen L, McVay J, Hoorfar A, Kivikoski M (2008) Characterization of passive UHF RFID tag performance. IEEE Antennas Propag Mag 50(3):207–212
5. Yen C-C, Gutierrez AE, Veeramani D, van de Weide D (2007) Radar cross-section analysis of backscattering RFID tags. IEEE Antennas Wirel Propag Lett 6:279–281

6. Bolomey JC, Gardiol F (2008) Optimization of passive RFID tag antennas. In: Proceedings of IEEE international symposium on antennas and propagation, pp 1–4
7. Kwon H, Lee B (2005) Meander line RFID tag at UHF band evaluated with radar cross sections. Proceedings of the IEEE Asia-Pacific microwave conference
8. Kurokawa K (1965) Power waves and the scattering matrix. IEEE Trans Microw Theory Tech 13(2):194–202
9. Nikitin P, Rao KVS, Martinez RD (2007) Differential RCS of RFID tag. IET Electron Lett 43(8):431–432
10. Dobkin DM (2008) The RF in RFID: passive UHF RFID systems in practice. Elsevier, New York
11. Penttila K, Keskilammi M, Sydanheimo L, Kivikoski M (2006) Radar cross-section analysis for passive RFID systems. IEE Proc Microw Antennas Propag 153(1):103–109
12. Zhang J, Babar A, Ukkonen L, Sydänheimo L, Elsherbeni A, Yang F (2008) Performance of RFID bowtie tag antenna with different impedance matching. Proceedings of the Asia Pacific microwave conference, Hong Kong and Macau E1–06, 16–20 Dec 2008, 4p
13. http://www.voyantic.com/index.php?trg=browse&id=64

Chipless Tags, the Next RFID Frontier

S. Tedjini, E. Perret, V. Deepu, and M. Bernier

1 Introduction

Even if the concept of RadioFrequency IDentification (RFID) was introduced many decades ago [1], it is still very attractive and fertile in term of R&D and new applications [2]. Today, RFID is seen as a very enabling technology and it is under consideration for thousands of applications covering a large variety of domains among them: ID papers, security, access control, road toll, ticketing, pharmacy, logistic, manufacturing, gambling, etc.[3] Due to their relative low cost and large distance of communication, passive (i.e., batteryless) UHF tags are very promising. The crucial issues of cost, efficiency, reliability, security, and standards are under consideration by several groups worldwide [3].

Research on passive tags, particularly UHF tags, is still very active in order to ensure interoperability, low-cost requirement, and data security. The interoperability is needed since there are three frequency bands worldwide. Roughly, the operating frequency bands 865–869 MHz for Europe, 902–928 MHz, for Americas and USA, 952–954 MHz in Japan, China, and most of Asia [4]. The interoperability requires the development of efficient miniaturized antenna able to cover the three RFID UHF Bands. Privacy is highly dependant on the security of the data contained in the RFID Chip. Today, different microelectronic technologies are available for manufacturing RFID chips: CMOS, ASIC, and EEPROM are the best known examples [4, 5]. As the IFF [6] (Identification of Friend or Foe) was the first application of RFID, during World War II, the Allies used the cryptography on IFF transponders. In today's RFID, chip on-tag cryptography is generally desirable and can be implemented in many applications. However, on-tag cryptography is still prohibitive when it violates application requirements, such as power or cost constraints [7]. The third issue concerns the cost of the tag. In addition to the cost of the RFID IC and the antenna, the cost of the tag will depend on the connection manufacturing process between the antenna and the IC RFID Chip. Even if much progress has been made in antenna

S. Tedjini (✉), E. Perret, V. Deepu, and M. Bernier
Grenoble-inp/LCIS, ESISAR, F 26902 Valence, France
e-mail: smail.tedjini@lcis.grenoble-inp.fr

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_23,

cost reduction using low-cost substrate and ink printing [8] of at least a part of the antenna, the interconnection between antenna and RFID IC ports still demands better reliability [9, 10].

One alternative for better data security and low-cost objective is to consider Chipless solutions. Recently many designs of chipless RFIDs in the microwave region have been reported in literature. Microwave tags or RF tags have obvious advantages of longer range compared to optical barcodes and are easy to be fabricated by conventional lithographic techniques. This is a very interesting approach under development in many labs.

In this communication, we discuss different approaches and advances for UHF tag design. Section 2 will discuss the traditional tags based on RFID chip. Section 3 is dedicated to chipless. Both RF and THz chipless possibilities are presented and discussed. Some simulation results of newly proposed structures are reported.

2 Traditional RFID Tag

Nowadays, UHF Radio Frequency Identification (RFID) technology is generating much interest in industrial and academic institutions owing to its immense potential in tracking. This is why the design of UHF RFID tags with high performances and low cost is a very active field of research.

UHF tags with relatively low cost have been in place for the last 10 years. However, the deployment in high-volume of this technology is still held back by the relatively high cost of these tags. This is a part of the situation, but it is not the only reason. Indeed, if we consider the case of HF RFID, one can notice that such tags are relatively close to the UHF tags regarding manufacturing process, technology and in situ materials. These tags are also composed of a silicon chip and metal strips for communicating with the reader, and thus the costs of both of these tags (UH and UHF) are comparable.

But the field of application is quite different. The HF tags are used for short-range applications. However, if we consider the mass market, in spite of this real handicap (for maintenance or handling reason, it is generally better to be able to read all tags involved in a specific area remotely), HF tags are still the most involved in this sector. The reason for this is not economical (tags prices are comparable) but technical. Indeed, in real situation, the HF tags are more robust than the UHF tags. This has an extremely important impact for the final user: without preliminary test, a unique HF tag can be used with a very great number of objects. For example, in practice, the RFID system operation will not be disturbed if the object, on which the tag is placed, contains metals or not.

In the case of UHF, the environment (the object on which the tag is placed, as well as the close environment) in which the tags are used affects their characteristics considerably. In particular, when the tag is placed in an environment different from which it was specifically designed for, the performances of the system can be rapidly damaged, and thus the potential for this technology is limited. However,

we remind that this application would be viable for detection in about 99% cases. General observations could lead one to notice that tracking objects have neither the same electromagnetic properties, nor the same geometrical size. Moreover, considering the trend of new applications, it is expected that the tags will be integrated directly in the object itself. Indeed, it is interesting to follow the object throughout its different manufacture steps, but also throughout the supply chain.

Therefore, it is important to develop new RFID tag design tools to promote the deployment in mass-volume of the UHF RFID technology. The tool, on which we are working [11], is able to automatically design the tag according to user specifications. The inputs by the user are the geometrical size of the final tag and the properties of the UHF RFID chip frontend. Depending on the user requirements, on where the tag will be used (permittivity, metal, etc.), the application returns the topology of the tag, which can be used directly for the manufacturing.

Antennas design solutions rest on purely empirical design approaches, based on folded dipole antennas and current loop (for matching and near field considerations). We develop a new design approach where the user requirements take part in the design of the antenna. The antenna topology is not defined a priori as in the classical methods. Original topologies of antennas are generated automatically and selected according to the constraints. Some examples are given in Fig. 1. This approach does not simply consist in optimizing a topology already selected, but in designing the antenna from beginning to end. Our approach combines general purpose software (for example, MATLAB®) with software dedicated to EM simulations (for example, ANSOFT DESIGNER®). The main motivation for the integration of these two types of software is to benefit from data processing, the very broad panel of functions (for example functions of optimization), along with powerful and very general-purpose electromagnetic simulators. All the commands are controlled from MATLAB® what makes this approach very flexible. So, the designer can avoid manual repetitive tasks as well as the tedious parameterizations.

We use an optimization process based on the concept of Genetic Algorithms (GA) to satisfy the constraints set during the design process. The optimization consists of an iterative process, which first generates the antenna shape, then simulates it and finely evaluates its performance according to the imposed constraints. Thus, the antenna shape changes during iteration on an evolutionary principle. This is repeated until an antenna design, which satisfies the project specification (as good

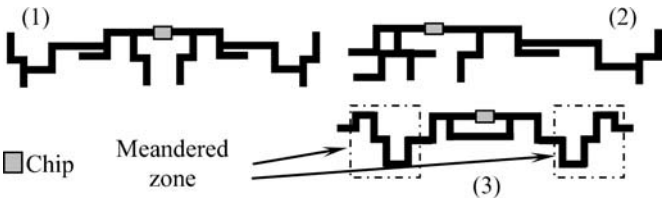


Fig. 1 Examples of automatic antenna designs: (1) symmetric (2) asymmetric (3) sectorization method (meandered zone in dash line)

as possible) is obtained. This very flexible approach makes it possible to take into account the different issues throughout the design process, and specially, the physical environment of the tag [11].

3 Chipless Solutions

Chipless RFID is an emerging area of RFID technology for ultra-low-cost RFID applications. However, it is currently confined to the unlicensed radio frequency bands. In this section, we present different approaches for chipless solutions, including some alternative to consider higher frequencies, namely the THz domain.

3.1 Chipless Methods

The most popular tags on the market (the most sold too) are the passive tags. This family of tags has achieved tremendous growth, although unit costs still remain high, hindering their development [1]. This is why tags without chip (chipless) have appeared and making it possible to reduce the cost in a large way to make them compatible with the Auto ID Center recommendations. In addition to the price of the chip, this approach allows to drive down chip-assembly cost. These elements make up to more than half of the price of the traditional tag [12].

Chipless tags, also named "RF barcode", are usually manufactured with low cost materials, generally electromagnetic reflective or absorptive materials. Chipless tags, compared to passive tags, generally have the following characteristics:

- low cost, less than 5 cents in volume,
- contactless, short ranges less than 1 m,
- better reliability: thermal and mechanical behaviours

However, these advantages should be balanced with the limited storage capacity (a few tens of bits) and the non-rewriteable characteristic (Read-Only Tags) of these devices. Another drawback is the cost of the reader, which could be higher compared to chip-based readers.

Chipless tags are composed of different families, based on various approaches. The most promising are based on:

1. The acousto-optics properties of materials, more precisely on surface acoustic wave (SAW) [13]. This approach already commercialized is by far the most mature chipless RFID technology.
2. Printed organic transistors. This prospective approach is mainly based on the same principle of passive RFID [14],
3. The electromagnetic properties of RF in passive microwave integrated circuits [15–21]. This approach is very promising but still in the developing stage.

The question is: how is it possible to encode information using ultra low-cost passive RF devices? The principle of the information encoding, which consists in encoding

the identification number of the tag, is based on the generation of a specific temporal or frequency footprint. This temporal footprint can be obtained by the generation of echoes due to the reflection of the incidental impulse. In the frequency domain, one can characterize the spectrum of the tag backscatter signal. There are several ways to encode binary data. Two easy-to-implement approaches for information encoding consist of:

- locating the presence or absence of a specific signal which is known to occur at a given time or frequency (such as On-Off Keying modulation (OOK)).
- measuring the gap (in time or in frequency) between two characteristic signals (such as pulse position modulation (PPM)).

The signals are generally electromagnetic waves, one can use the amplitude or the phase to encode the information. In the temporal domain, the design of devices rests on the concept of reflecting signals due to discontinuities. These discontinuities can be typically due to a rough variation of the geometries of the transition line (microwave approach) or of the medium (optic approach). A simple technique is to place a number of discontinuities at different distances in order to obtain a specific signal where the information is encoded by the temporal gap between the impulses. These discontinuities can be easily realized with localized [15] or distributed [16] capacitances placed on a transmission line.

In the frequency domain, it is possible to encode the information by taking into account the amplitude variations in the frequency of the backscattering wave. Such work has been done by placing resonating elements near a transmission line [17,18] or by exploiting the resonance frequency of a network of dipoles [19, 20]. Some studies have shown that it is particularly interesting to encode information by the phase wave variations [21, 22].

The great advantage of these devices is that they can be manufactured on the top of low-cost dielectric substrates. However, it has low data capacity and dimensions are usually quite larger (about tenth of cm^2).

The introduction of 2D structures could tackle these limitations. We think also that these different principles presented earlier can be transferred to higher frequencies in order to offer miniaturized tag solutions with higher capacities. Some years ago, devices based on holographic principles [23] have been investigated. Such a solution requires imaging technique in order to read the information.

3.2 RF Chipless

In the amplitude approach, an array of microstrip dipoles behaving as band pass or band stop filters tuned to certain predetermined frequencies is used to represent data as given in [19]. Another method is to use capacitive tuned split microstrip resonators. Here, the capacitance of each element of the dipole array is varied by changing the dimension of the split at the center, thereby obtaining the desired tuning [20]. The aforementioned two techniques are based on the bistatic S21

measurements. These methods posse certain difficulties owing to multipath effects, mutual coupling, requirement of large bandwidth and few number of data bits.

Preradovic et al. has presented a fully passive chipless RFID system using both the amplitude and phase of the spectral signature [17]. This system uses a pair of orthogonally polarized dual band antennas with wide bandwidth for the transmission and reception of signals. A multi resonator circuit is used to encode the multi frequency encoder signal from the antenna. By varying the dimensions of each of the spiral resonator, the corresponding frequency can be varied. But in practical cases, it has to be seen whether the signal to noise ratio is well maintained. This method requires a reference for performing the amplitude and phase of the signal. Mukherjee et al. [22] has proposed a method based on the phase – frequency signature by the reactive termination of the tag antenna. A microstrip based L-C ladder is used to encode the bits in phase – frequency profile. Various fully printable chipless RFID tags with reduced cost are also available in literature. Inkjet printable eight bit tags have been realized in [16]. This method uses a transmission line with capacitive discontinuities using SMT (Surface Mountain Technology) technology. Based on whether the capacitors are connected or not connected to the transmission lines, there will be reflections or no reflections. This idea is used to code data.

Inksure technologies has proposed an easily printable SAR based RFID tag which has a read range of one foot [24]. Post processing is done in the received data to account for the diffraction effects between elements in the tag in order to increase the data accuracy.

But even with all the aforementioned techniques, the only commercial successful chipless RFID system is the one developed by RFSAW based on surface acoustic waves (SAWs) [13]. An acoustic wave device uses a piezoelectric material to generate the acoustic wave. When an oscillating electric field is applied to a piezoelectric acoustic wave sensor, an acoustic wave is created and propagates through the substrate. This is then converted back to an electric field. Based on the discontinuities, the reflected wave is modified, which can be analyzed to obtain the stored information. SAW tags are cost effective with large capacity of data storage.

Although, SAW tags are fully functional and can replace the chipped tags, they do not provide a fully printable solution due to their piezoelectric nature. Thus, they cannot be applied for low-cost substrates such as paper. This has further intensified the need of an all printable and compact tag with large data storage capability and small bandwidth so as to work within the allocated RFID bands.

We are working toward the design of an all printable and compact RF-RFID tag. As an example, a two bit chipless RFID tag based on the variations in phase has been studied. The phase variations are obtained by varying the topology of the reflecting element. Depending on the presence or absence of the tag, the phase of the reflected wave is altered. Figure 2 shows the phase of the reflected signal. It can be seen that large noticeable variations in phase is produced for various configurations of the reflecting element, due to changes in the current distribution.

It is also interesting to note that there are variations in the resonance (Fig. 3), which is obvious owing to changes in current distribution. This data can also be used to double check the received information.

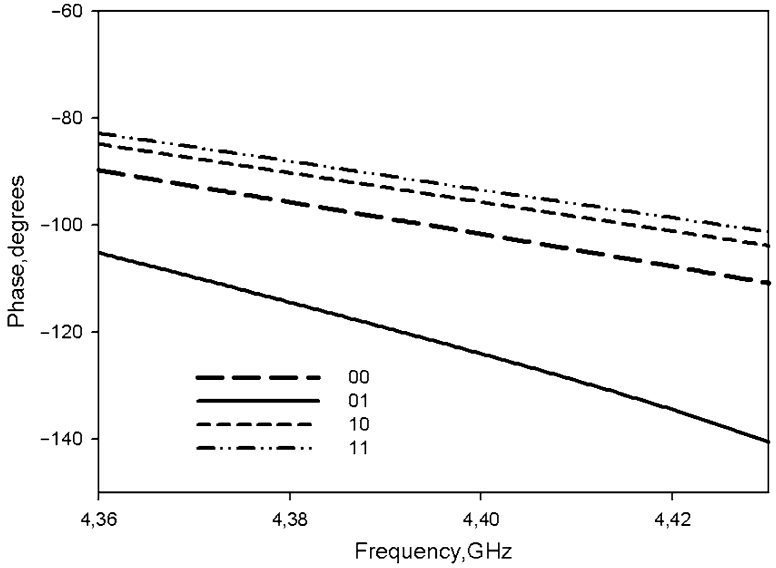


Fig. 2 Phase variations with different configurations corresponding to different codes

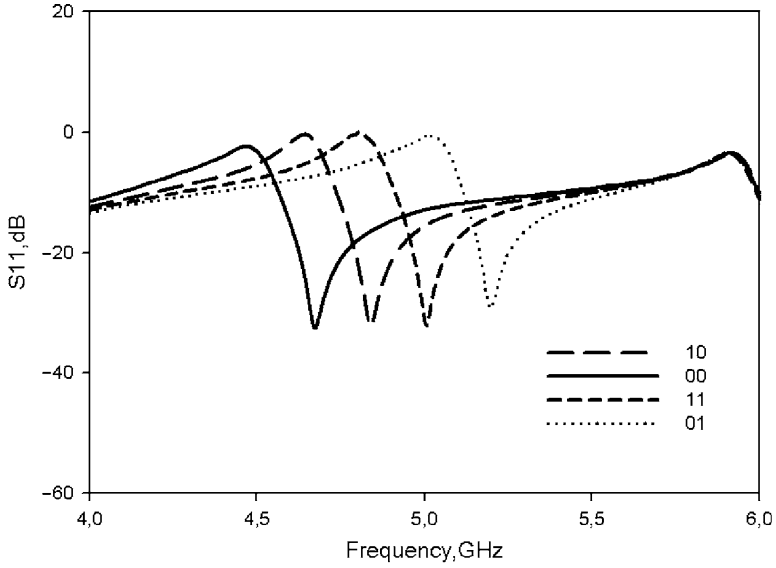


Fig. 3 Variation in resonance for different configurations

3.3 Towards THz Chipless

As shown previously, surface information would be carried out by RF structure that would be read by RF signals. We discuss in this part a different approach of the RFID where the information is carried out by a multi-layer structure whose dimensions would be compatible with the terahertz (THz) frequencies. In this case, the information coded in volume would be read by THz signals. The user will have at his disposal three possibilities of memorizing his information: either on the surface of the tag using RF signal (RFID), or in the volume or both. This latter solution brings flexibility and permits to deal with much more protected data. This new family of tags constitutes powerful communicating objects and ensures an optimal management of energy, since they are passive devices based on reflections. Moreover, their passive character makes them no forgeable, and the THz information coded in their volume makes them unreadable unless you are authorized.

It exists different approaches to code the THz information. This can be achieved either using the temporal response of the multilayer structure, or its frequency signature. Indeed, a reflection coefficient appears at each interface of the structure since the characteristic impedance of a dielectric material depends (in the THz domain) on its optical properties (permittivity and magnetic permeability), then an electromagnetic (EM) signal, emitted by a pulse source, is partially reflected at each interface of the multilayer structure generating echoes whose time delay and magnitudes depend on the geometrical and optical properties of each layer. In turn, a detector reads a temporal signature: the tag information is coded in the time domain. The reader is then able to detect two consecutive echoes as long as the pulse duration of the incoming wave is smaller than its back and fourth travel time within the thinner layer. Taking this into account, the source used must generate a pulsed signal with pulse durations not greater than few tens of femtoseconds for a 1 mm-thick tag. For such tag, the system requires a THz source.

Nevertheless, it is still possible to get specific information about the tag even if the THz source provides a continuous signal (CW source). Indeed the total EM wave reflected by the whole multilayer structure is the sum of all the reflected waves having been reflected at each interface. Since these reflected waves have the same frequencies but different magnitudes, the intensity of the total reflected (or transmitted) wave is due to the interferences between these phase-shifted reflected waves. As long as the relative phase shifts between the multiple reflection depend on the frequency of the CW incoming wave, on the one hand, and on the geometrical and optical properties of the tag, on the other hand, a frequency sweep of the CW source involves specific intensity modulation of the total reflected (or transmitted) EM wave. In turn, a detector reads a spectral signature: the tag information is coded in the frequency domain. Below we present some results obtained with this latter approach, and more specifically how it is possible to code several bits with a simple multilayer structure as a RFID chipless tag in the THz domain.

The studied THz tag on Fig. 4 consists of non-magnetic dielectric media arranged in a well-defined order, ensuring three different functions, which are required to identify precisely the tag and its information. The periodical stack of layers

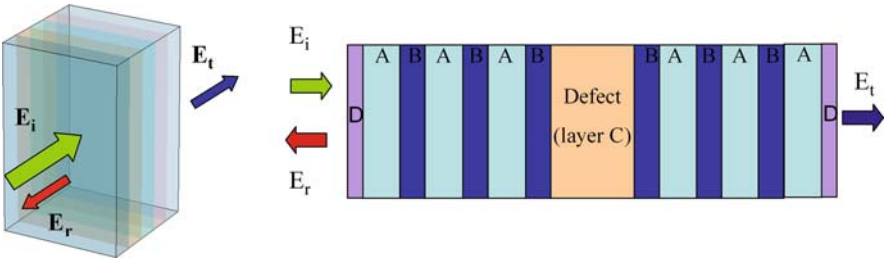


Fig. 4 Schematic of the THz chipless tag structure

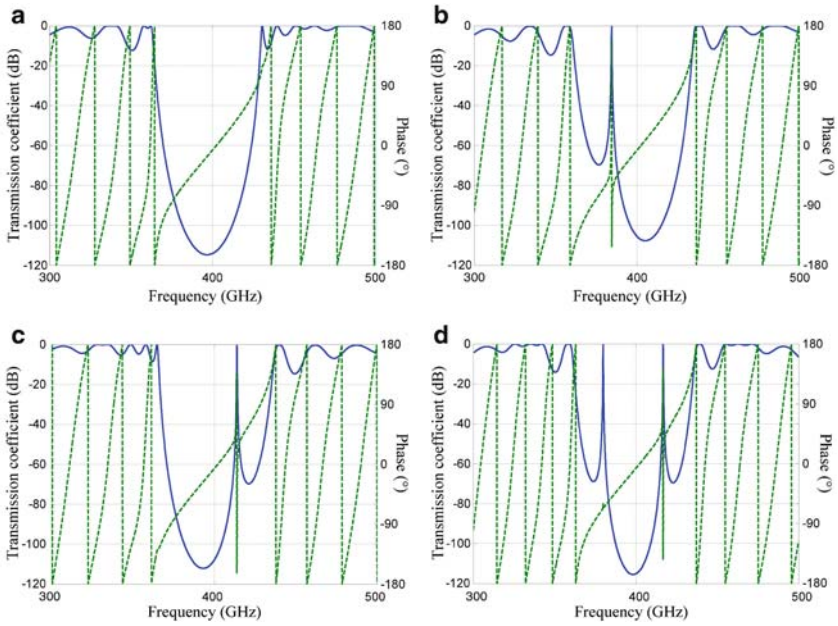


Fig. 5 Transmission coefficient and phase (dashed line) of a TEM ($\theta = 0°$) plan wave out coming from the THz tag, for different defect layer thicknesses

A and B is well-known as Bragg mirror. This periodic structure has a transmission coefficient, which depends on the frequency of the incident signal.

As depicted in Fig. 5b, this Bragg mirror prevents an incoming wave from being transmitted through the structure if its wavelength is included within a certain bandwidth whose spectral characteristics (rejection level, position and width) depend on the dielectric and geometrical properties of each periodically stacked layers A and B. The Bragg mirror is a 1-D photonic crystal, presenting a Photonic Band Gap (PBG) that allows coding spectral information. Indeed, introducing a layer C (defect layer Fig. 4) embedded by two 1-D photonic crystals, one creates a Fabry-Pérot

cavity [25] having frequency-dependent reflectivity since the bandwidth separating two consecutive transmission peaks depend on the optical length of the defect. Thus, either none or several peaks occur within the PBG. Then, presence or absence of those transmission peaks, called defect mode, code the useful information. As example of results, we consider the multi-layer structure on Fig. 4, to develop a THz tag whose spectral signature presents an orientation-free dependence: the reader must identify the tag regardless its relative orientation. The spectral response of the developed structure must be also independent of the polarization state of the incoming THz EM wave in order to fit the RFID applications requirements. The transmitted and reflected EM waves (Et and Er, respectively) are numerically calculated with the transfer matrix method [26], considering TE or TM planar incoming wave Ei, with any orientation θ about the tag surface. The simplest way to code the information is to read the presence/absence of a defect mode within the PBG. Figure 5 shows the spectral signatures of 4 tags having the same layers dimensions except for defect thickness.

The absence of defect mode in the both half bandwidth of the PBG (Fig. 5a) could be interpreted as "00". Therefore Fig. 5b sets for "10", Fig. 5c for "01" and Fig. 5d for "11". It means that, the number of states one can identify is given by the number of defect peaks measurable by a reader having a certain spectral resolution. In turn, to improve the number of states, one should first improve the spectral resolution of the reader and/or broaden the PBG bandwidth. This latter solution can be achieved developing, for example, Bragg mirrors with metamaterials.

4 Conclusion

The RFID technology is expanding rapidly and applied in many domains. It is becoming a part of our everyday life. The variety of applications and environments requires the development of quite different tags in order to meet the needs and the constraints of each situation. The tags can be grouped in two families regarding their composition. The first family is the chipped tag, based on the use of an IC chip, which contains the information. These tags are quite developed and are available in many different formats. Chipped tags continue to be investigated in order to overcome some limits such as data security, reliability and low-cost needs. The second family is the chipless form. These tags do not use IC chip and the information is directly coded on the surface and/or in the volume of the structure. These tags, also known as RF barcodes, are very attractive in term of cost and data security. Many research projects worldwide are dedicated to the development of efficient and versatile chipless tags.

Chipless tags are less than 1% of the RFID market today. Due to their low cost and great flexibility, some market projections show that chipless tags will reach 60% of the RFID market before the end of the next decade. This is the reason why we can consider chipless as the next RFID frontier.

Acknowledgments The authors would like to thank Prof. L. Duvillaret and Dr. F. Garet, for guidance and fruitful discussions on part of this work and H. Chaabane for his help. The authors are grateful to Grenoble-inp for supporting this project via the BQR.

References

1. Stockman H (1948) Communication by means of reflected power. Proceeding of the IRE, pp 1196–1204
2. Landt J (2001) Shrouds of time: the history of RFID. www.aimglobal.org/technologies/rfid/resources/shrouds_of_time.pdf
3. IDTechEx Knowledgebase. www.IdtechEx.com
4. Finkenzeller K (2004) RFID handbook: fundamentals and applications. Wiley
5. Preradovic S, Karmakar NC, Balbin I (2008) RFID transponders. IEEE Microw Mag 2: 90–103
6. (2005) Identification friend or foe IFF systems: IFF questions & answers. Dean Boys. www.dean-boys.com/extras/iff/iffqa.html
7. Rieback MR, Crispo B, Tanenbaum AS (2006) The evolution of RFID security. IEEE Pervasive Comput 5: 62–69. Published by the IEEE CS and IEEE ComSoc
8. Bechevet D, Vuong TP,Tedjini S (2005) Design and measurements of antennas for RFID, made by conductive ink on plastics. IEEE AP Symposium, 3–8 July 2005
9. Bechevet D (2005) Contribution au développement de tags RFID en UHF et Microondes, sur matériaux plastiques. Ph.D. Grenoble-inp
10. Ghiotto A (2008) Conception d'antennes de tags RFID UHF, application à la réalisation par jet de matière. Ph.D. Grenoble-inp
11. Chaabane H, Perret E,Tedjini S (2009) Conception automatisée d'un tag RFID UHF robuste, 16émes Journées Nationales Micro-Ondes, 2009, Grenoble (France), 27–29 Mai 2009
12. Das R (2005) Chip versus chipless for RFID applications. ACM international conference proceeding series, Grenoble France, vol 121, pp 23–26
13. Hartmann CS (2002) A global SAW ID tag with large data capacity. IEEE Ultrasonics Symp 1: 65–69
14. Subramanian V et al (2005) Progress toward development of all-printed RFID tags: materials, processes, and devices. Proc IEEE 93:1330–1338
15. Zhang L et al (2006) An innovative fully printable RFID technology based on high speed TDR. High Density Microsys Des Pack Comp Fail:166–170
16. Zheng L et al (2008) Design and implementation of a fully reconfigurable chipless RFID tag using Inkjet printing technology. IEEE (ISCAS 2008), 18–21 May 2008, pp 1524–1527
17. Preradovic S et al (2008) A novel chipless RFID system based on planar multiresonators for barcode replacement. Proceeding of the 2008 IEEE RFID, Las Vegas, April 2008, pp 289–296
18. Preradovic S, Balbin I, Karmakar NC, Swiegers G (2008) Chipless frequency signature based RFID transponders. 38th EuMC, Amsterdam, Oct 2008, pp 1723–1726
19. Jalaly I, Robertson ID (2005) Capacitively-tuned split microstrip resonators for RFID bar codes. European Microwave Conference 2005, vol 2, pp 4–6
20. Jalaly I, Robertson ID (2005) RF barcodes using multiple frequency bands. IEEE MTT-S Digest, June 2005
21. Mukherjee S (2007) Chipless RFID device. RFID Eurasia 2007, Istanbul, pp 1–4
22. Mukherjee S (2007) Chipless radio frequency identification by remote measurement of complex impedance. European Microwave Conference Munich, pp 1007–1010
23. Cumming DRS, Drysdale TD (2003) Security tag. Patent, GB 0305606.6, 12 March 2003
24. Taylor D (2009) Introducing SAR code – an unique Chipless RFID technology. RFID journal, 7th Annual conference, Orlando April 2009
25. Fabry C, Perot A (1897) Sur les franges des lames minces argentées et leur application a la mesure de petites épaisseurs d'air. Ann Chim Phys 12: 459–501
26. Nlmec H, Kužel P, Garet F et al (2004) Time-domain terahertz study of defect formation in one-dimensional photonic crystals, Appl Opt 43:1965–1970

Backscatter Communication Using Ultrawide Bandwidth Signals for RFID Applications

F. Guidi, D. Dardari, C. Roblin, and A. Sibille

1 Introduction

RFID technology for use in real-time object identification is being rapidly adopted in several fields such as logistic, automotive, surveillance, automation systems, etc. [1]. A radiofrequency identification (RFID) system consists of readers and tags applied to objects. The reader interrogates the tags via a wireless link to obtain the data stored on them. The cheapest RFID tags with the largest commercial potential are passive or semi-passive, and the energy necessary for tag–reader communication is harvested from the reader's signal. Passive RFID tags are usually based on backscatter modulation, where the antenna reflection properties are changed according to information data [2].

Future advanced RFID systems are expected to provide both reliable identification and high-definition localization of tags. Accurate real-time localization, high security, management of large number of tags, in addition to extremely low power consumption, small size and low cost, will be the new important requirements [3]. Unfortunately, most of these requirements cannot be fulfilled by the current first and second generation RFID [1] or wireless sensor network (WSN) technologies such as those based on ZigBee standard [4]. In fact, RFID systems using standard continuous wave (CW)-oriented communication in the ultra-high frequency (UHF) band have an insufficient range resolution to achieve accurate localization, are affected by multipath signal cancellation (due to the extreme narrow bandwidth signal), are very sensitive to narrowband interference and multiuser interference, and have an intrinsic low security [5].

A promising wireless technique for next generation RFID is the ultrawide bandwidth (UWB) technology characterized, in its impulse radio UWB (IR-UWB) implementation, by the transmission of subnanosecond duration pulses.

F. Guidi (✉) and D. Dardari
WiLAB, DEIS, University of Bologna at Cesena, via Venezia 52, 47023 Cesena (FC), Italy
e-mail: francesco.guidi5@unibo.it; ddardari@ieee.org

C. Roblin and A. Sibille
ENSTA-ParisTech, 32 Boulevard Victor, 75739 Paris Cedex 15, France
e-mail: christophe.roblin@ensta.fr; alain.sibille@ensta.fr

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_24,

The potential advantages of UWB include, but are not limited to, low power consumption at the transmitter side, extremely accurate ranging and positioning capability at submeter level, robustness to multipath (better area coverage), low detection probability (higher security), large number of devices operating and coexisting in small areas (efficient multiple channel access and interference mitigation) [6–8].

Passive tags based on backscatter modulation appears very promising, especially in the perspective of the adoption of efficient energy scavenging techniques for tag's control logic alimentation. Recently, some applications of the UWB technology in tags based on backscatter modulation have been proposed in [9–11].

In this chapter, the performance of general UWB–RFID system architecture proposed in [9] and [11], based on backscatter modulation, is characterized using experimental data with the purpose of investigating the performance in realistic environment, where strong clutter is present.

2 UWB Backscatter Propagation

Passive RFID tags are based on backscatter modulation, where the antenna reflection properties are changed according to information data [2]. In general, when an electromagnetic (e.m.) wave encounters an antenna, it is partially reflected back depending on antenna configuration. Antenna scattering mechanism is composed of structural and antenna mode scattering [12]. Structural mode occurs due to the antenna's given shape and material and it is independent on how the antenna is loaded. On the contrary, antenna mode scattering is a function of the antenna load, thus data can be sent back to the reader through a proper variation of the antenna load characteristic without requiring a dedicated power source (backscatter modulation). This property is currently adopted in traditional passive UHF-RFID tags based on CW signals to carry information from the tag to the reader.

While UHF-RFIDs have an extensive literature dealing with this issue [2, 5, 12], the characterization of backscatter properties when operating with UWB signals deserves further investigations especially in realistic environments [13]. In the following paragraphs, we will give an overview of the UWB antenna backscatter properties.

2.1 UWB Antenna Backscattering

When a UWB pulse is transmitted and UWB antennas are employed, the reflected signal takes the form reported in Fig. 1, where the structural and antenna modes scattered components are plotted separated for convenience. The antenna mode scattered signal can be varied according to the antenna load Z_{L}, whereas the scattering of the structural mode will remain the same. Among the various possibilities, three

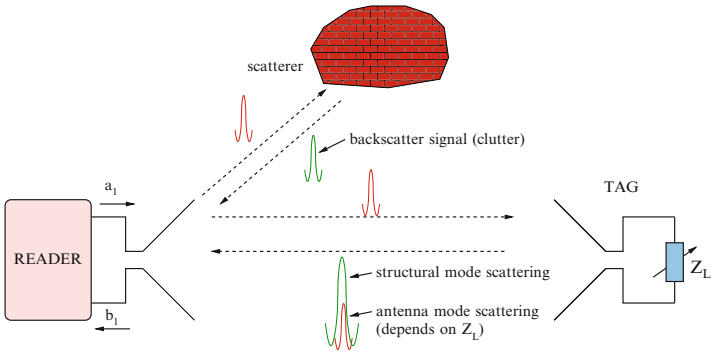


Fig. 1 Example of backscatter mechanism of the transmitted pulse due to tag's antenna and the presence of scatterers

particular choices are of interest for passive UWB RFID: $Z_{\mathrm{L}} = 0$ (short circuit), $Z_{\mathrm{L}} = \infty$ (open circuit) and $Z_{\mathrm{L}} = Z_{\mathrm{A}}$ (matched load), where Z_{A} is the antenna impedance. Ideally, antenna mode scattered waveforms have a phase difference of 180° between the case of open and short circuit loads, whereas no antenna mode scattering exists in case of perfect matched load. In UWB antennas, the structural mode component takes a significant role in the total scattered signal; in fact, it is typically 1 or 2 orders of magnitude higher than that of the antenna mode [12, 13]. In addition, signals scattered by the surrounding environment (clutter) are inevitably present and superimposed to the useful signal.

2.2 The Round-Trip Channel Transfer Function

To analyze the performance of backscatter modulation schemes, a proper model for the reader–tag–reader interaction is needed.

Let us consider a reference scenario as shown in Fig. 2, where a couple of UWB antennas, acting as tag and reader, located at distance d are present. The e.m. wave at the tag's antenna is partially backscattered according to the antenna scattering characteristics which depend on the antenna load (different load configurations will be referred to as tag status X) and reader–tag orientation in the 3D space.

Generally the reader's antenna emits a very short pulse $p(t)$. We denote $w(t; X)$ the backscattered signal, received back by the reader's antenna, due to the tag's antenna mode, whose shape and energy are a function of the tag status X (open, short, loaded) as well as on the tag orientation, distance, and e.m. polarization. In the frequency domain, it is

$$W(f; X) = P(f)\, H(f; X), \tag{1}$$

where $H(f; X)$ is the transfer function of the round trip channel (reader–tag–reader) and the dependence on the tag status X is shown explicitly.

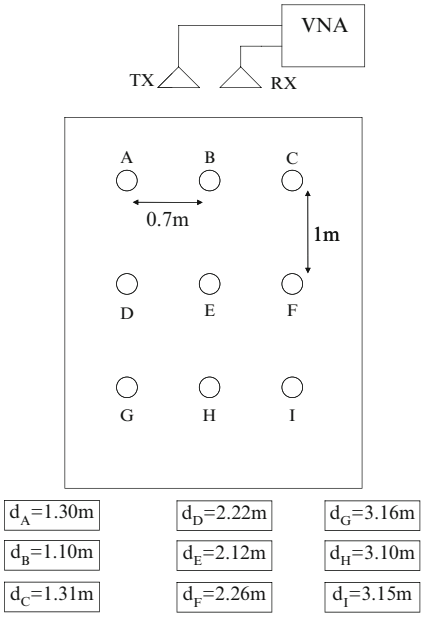


Fig. 2 Indoor scenario considered for the measurement campaign at ENSTA-ParisTech. The distances between each point and the antennas connected to the VNA are also reported

3 Experimental Characterization of UWB Antenna Backscatter

The backscattering signal from a UWB antenna was measured both in anechoic chamber and in a typical indoor environment at ENSTA-ParisTech laboratory. Measurements were performed in the frequency domain, by means of a vector network analyzer (VNA) in the 2–12 GHz band with steps of 5 MHz. Two Horn Lindgren 3,117 antennas were employed as reference antennas. They were placed in a quasi-monostatic configuration, separated by 18 cm, guaranteeing a high isolation between the transmit and receive channels as shown in Fig. 2. The scattering from a UWB monopole dual feed stripline (DFMS) [14], which is a small planar antenna, was measured. Its dimensions ($40 \times 24 \times 3\,\text{mm}^3$) make this antenna quite suitable for RFID tags. Three different load conditions were considered. Inside the anechoic chamber, the antenna under test (AUT) was placed on a 3D positioner at $d_{\text{ref}} = 1.46\,\text{m}$ from the reference antennas.

In indoor environment, a rectangular grid of nine points as showed in Fig. 2, spaced out of about 1 m in depth and 70 cm in width, was defined in a room with furniture and having dimensions ($5.13 \times 4.49\,\text{m}^2$). The tag was positioned alternatively in each point on a vertical support. For both cases, a simple data processing was performed to obtain the antenna backscattering response from the measured S_{21}. The collected data was first filtered in the frequency domain to avoid ringing effect. Then, applying the inverse Fourier transform, the signal in the time domain was derived.

In Fig. 3, an example of backscattered signal received from the tag placed at point H (distance 3.10 m) is reported. As can be noted, several clutter components

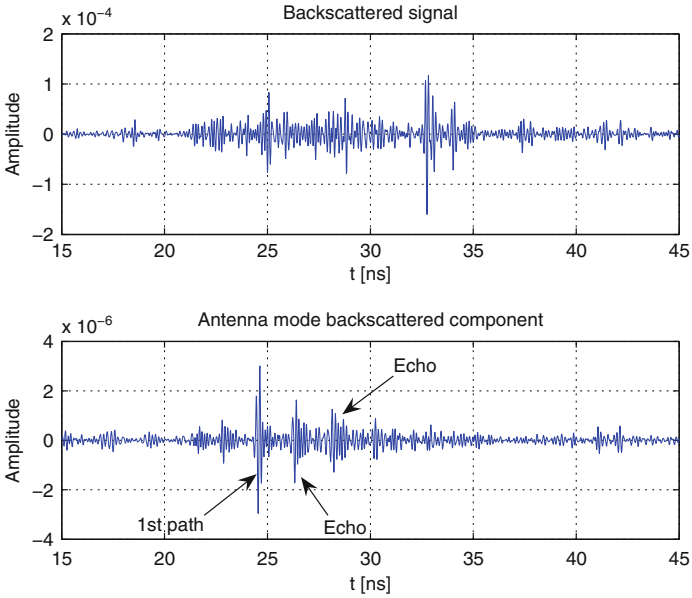


Fig. 3 Example of the laboratory impulse response (grid point H, at distance of 3.10 m) and of the antenna mode contribution (after clutter removal). In particular, it is evident the presence of some echoes after the first direct path

(including the antenna structural mode) are present. The second plot shows the antenna mode backscattered signal (that of interest) after clutter removal. Due to its small amplitude (about 2 orders of magnitude less than the clutter), it is completely buried below the clutter component.

Then, clutter and the antenna structural mode scattering have a significant impact at the reader's antenna, thus making the detection of the antenna mode scattered signal (which carries data) a main issue in passive UWB RFID systems which has not been widely addressed yet. To this purpose, ad hoc robust backscatter modulation schemes have to be designed as will be illustrated in Sect. 4.

4 Tags–Reader Backscatter Communication Using UWB Signals

In Fig. 4, the architectures proposed in [9, 11] for tag and reader are reported. The reader is composed of a transmitter and a receiver section both connected to the same UWB antenna through a TX/RX switch. During the interrogation phase, the reader transmits a sequence of UWB pulses, each having energy E_t, modulated by a periodic binary sequence $\{d_n\}$ of period N_s, with $d_n \in \{-1, +1\}$, specific of

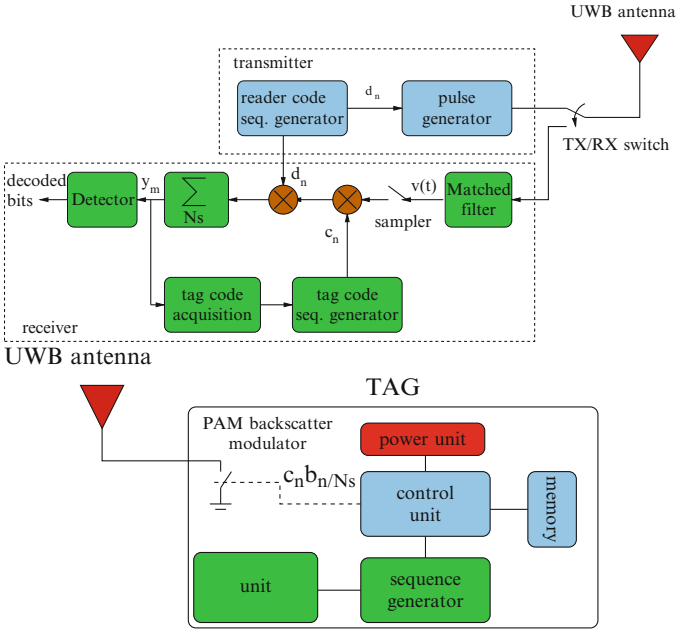


Fig. 4 The considered scheme of the tag and the reader [11]

that particular reader (reader's code sequence). Without loss of generality, we consider an infinite sequence of pulses spaced apart of T_f seconds (frame time), that is,

$$s(t) = \sum_{n=-\infty}^{\infty} d_n \cdot p(t - nT_\mathrm{f}) \tag{2}$$

characterized by a transmission power P_t. The frame time T_f has to be chosen to make the intersymbol interference due to multipath, negligible. In indoor scenario, $T_\mathrm{f} = 50$–$100\,\mathrm{ns}$ is usually sufficient.

During the transmission of each pulse, the antenna is connected to the transmitter section while it is kept connected to the receiver section during the remaining time until the successive pulse is transmitted. Each pulse in (2) is backscattered by the tag's antenna as well as by all the surrounding scatterers present in the environment which form the clutter component. If the frame time T_f is properly designed, all backscattered signals are received before the transmission of the successive pulse. The main task of the receiver section of the reader is to detect the useful backscattered signal (i.e., the antenna mode scattering dependent on antenna load changes) from those backscattered by the antenna structural mode and other scatterers (clutter) which are, in general, dominant. To this purpose, in [9], a quite general backscatter modulator architecture is proposed allowing for different signaling schemes such as pulse amplitude modulation (PAM), pulse position

modulation (PPM), and ON–OFF keying. We analyze the performance of a simplified version of the tag architecture in [9] by considering the 2-PAM signaling. In this case, the backscatter modulator reduces to a simple switch as shown in Fig. 4. To make the uplink communication robust to the presence of clutter, interference and to allow multiple access, the tag is designed to change its status (short or open circuit) at each frame time T_f according to the data to be transmitted and a zero mean periodic tag's code sequence $\{c_n\}$, with $c_n \in \{-1, +1\}$, of period N_s [9, 11]. Specifically, each tag information symbol $b_k \in \{-1, +1\}$ is associated to N_s pulses, thus the symbol time results $T_\mathrm{s} = T_\mathrm{f} N_\mathrm{s}$. Therefore, the tag status x_n at the nth frame time is $x_n = (c_n\, b_{\lfloor n/N_\mathrm{s} \rfloor} + 1)/2$. It can take the values $x_n = 0$ (open circuit) and $x_n = 1$ (short circuit). In this way, the polarity of the reflected signal changes according to the tag's code sequence during a symbol time, whereas the information symbol affects the entire sequence pulse polarity at each symbol time. The received signal at the reader is

$$r(t) = \sum_{n=-\infty}^{\infty} d_n\, w(t - nT_\mathrm{f}; x_n) + \sum_{n=-\infty}^{\infty} d_n\, w^{(\mathrm{c})}(t - nT_\mathrm{f}) + n(t), \tag{3}$$

where $n(t)$ is the additive white gaussian noise (AWGN) with two-sided power spectra density $N_0/2$. The function $w(t; x_n)$ represents the backscattered pulse modulated by the tag's antenna load and the signal $w^{(\mathrm{c})}(t)$ contains the backscattered version of the pulse $p(t)$ due to the clutter component as well as antenna structural mode scattering. When a symmetric 2-PAM signaling scheme is adopted and perfect pulse symmetry is considered (i.e., $w(t; 1) = -w(t; 0)$), it is $w(t; x_n) = c_n \cdot b_{\lfloor n/N_s \rfloor} \cdot w(t; 0)$ and (3) becomes

$$\begin{aligned} r(t) = &\sum_{n=-\infty}^{\infty} d_n\, c_n\, b_{\lfloor n/N_s \rfloor}\, w(t - nT_\mathrm{f}; 0) + \\ &+ \sum_{n=-\infty}^{\infty} d_n\, w^{(\mathrm{c})}(t - nT_\mathrm{f}) + n(t)\,. \end{aligned} \tag{4}$$

Considering the receiver scheme reported in Fig. 4, where a matched filter (MF) demodulator is adopted, the received signal is first passed through a filter having impulse response $h(t)$ specified later, to obtain the signal $v(t) = r(t) \otimes h(t)$, where $\otimes$ is the convolution operator.

Supposing perfect code and synchronization to the time instant τ_0 corresponding to the maximum peak of the filter output,[1] $v(t)$ is sampled at sampling intervals $t_{i,m} = i\, T_\mathrm{f} + m\, N_\mathrm{s}\, T_\mathrm{f} + \tau_0$, with $i = 0, 1, ..., N_\mathrm{s} - 1$, thus obtaining the samples

[1] In general, tag's internal clock is asynchronous with respect to reader's internal clock. For this reason a suitable synchronization scheme has to be considered at the reader as will be investigated in the following up paper. Time delay τ_0 accounts for filter delay as well as round-trip propagation delay. Its estimate can be useful to measure the distance (ranging) between tag and reader [15].

$$
\begin{aligned}
v_{i,m} &= v(t_{i,m}) \\
&= \sum_{n=-\infty}^{\infty} d_n\, c_n\, b_{\lfloor n/N_s \rfloor}\, \gamma(i\,T_\mathrm{f} - n T_\mathrm{f} + m N_\mathrm{s} T_\mathrm{f} + \tau_0) \; + \\
&\quad + \sum_{n=-\infty}^{\infty} d_n\, \gamma^{(\mathrm{c})}(i\,T_\mathrm{f} - n T_\mathrm{f} + m N_\mathrm{s} T_\mathrm{f} + \tau_0) \; + \\
&\quad + z(i\,T_\mathrm{f} + m N_\mathrm{s} T_\mathrm{f} + \tau_0),
\end{aligned}
\tag{5}
$$

where $\gamma(t;x) \triangleq w(t;x) \otimes h(t)$, $\gamma^{(\mathrm{c})}(t) \triangleq w^{(\mathrm{c})}(t) \otimes h(t)$ and $z(t) \triangleq n(t) \otimes h(t)$.

Looking at (5), it can be noted that only the antenna mode scattered signals result to be modulated by the combination of the tag's and reader's code sequences $\{c_n\}$ and $\{d_n\}$, whereas all clutter signals components (included the antenna structural mode scattering) are received modulated only by the reader's code sequence $\{d_n\}$. Then, as shown in Fig. 4, to remove the clutter component at the receiver, the sampled signal $v_{i,m}$ is multiplied by the composite sequence $\{c_n \cdot d_n\}$, which identifies both the reader and the desired tag.[2] All N_s resulting samples composing a data symbol are summed up to form the decision variable at the detector input

$$
\begin{aligned}
y_m = &\sum_{i=0}^{N_\mathrm{s}-1} c_i\, d_i\, c_{i+mN_\mathrm{s}}\, d_{i+mN_\mathrm{s}}\, b_{\lfloor m+i/N_s \rfloor}\, \gamma(\tau_0) \; + \sum_{i=0}^{N_\mathrm{s}-1} \sum_{n=-\infty}^{\infty} c_i\, d_i\, d_{i+mN_\mathrm{s}} \\
&\gamma^{(\mathrm{c})}(i\,T_\mathrm{f} - n T_\mathrm{f} + m N_\mathrm{s} T_\mathrm{f} + \tau_0) \; + \; z_m,
\end{aligned}
\tag{6}
$$

where $z_m \triangleq \sum_{i=0}^{N_\mathrm{s}-1} d_i\, c_i\, z(i\,T_\mathrm{f} + m N_\mathrm{s} T_\mathrm{f} + \tau_0)$ is Gaussian distributed random variable (r.v.) with zero mean and variance $\sigma_z^2 = N_\mathrm{s}\, N_0\, E_\mathrm{h}/2$ where E_h is the energy of $h(t)$. Considering the clutter component time-invariant within a symbol time (i.e., quasistatic scenario) and that $c_{i+mN_\mathrm{s}} = c_i$, $b_{\lfloor m+i/N_\mathrm{s} \rfloor} = b_m$ and $d_{i+mN_\mathrm{s}} = d_i$ with $i = 0, 1, ..., N_\mathrm{s} - 1$, the decision variable for the mth symbol b_m becomes

$$
y_m = b_m\, \gamma(\tau_0)\, N_\mathrm{s} \; + \gamma^{(\mathrm{c})}(\tau_0) \sum_{i=0}^{N_\mathrm{s}-1} c_i \; + z_m \,. \tag{7}
$$

As can be deduced from (7), to completely remove the clutter component, it is sufficient that the tag's code sequence $\{c_n\}$ has zero mean, i.e., $\sum_{n=0}^{N_\mathrm{s}-1} c_n = 0$. In such a case, (7) can be further simplified leading to

$$
y_m = b_m\, N_\mathrm{s}\, \rho\, \sqrt{E_\mathrm{r}}\, \sqrt{E_\mathrm{h}} + z_m \,, \tag{8}
$$

[2] Multiple readers can access the same tag by using different reader sequences provided that they are designed with good crosscorrelation properties.

where ρ is the normalized crosscorrelation between pulses $w(t;0)$ and $h(t)$, which accounts for the mismatch due to pulse distortion. Parameter E_r represents the received energy. Using (1) E_r is given by

$$E_r = \int_{-\infty}^{\infty} |w(t;0)|^2 \, dt \, . \tag{9}$$

In a single-user scenario, the bit error probability is given by

$$P_b = \frac{1}{2}\text{erfc}\sqrt{\frac{E_r N_s \rho^2}{N_0}} = \frac{1}{2}\text{erfc}\sqrt{\frac{P_t G \rho^2}{N_0 R_b}}, \tag{10}$$

where $R_b = 1/(N_s T_f)$ is the data rate (symbol rate), ρ accounts for the ability of the receiver to collect the energy coming from different propagation paths, $G \triangleq E_r/E_t$ is the round-trip channel average power gain and P_t is the transmitted power.

The simplest UWB receiver is the single-path matched filter (SPMF) [8], where the received signal is correlated with a local replica (template) of the transmitted pulse $p(t)$ or, equivalently, is filtered by a filter matched to $p(t)$. The performance of the SPMF, or any other receiver solutions such as those based on Rake structures, are bounded by that of the ideal matched filter (IMF) receiver, where all the energy deriving from the backscattered signal is supposed to be captured (i.e., $h(t) \equiv w(-t;0)$ and hence $\rho = 1$).

5 Numerical Results

We now investigate the bit error probability of the UWB–RFID system described in this paper as a function of the data rate R_b. The transmitted pulse shape and power have been chosen so that the transmitted signal is compliant with the 3.1–10.6 GHz FCC mask. Specifically, the 6th derivative Gaussian monocycle has been considered. The other parameters are $T_f = 100$ ns and the effective isotropically radiated power is EIRP$= -6.70$ dBm. Results related to the SPMF receiver in AWGN channel (obtained using measurement data in anechoic chamber considering a distance $d = 1.46$ m) and in every grid point location inside the laboratory are shown in Figs. 5 and 6, respectively. As expected, the performance of the IMF receiver is significantly better than that obtained using the simple SPMF receiver because all useful energy coming from the multipath is captured and hence the performance depends only on the received power. In fact, the SPMF receiver is not able to collect the energy from multipath and suffers from pulse distortion due to propagation as well as antenna effects. Their problem can be mitigated by considering more complex receiver structures such as those based on Rake solutions. The performance obtained with the tag located at point B is better than AWGN condition because

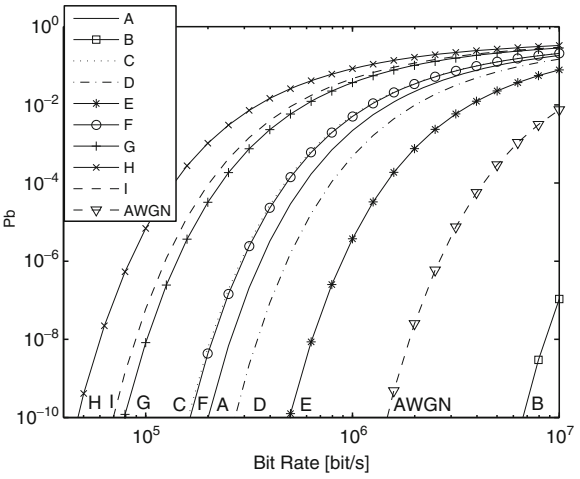


Fig. 5 Bit error probability as a function of the bit rate R_b in different tag locations. SPMF receiver is considered

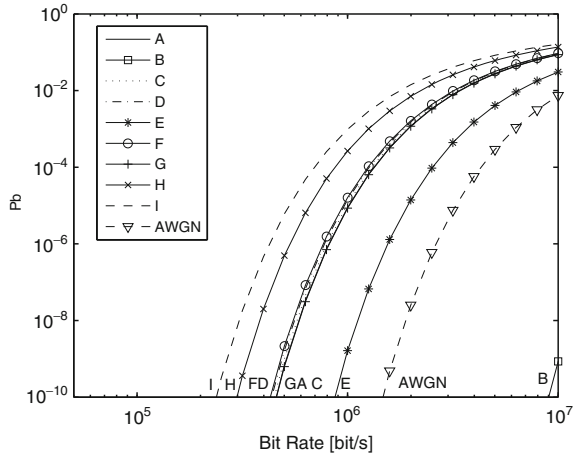


Fig. 6 Bit error probability as a function of the bit rate R_b in different tag locations. IMF is considered

of the shorter distance (1.10 m vs. 1.46 m). It can be noted that in some cases, tags placed at larger distances correspond to higher performance. This result depends on the higher amount of energy that can be collected in some locations due to the presence of richer multipath.

If we fix the target bit error probability (e.g. $P_b = 10^{-3}$), results show that data rates up to 200 kbit/s at a distance of 3.10 m with a transmitted power lower than 1 mW are feasible in a realistic environment. It is important to remark that similar performance is achievable by the conventional UHF RFID technology only using a transmitted power higher than 1 W (i.e., more than 3 orders of magnitude higher level).

6 Conclusions

In this paper, the attainable performance of passive UWB–RFID based on backscatter modulation has been analyzed using measured data in controlled and realistic environments. It has been shown that, with proper signal design and processing, the clutter due to scatters and tag's antenna structural backscattering can be easily removed. Results indicate that, even using simple receiver structures, distances up to 3 m can be covered with data rates larger than 200 kbit/s with a transmitted power less than 1 mW.

Acknowledgment This work has been performed within the framework FP7 European Project EUWB (grant no. 215669).

References

1. Finkenzeller K (2004) RFID handbook: fundamentals and applications in contactless smart cards and identification, 2nd edn. Wiley, New York
2. Chawla V, Ha DS (2007) An overview of passive RFID. IEEE applications & practice, pp 11–17
3. To strengthen european technology: The support of RFID within FP7 (2007) European Commission DG Information Society and Media/Unit G2 "Microsystems," September 2007
4. Verdone R, Dardari D, Mazzini G, Conti A (2008) Wireless sensor and actuator networks: technologies, analysis and design. Elsevier, Amsterdam
5. Kim D, Ingram M, Smith W (2001) Small-scale fading for an indoor wireless channel with modulated backscatter. In: Vehicular technology conference, 2001. VTC 2001 Fall. IEEE VTS 54th, vol 3. Atlantic City, NJ, pp 1616–1620
6. Proceedings of IEEE, Special issue on UWB technology & emerging applications (2009)
7. Ha D, Schaumont P (2007) Replacing cryptography with ultra wideband (UWB) modulation in secure RFID. In: 2007 IEEE international conference on RFID, Grapevine, TX
8. Win MZ, Scholtz RA (1998) Impulse radio: how it works. IEEE Commun Lett 2(2):36–38
9. Dardari D (2008) Metodo e apparato per la comunicazione in sistemi RFID a banda ultra-larga (methods and apparatus for the communication in ultrawide bandwidth RFID systems). Italy Patent Application MO2008A000053, 29 February 2008
10. Dardari D (2004) Pseudo-random active UWB reflectors for accurate ranging. IEEE Commun Lett 8(10):608–610
11. Dardari D, D'Errico R (2008) Passive ultrawide bandwidth RFID. In: IEEE Global communications conference (GLOBECOM 2008), New Orleans, LA
12. Penttila K, Keskilammi M, Sydanheimo L, Kivikoski M (2006) Radar cross-section analysis for passive RFID systems. IEE Proc Microw Antennas Propag 153(1):103–109
13. Hu S, Law CL, Shen Z, Zhu L, Zhang W, Dou W (2007) Backscattering cross section of ultrawideband antennas. IEEE Antennas Wirel Propagat Lett 6:70–72
14. Bories S, Ghannoum H, Roblin C (2005) Robust planar stripline monopole for UWB terminal applications. Ultra-wideband, 2005 IEEE international conference, pp 80–84
15. Dardari D, Conti A, Ferner U, Giorgetti A, Win MZ (2009) Ranging with ultrawide bandwidth signals in multipath environments. Proc IEEE (Special issue on UWB technology & emerging applications) 97(2):404–426

Passive RFID Integrated Transponders for Automotive Applications

Alberto Toccafondi, Cristian Della Giovampaola, Paolo Braconi, and Alessio Cucini

1 Introduction

Radio frequency identification (RFID) is a promising technology which uses radio frequency electromagnetic waves for the automatic identification of objects or items [3]. RFID systems are typically composed of a reader station and small transponders or tags located on the objects to be identified. The identification process takes place through the RFID reader which interrogates a specific volume and collects information about the objects, exchanging wireless data with the object's tags located in the mentioned volume.

RFID systems with a variety of operational frequency, country regulations and techniques have been introduced. Among them, HF (3–30 MHz, ISO 15693) and UHF (865–956 MHz) system are probably the most used all over the world. Passive RFID systems operating at HF band (13.56 MHz) are widely used in the area of object tracking and access control. They use a near field inductive coupling to transfer both power and binary data between the reader and the tags. These systems offer an excellent immunity to environmental noise and electrical interference and exhibit a minimal shielding effect from adjacent objects and human body. However, due to country regulations on the permitted magnetic field strength, the maximum achievable reading range is limited to approximately 1 m.

Passive RFID systems operating at UHF band have recently drawn a great deal of attention because of their numerous benefits such as low cost, middle to long reading range, good reliability of the communication links and a large data storage capability of the tags. Typical commercial applications of these systems are in access control services as well as in process or life cycles traceability. UHF passive RFID

A. Toccafondi (✉) and C.D. Giovampaola
Department of Information Engineering, University of Siena, Via Roma 56, Siena 53100, Italy
e-mail: albertot@dii.unisi.it; dellagiovampaola@dii.unisi.it

P. Braconi and A. Cucini
Wavecomm S.r.L., Loc. Belvedere, 53034 Colle Val d'Elsa, Siena 53100, Italy
e-mail: braconi@wavecomm.it; cucini@wavecomm.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_25,

systems (865–870 MHz in Europe, 902–928 MHz in North and South America and 950–956 MHz in Japan) and microwave (2.45 GHz) bands use the modulated scattered technique to establish a radio link between the reader and the tags. Here, the reflected signal from the tag is modulated by an integrated microchip (IC) directly connected to the antenna. As a consequence, a good RFID transponder performance, regarding the reading range and data rate transmission, may be obtained only with a correct conjugate matching between the tag antenna and the IC. Depending on both the IC sensitivity and the tag antenna performance, typical ranges of 4–6 m can now be achieved using passive UHF backscatter transponders.

Recently, significant efforts have been made in developing RFID tags and intelligent transponders for vehicle to road-side communications, to automate the vehicular control access and road-tolling operations [1, 4]. In the identification of moving vehicles, active tags are the typical choice due to their very extended communication range and high operational efficiency [2]. However, the use of active tags presents some impairments such as the high cost for mass production, when compared to passive ones, and the life expectancy of the battery. Moreover, the long reading distance may cause some problems in controlling the reader detection volume, which may result in possible wrong detections by the reader when it operates in a multilane identification environment. The use of passive tags is more advantageous due to their low cost, compactness and maintenance free.

This paper presents and analyzes an RFID passive system for the identification of moving vehicles within a monolane scenario. The system in characterized by the use of transponders which integrate the two leading passive technology for RFID identification, namely an ISO 15693 tag operating at HF band, and an EPC GEN2 tag operating at UHF band. The typical applications of such integrated transponders are in multiservice systems, where a high level of interoperability between different systems is required [5, 10]. In this case, a single removable card equipped with the two transponders is conceived to be used for both vehicle identification and personal access to controlled areas (see Fig. 1). To this end, a single-lane vehicle identification scenario has been simulated and analyzed, in order to assess the effects of the various multipath contributions originated by the presence of the car body, since they may reach the passive tag. It is found that in some conditions the presence

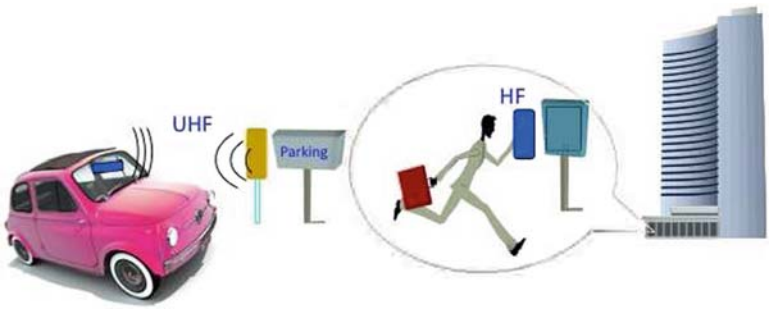


Fig. 1 Multi-system scenario

of multiple signal paths tends to enhance the minimum field intensity required to activate a passive tag, thus improving the achievable reading distance and the interrogating volume. Next, an integrated UHF-HF RFID passive tag has been designed. It is composed of a UHF (European band 865–870 MHz) and an HF (13.56 MHz) ISO 15693 commercial tags arranged in two separate sections on an equivalent ISO 7810 ID-1 card space. The UHF tag antenna has been conceived to provide small size and proper conjugate matching to the high-capacitive input impedance of the tag IC. The antenna design has been optimized taking into account both the presence of the commercial ISO 15693 tag located in the HF section and the presence of the car windshield. Finally, a prototype of the integrated transponder printed on FR4 substrate has been developed and tested in a real scenario.

2 Vehicle Identification Scenario

The identification of a moving vehicle in a real environment can be quite difficult, because of multiple path signals interferences that may occur during the interrogating operations. The main cause of multiple paths is reflections and diffractions that the illuminating field undergoes at the metallic surfaces and edges of the car body. In operative conditions, when a passive tag located on the windshield of a moving car is illuminated by the reader antenna, it is expected that the dominant phenomena are associated with the reflection from the hood of the car and with the diffractions at the surrounding metallic edges of the windshield. Figure 2 depicts the single-lane identification scenario modeled using a commercial numerical code (Feko) to estimate the multipath effects on the performance of the passive tag. The model consists of a part of the metallic car body with a planar glass windshield of thickness $t_g = 5$ mm and dielectric permittivity $\varepsilon_r = 6.5$, illuminated by a circularly

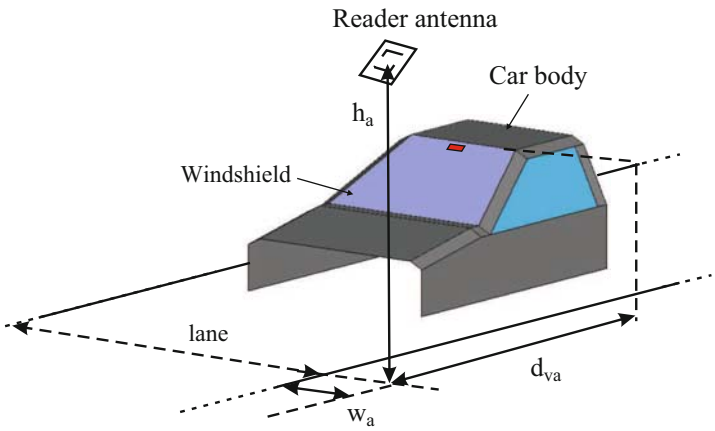


Fig. 2 Simulated single-lane vehicle identification scenario

polarized patch antenna operating at 867 MHz. To limit the computational burden, the car model is constituted of solely the parts that contribute to the reflected or diffracted field at the tag position. The reader antenna is laterally displaced with respect to the lane, at a distance of $w_a = 0.8$ m from the side and at a height of $h_a = 3$ m over the lane. The beam axis of the antenna is tilted from the horizontal by a $\theta_a = 60°$ angle, pointing towards the center of the lane. The transmitted ERP has been set to 2 W and the total electric field has been observed on a rectangular plane located inside the car parallel to the windshield at a distance of about 2 mm. Figures 3 and 4 show the simulated total electric field intensity at the observation plane for a distance $d_{va} = 3$ m. In Fig. 3 the map of the total field intensity when both the car and lane are not present is shown. It is worth noting that the field footprint is mainly dictated by the radiation pattern of the reader antenna, and that the field intensity in the most part of the windshield is quite uniform. A completely different situation is shown in Fig. 4 where the car body is present. Diffractions and reflections contributions arising at the various parts of the car body can increase the field intensity in regions near the upper portion of the windshield. In these parts the total electric field intensity may be significantly higher compared to the free space case. Moreover, due to the particular geometry configuration, these effects tend to remain rather constant also when the car is moving inside the beam footprint. In the design of the RFID system we will benefit from this phenomenon, providing that the activation threshold of the designed passive tag is smaller than the electric field predicted in the free space case.

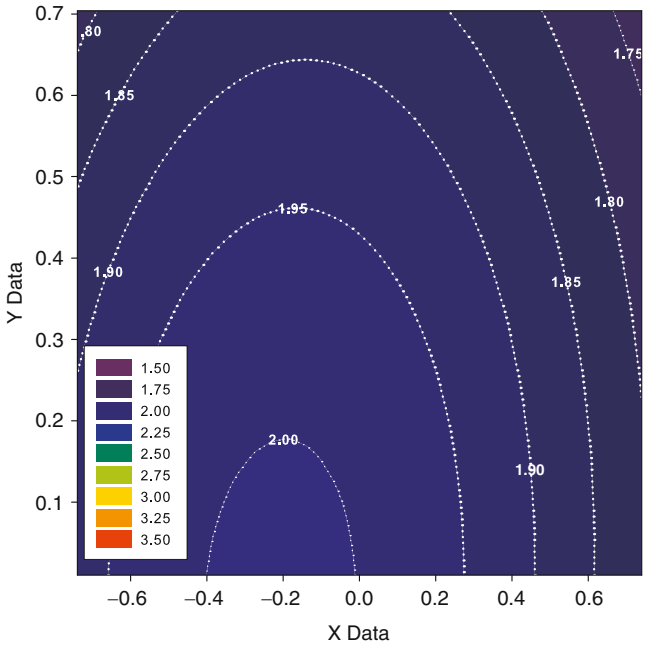


Fig. 3 Electric field intensity (V/m) at the observation plane without the car body

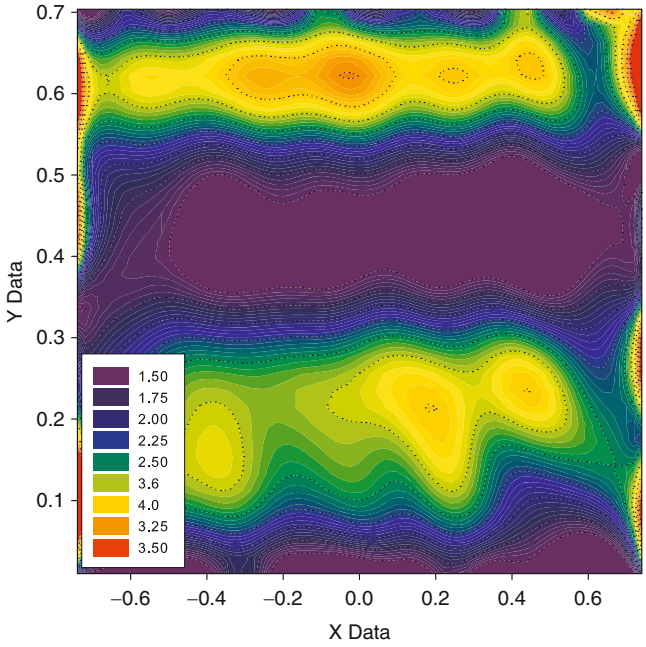


Fig. 4 Electric field intensity (V/m) at the observation plane in presence of the car body

3 Transponder Design and Analysis

As mentioned above, the integrated transponder is constituted of two different passive tags operating at HF and UHF bands allocated on a single ISO 7810 ID-1 Card. The transponder layout is shown in Fig. 5.

Concerning the HF section, due the wide availability of commercial inlays, we chose to resort to a commercial ISO 15693 Tag constituted by a copper wire deposited antenna coil connected to a Philips I-Code RFID chip. This commercial Tag is 52 mm × 41 mm thus leaving an available space of about 81 mm × 12 mm for the UHF antenna. However, to take into account the presence of the HF antenna coil while designing the UHF tag, a six-turn rectangular loop coil is considered in the numerical simulations, as illustrated in Fig. 5.

The UHF tag is constituted by a T-match dipole antenna with meandered arms connected to a UHF Gen2 Strap integrated chip (IC) kindly provided by Texas, with an estimated packaged chip input impedance of $Z_c = (11 - j63)\Omega$ and a power sensitivity $P_{th} = -13\,\text{dBm}$ at $f = 867\,\text{MHz}$. It is found in [7, 9, 10] that this type of antenna may provide a very good tradeoff between the space occupied by the tag and the desired correct conjugate impedance matching between antenna and tag IC.

The UHF antenna has been designed and optimized using a commercial electromagnetic simulation software (CST-MicroWave Studio). The main goal of the design has been to tune the antenna input impedance in such a way to be the

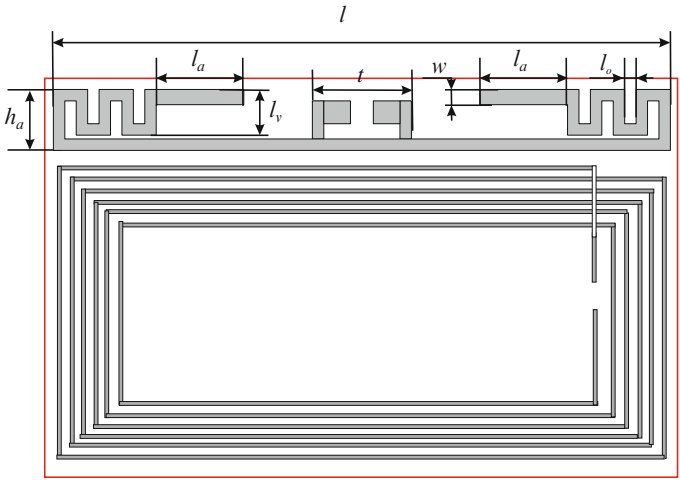


Fig. 5 UHF-HF integrated transponder layout

conjugate of the microchip characteristic impedance when the tag is in the operative environment. To this end, numerical simulations are carried out by modeling the presence of the car windshield near the tag antenna. The glass has been modeled as a planar dielectric layer of thickness $t_g = 5$ mm, with relative dielectric permittivity $\varepsilon_r = 6.5$ located at a distance $h_g = 2$ mm from the antenna.

3.1 Return Loss

A proper impedance match between the antenna and the IC is very important in order to maximize the tag performance. As is well known, the maximum reading range D_{max} of an RFID system can be computed using the Friis free-space formula as

$$D_{max} = \frac{\lambda}{4\pi}\sqrt{\frac{P_{eirp}G_r}{P_{th}}p(1-|s|^2)}, \tag{1}$$

where λ is the wavelength, P_{eirp} is the equivalent isotropic radiated power transmitted by the reader, G_r is the tag antenna gain, P_{th} is the minimum power required to activate the chip, p is the polarization loss factor, and s is the Kurokawa's power reflection coefficient [6,8]. Since the maximum reading distance is directly affected by the power reflection coefficient, in order to minimize the term $|s|^2$, a better conjugate impedance match between the antenna and the IC is required. However, it is found that [10] for a given tag antenna gain, the reading range may not be significantly increased by increasing the return loss over about -15 to -17 dB. These return loss values can be considered as the impedance match requirement in the design process in order to obtain good tag performance.

3.2 UHF Antenna Design

A prototype transponder has been designed on a single grounded low cost FR4 dielectric slab ($\varepsilon_r = 4.4$, $h = 0.8\,\text{mm}$) and has been designed to operate within the European licensed UHF band (center frequency $f = 867\,\text{MHz}$).

The design parameters of the antenna were initially set such that the total length of the dipole measured at the median line was half wavelength at $f = 867\,\text{MHz}$, provided that the dipole could fit the space available for the UHF section. A major tuning of the antenna input impedance has been then obtained by modifying the length of the vertical l_v and the horizontal l_o length of the branches of the meander line. To better select these values, once the total length of the dipole is fixed, we analyzed the behavior of the antenna input impedance for several ratios $\eta = l_v/l_o$. To this end, in Fig. 6, the magnitude of the input reflection coefficient as a function of η is reported. We can observe that increasing η corresponds to increase the resonant frequency of the dipole antenna. A good conjugate match is obtained at $f = 865\,\text{MHz}$ for $\eta = 4$. In the last step, the fine tuning to the frequency $f = 867\,\text{MHz}$ can be achieved by adjusting the length l_a of the final arms of the dipole without significantly changing the impedance matching between antenna and microchip. The final geometrical dimensions of the meandered dipole antenna are $w = 1.5\,\text{mm}$, $l = 81\,\text{mm}$, $h_a = 8\,\text{mm}$, $l_o = 1.5\,\text{mm}$, $l_a = 11\,\text{mm}$, $l_v = 6\,\text{mm}$, $l_t = 5\,\text{mm}$, $t = 13.24\,\text{mm}$. For this antenna, the $-10\,\text{dB}$ bandwidth is approximately 45 MHz, due to the low input impedance quality factor of the IC (~ 5.8).

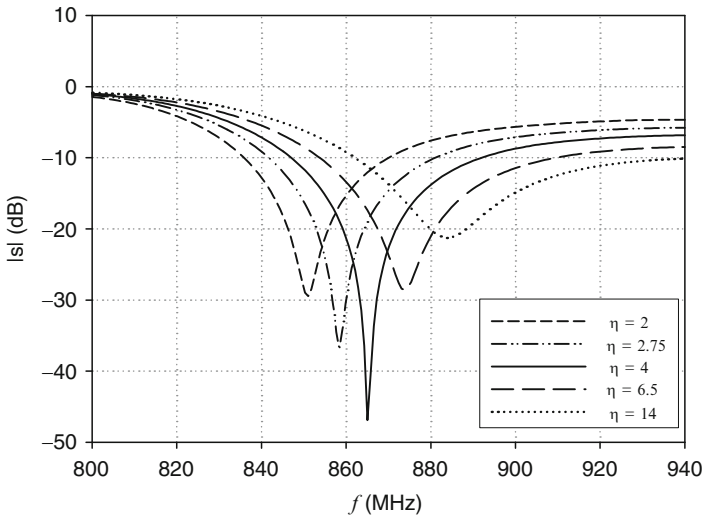


Fig. 6 Antenna input reflection coefficient

4 Experimental Results

A prototype of the integrated transponder at the European band ($f_o = 867$ MHz) with the proposed antenna has been built using FR4 ($\varepsilon_r = 4.4$, thickness = 0.8 mm) as a substrate and copper for the traces. A picture of the prototype is shown in Fig. 7. We can easily recognize the commercial ISO 15693 Tag which consists of a copper wire antenna coil deposited on a plastic flexible substrate. The reading range for the tag has been measured using a setup composed of a UHF reader and a circularly polarized antenna with gain $G_t = 9$ dBc. In order to test the maximum reading range of the integrated transponder when placed on the windshield, the reader antenna was first positioned in front of the car under test, such that the tag was illuminated directly by the reader antenna beam. Table 1 shows the measured maximum reading distance of the UHF tag of the integrated transponder when it is equipped with the HF tag. Test results have been obtained for two different types of car and two different positions of the transponder on the windshield. The two positions A and B refer to an upper central and upper lateral position of the transponder on the windshield, respectively.

It is worth noting that that the maximum reading distance is significantly higher than the one expected in free space, thus substantiating an enhancement of the field intensity near the glass–metal junction of the windshield. An uninterrupted reading capability is observed when reducing the distance as the vehicle moves. The measurements also show that the presence of the HF coil does not significantly degrade

Fig. 7 Prototype of the integrated transponder

Table 1 Maximum reading distance

Car	Position A (m)	Position B (m)
Car type 1	9.1	8.5
Car type 2	7.6	8.9

the tag performance. Preliminary tests have also been conducted in a real scenario showing that the car type 1 equipped with the proposed tag in position A can be identified at least 2 times, up to a vehicle speed of about 40 km/h.

5 Conclusions

In this paper an RFID passive system for the identification of moving vehicles within a monolane scenario for nonstop road-toll operations has been presented and analyzed. A single-lane identification scenario has been implemented to analyze the effects on the identification operations of the various reflection and diffraction phenomena originating at the car body. Simulation results have shown that, for the chosen configuration of the system, multiple path phenomena may cause an enhancement of the field intensity near the upper glass–metal junction of the windshield. This phenomenon may be used to obtain a reading zone useful for the identification of a moving vehicle approaching a road-toll system with a relatively slow speed. The system uses passive transponders operating within the European band (865–870 MHz). This latter is composed by both a UHF tag and an ISO 15693 commercial tag, arranged in two separate sections on a single ISO 7810 ID-1 card. Numerical simulation have shown an antenna gain of about 1.5 dB and a power reflection coefficient with a −10 dB bandwidth of about 45 MHz. Test results using a prototype of the integrated transponder have confirmed a good performance of the system with an uninterrupted identification capability of a tagged-car up to distances greater then the ones expected for a free space scenario.

References

1. Almanza-Ojeda DL, Hernandez-Gutierrez A, Ibarra-Manzano MA (2006) Design and implementation of a vehicular access control using RFID. Electronics and Photonics, MEP 2006. Multiconference, pp 223–225
2. Blythe P (1999) RFID for road tolling, road-use pricing and vehicle access control. RFID Technology (Ref. No. 1999/123), IEE colloquium, pp 8/1–8/16
3. Finkenzeller K (2003) RFID handbook, 2 edn. Wiley, New York
4. Foina AG, Barbin SE, Ramirez FJ (2007) A new approach for vehicle access control using active RFID tags. Microwave and optoelectronics conference, IMOC 2007. SBMO/IEEE MTT-S international, pp 90–93
5. Hirvonen M, Pesonen N, Vermesan O, Rusu C, Enoksson P (2008) Multi-system, multi-band rfid antenna: Bridging the gap between HF- and UHF-based RFID applications. Wireless technology, EuWiT 2008. European Conference, pp 346–349
6. Kurokawa K (1965) Power waves and the scattering matrix. IEEE Trans Microw Theory Tech 13(3):194–202
7. Marrocco G (2003) Gain-optimized self-resonant meandered line antennas for RFID applications. IEEE Antennas Wirel Propag lett 2:302–305

8. Nikitin PV, Rao KS, Lam SF, Pillai V, Martinez R, Heinrich H (2005) Power reflection coefficent analysis for complex impedances in RFID TSG design. IEEE Trans Microw Theory Tech 53(9):2721–2725
9. Rao KS, Nikitin PV, Lam SF (2005) Antenna design for UHF RFID tags: a review and a practical application. IEEE Trans Antennas Propag 53(12):3870–3876
10. Toccafondi A, Braconi P (2007) Compact meander line antenna for HF-UHF tag integration. In: Proceedings of IEEE AP-Symposium, Honolulu, HI

Sensor-Oriented Passive RFID

Gaetano Marrocco, Cecilia Occhiuzzi, and Francesco Amato

1 Introduction

The recent advances in Wireless Sensor Networks (WSNs) [1] for applications in mobile and time-varying environments are generating a growing attention to low-cost and low-power wireless nodes equipped with radio/sensing ability, spatially distributed to ensure a cooperative monitoring of physical or application-specific conditions and parameters. Typical fields of applications for WSNs include environmental and habitat monitoring, disaster relief [2] healthcare, inventory tracking and industrial processing monitoring, security and military surveillance, and smart spaces applications. A novel technological trend is the integration among WSNs and Radio Frequency Identification (RFID) technologies. Such a convergence of sensing and identification features may enable a wide range of heterogeneous applications, which demand a tight synergy between detection and tagging.

A new frontier is the wireless monitoring of people within Mobile Healthcare Services [3] with the purpose to reduce the hospitalization of patients, to support disaster relief, or to get an epidemic under control. An RFID system could provide real-time bio-monitoring and localization of patients inside hospitals or domestic environments, as well as in extreme conditions such as a Space Capsule. In these cases, the tag should be placed on the human body and equipped with biosensors (temperature, blood pressure, glucose content, motion) and, when activated by the reader, tag ID and bio-signals could be transferred to remote units and then stored and processed.

Up to date, several approaches have been proposed to provide RFID devices with enhanced sensing and detection capabilities. The main solutions make use of active or passive RFID transponders and Surface Acoustic Wave (SAW) devices [4]. A significant example of enhanced passive RFID system is given by the Wireless Identification Sensing Platform (WISP) project [5], which introduced the concept of ID modulation.

G. Marrocco (✉), C. Occhiuzzi, and F. Amato
University of Roma Tor Vergata, Dipartimento di Informatica Sistemi e Produzione,
Via del Politecnico, 1-00133 Rome, Italy
e-mail: marrocco@disp.uniroma2.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_26,

These devices could be *passive*, harvesting energy from the interrogating system, *semi-active* when a battery is included only to feed the sensors, or fully *active* where a local source directly feeds a microcontroller besides the transmitting radio. However, the large battery packs required for active techniques, in addition to the use of protruding antennas, could be suboptimal for some applications and additional issues have to be considered such as the compromise between a long battery life and a miniaturized design.

In passive RFIDs, together with the microchip sensitivity, the tag antenna plays a key role in the overall system performance, such as the reading range and the compatibility with the tagged object. In case of RFIDs with sensing capability, the antenna should be additionally suited to electrical and physical integration with sensing electronics.

The most challenging issues for passive RFID-sensor are:

- The design of tag antennas suited to application over (or even inside) the human body (or any other high-dielectric object). It generally produces a strong pattern distortion and reduces the tag efficiency due to energy dissipation and scattering.
- The extraction of the sensed data out of the backscattering response of a passive tag without the use of any dedicated microcontroller.

The first goal requires the tag design to directly include the presence of a human body model. Data extraction may be accomplished by consideration that the RFID tag acts as a digital device, which, when interrogated and energized by a reader, send back its own ID. Moreover, the tag may be also considered as an analog component whose strength of backscattered power is sensibly affected by the electrical feature of the antenna, which is in turn dependent on the tagged object change.

Within this scenario, this contribution is aimed at introducing a procedure to extract sensing data independently on the observation angle and on the reader-tag distance. This idea is discussed here for what concerns the sensing of discrete and continue data. The subject is approached theoretically and experimentally and corroborated by preliminary prototypes for the sensing of material changes and for the detection of human motion by means of a properly developed wearable tag.

2 Basic Definitions for RFID Systems

At the beginning of the reader-to-tag communication protocol [6], the reader first *activates* the tag, placed over a target object, by sending a continuous wave, which, on charging an internal capacitor, provides the required energy to perform actions. During this *listening mode*, the microchip (IC) exhibits an input impedance $Z_{\text{chip}} = R_{\text{chip}} + jX_{\text{chip}}$, with X_{chip} capacitive, and the antenna impedance $Z_{\text{A}} = R_{\text{A}} + jX_{\text{A}}$ should be matched to Z_{chip} $\left(Z_{\text{A}} = Z^{*}_{\text{chip}}\right)$ for maximum power transfer.

The two-way reader-tag link may be characterized by the equations of the power collected at the microchip (1) and the power backscattered by the tag toward the reader (2) and collected by this one:

$$P_{\mathrm{R}\rightarrow\mathrm{T}} = \left(\frac{\lambda_0}{4\pi d}\right)^2 G_{\mathrm{R}}(\theta,\varphi)G_{\mathrm{T}}(\theta,\varphi,\Psi)\tau(\Psi)\eta_{\mathrm{p}}P_{\mathrm{in}} \tag{1}$$

$$P_{\mathrm{R}\leftarrow\mathrm{T}} = \left(\frac{\lambda_0}{4\pi d}\right)^4 G_{\mathrm{R}}^2(\theta,\varphi)G_{\mathrm{T}}^2(\theta,\varphi,\Psi)\rho(\Psi)\eta_{\mathrm{p}}^2 P_{\mathrm{in}} \tag{2}$$

where "Ψ" generically indicates the "sensed quantity," e.g., a physical or geometrical parameter of the target which has to be monitored by the RFID platform (for instance, the dielectric permittivity, the thickness or the temperature, the motion). d is the reader-tag distance, G_{R} is the gain of the reader antenna, G_{T} is the gain of the tag's antenna when placed on the target. P_{in} is the power entering in the reader antenna, η_{p} is the polarization mismatch between the reader and the tag, τ is the power transmission coefficient of the tag, ρ is a function of the antenna impedance related to the tag's radar cross-section and to the modulation impedance Z_{mod} of the microchip to encode the low and high digital state:

$$\tau = \frac{4R_{\mathrm{chip}}R_{\mathrm{A}}(\Psi)}{\left|Z_{\mathrm{chip}} + Z_{\mathrm{A}}(\Psi)\right|^2} \tag{3}$$

$$\rho(\Psi) = \frac{4R_{\mathrm{A}}^2(\Psi)}{\left|Z_{\mathrm{mod}} + Z_{\mathrm{A}}(\Psi)\right|^2} \tag{4}$$

The tag is activated when the absorbed power exceeds the tag's microchip sensitivity threshold p_{T}, e.g., when $P_{\mathrm{R}\rightarrow\mathrm{T}} > p_{\mathrm{T}}$, and hence the maximum read distance is given, from (1), by

$$d_{\max}(\theta,\varphi) = \frac{c}{4\pi f}\sqrt{\chi\frac{EIRP_{\mathrm{R}}}{P_{\mathrm{chip}}}\tau G_{\mathrm{tag}}(\theta,\varphi)} \tag{5}$$

Both P_{in} and $P_{\mathrm{R}\leftarrow\mathrm{T}}$ are measurable quantities by the reader and they indirectly embed some physical information about the tag's features. However, the mutual reader-tag position may be generally unknown, preventing the direct use of the backscattered power to extract the sensing data.

3 Extraction of Sensing Data

A solution to extract some target's physical information out of the reader's measurements is the *ID-modulation* method [7], which is useful when the sensed quantity takes a discrete number of values or *events* $\{\Psi_1, .., \Psi_N\}$. The tag has to

be equipped with at least $\log_2 N$ microchips so that the Ψ_n event is discriminated through a particular combination of transmitted IDs according to a discrete coding. For instance, to discriminate two events a single microchip could be theoretically enough so that the first event is associated to the correct tag's response collected by the reader interrogation, while the second event to the absence of response. From a more general point of view, the tag's antenna has to be designed as a *multi-port* device, in which the ports' impedances are properly affected by the occurrence of the events [8]. An example of ID modulation is given later on in the case of wearable motion sensor.

A completely different approach may be followed to extract analog data, e.g., when the sensed parameter may vary with continuity within a given range. At this purpose, it is useful to process the ratio $P^2_{\mathrm{R}\to\mathrm{T}}/P_{\mathrm{R}\leftarrow\mathrm{T}}$ at the minimum reader's power $P^{\mathrm{to}}_{\mathrm{in}}$ activating the tag (*turn-on power*). In this case, the power $P_{\mathrm{R}\to\mathrm{T}}$ collected by the microchip will be, by definition, the microchip sensitivity p_{T} and the previous ratio becomes

$$\frac{p^2_{\mathrm{T}}}{P_{\mathrm{R}\leftarrow\mathrm{T}}} = P^{\mathrm{to}}_{\mathrm{in}}\frac{\tau^2(\Psi)}{\rho(\Psi)} \tag{6}$$

where the dependence on orientation and distance disappears. In case the tag's modulating impedance is the microchip impedance itself $\left(Z_{\mathrm{mod}} = Z_{\mathrm{chip}}\right)$, then the equation may be rewritten with respect to a reference condition $\Psi = \Psi_0$ as

$$\bar{P}(\Psi) \equiv \sqrt{\frac{P^{\mathrm{to}}_{\mathrm{in}}(\Psi)P_{\mathrm{R}\leftarrow\mathrm{T}}(\Psi)}{P^{\mathrm{to}}_{\mathrm{in}}(\Psi_0)P_{\mathrm{R}\leftarrow\mathrm{T}}(\Psi_0)}} = \frac{\left|Z_{\mathrm{chip}} + Z_{\mathrm{A}}(\Psi)\right|}{\left|Z_{\mathrm{chip}} + Z_{\mathrm{A}}(\Psi_0)\right|} \tag{7}$$

The term $\bar{P}(\Psi)$ is the geometrical average between the input threshold power and the backscattered power and is a measurable quantity, indirectly related to the variations of the target's parameter through the change of the tag impedance. It is therefore possible to numerically or experimentally develop a *measurement curve* $\bar{P} \leftrightarrow \Psi$ relating the variation of the measured response of the tag to the variation of the target, with no a-priori information about the tag-reader mutual position, regardless the presence of multi-path and of any change occurring in the nearby interacting scenario. The even complex link attenuation is in fact the same for the direct and the reverse link and hence the overall effect vanishes through the manipulation in (7).

4 Experimentations

Two experimental examples are here reported concerning the sensing of a continuous variable and the detection of discrete events. In the first case, the parameter Ψ is the material of the tagged object, while in the second case the event is any motion of the human body.

4.1 Sensing of Medium Changes

The RFID sensing technology is here applied to the wireless monitoring of the filling level h of plastic containers with both low- (sugar powder) and high-dielectric contrast (water). In these cases, the sensed quantity is the effective permittivity of the box container and, in turn, the change of the geometric shape of the target ($\Psi = h$). These examples may be representative, for instance, of the monitoring of post-surgery edema (water case) in some part of the human body and of the state of cellulites (sugar case).

The box used for the sugar-powder ($\varepsilon_s = 2.76, \sigma_s = 2.44 \cdot 10^{-2}$S/m) experiment is made of perspex ($\varepsilon_p = 2.7$, $\sigma_p \sim 0$). The tag antenna is here a conventional T-match dipole whose impedance is best matched to a NXP microchip transponder having input impedance $Z_{chip} = 15 - j135\,\Omega$ when the container is empty.

The estimation of the threshold input power in (6) was performed for three different mutual orientations $\phi_0 = \{0°, 45°, 180°\}$ between reader and tag on the horizontal plane so that $\phi_0 = 180°$ corresponds to have the box interposed between reader and tag. The sugar level is changed in the range $0 < h < 14$ cm.

The diagram in Fig. 1 gives the measurement curve computed as in (6) and (7) for the tag. It is worth noticing that the shape of the curves is not substantially dependent on the observation angle. The reader starts to detect a significant variation of the sugar only when its height exceeds 6–8 cm (corresponding to about half the antenna size). Beyond this condition, the measurement curve appears monotonic with a 3 dB dynamics.

This sensing modality could be also suited to provide on-off information about the filling state of the container so that a low measured data $\left(\bar{P}(h_S) < 0.75\right)$ reveals that the filling level is decreased below half the antenna size.

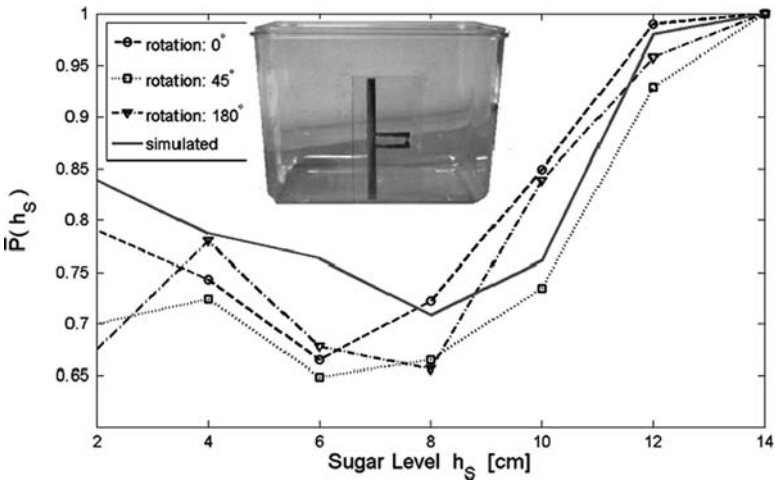


Fig. 1 RFID calibration curves for the sensing of the change in the level of sugar

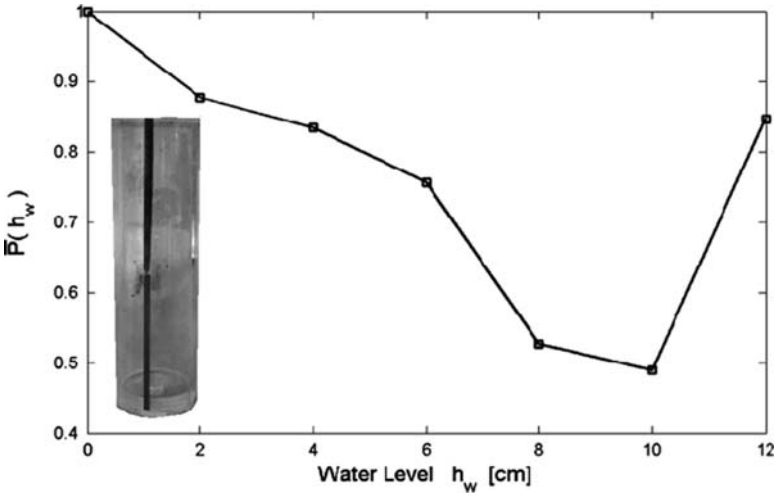


Fig. 2 RFID calibration curves for the sensing of the change in the level of water

A second test considers a smaller plastic cylindrical container (radius 22 mm, height 148 mm) with the purpose to detect the change in water level.

The tag antenna is again a simple dipole matched to the microchip when it is placed over the filled containers. The water level was changed from $h_W = 0$ cm to $h_W = 12$ cm. The sensing $\bar{P}(h_S)$ curve (Fig. 2) is nearly linear with the water level in the range 0–8 cm e.g., up to nearly half the dipole length.

4.2 Sensing of Body Motion

The human body motion may be detected through a wearable RFID tag equipped with a simple passive omnidirectional inertial switch to form a two-states (1 bit) RFID-sensor. The sensor reacts to the applied acceleration by changing its internal impedance: the main idea is to correlate the sensor state to the microchip activation with the purpose to earn information about the motion through the ID collected by the reader, according to an ID modulation paradigm [7]. More in detail, the tag will transmit its ID when it is at rest while does not transmit anything when is moving, or vice versa. The wearable tag has to minimize the interaction with the human body to prevent or at least to reduce the power absorption and therefore to achieve reasonable read distances. As evolution of the design in [9], the wearable tag considered here is the series-fed L-type patch sketched in Fig. 3. The rectangular plate has been folded around a dielectric slab of height h_S and the longest face, acting as a ground plane, is placed over the body through an optional dielectric insulator slab of thickness h_I. The polarization is linear, parallel to the antenna main-direction (x axis in the figure). Assuming that the thickness h_S of the inner dielectric is small compared with the wavelength, the radiation from the folding may be considered negligible

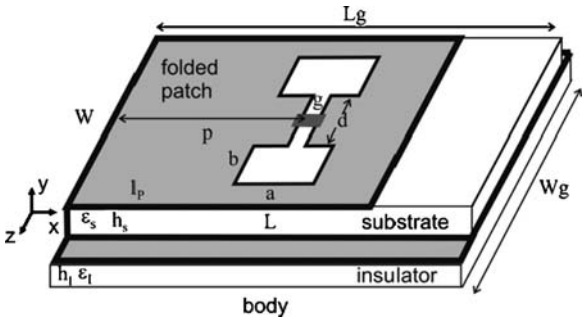


Fig. 3 Layout of the wearable tag for non-contacting sensor

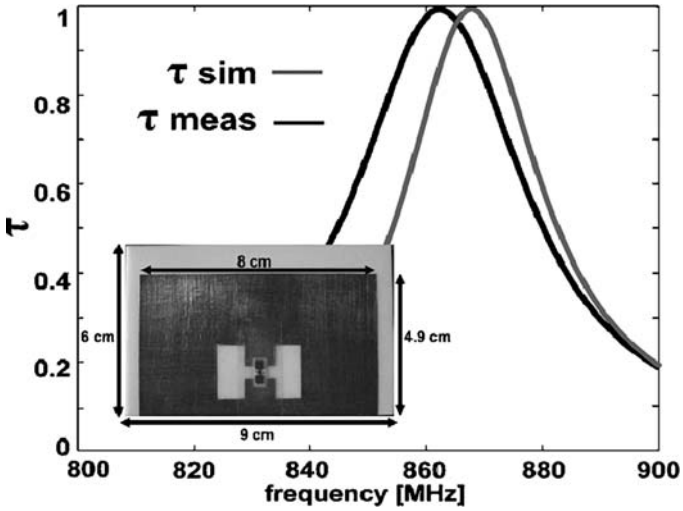


Fig. 4 Simulated and measured power transmission coefficient for the wearable tag antenna

and the gain and the matching features of the antenna mainly related to the slot and to the transmission line truncation. As for conventional patches, the increase in the horizontal size W produces a gain enhancement. Depending on the position of the tag over the body, and on the available space, it is possible to increase that dimension in order to achieve better radiation performances. The length L of the patch is chosen approximately equal to $\lambda/4$, where λ is the effective wavelength in the dielectric substrate. While the size of the slot's central gap is mainly fixed by the microchip packaging and by the eventual sensing electronics, different shape-factors and positions for the matching slot may be instead considered. In particular, the tag design may concentrate on the optimization of the only $\{a, p\}$ parameters having fixed the remainings.

A prototype matched to a microchip with input impedance $Z_{\text{chip}} = 15 - j135$ has been designed, fabricated and tested in real conditions (Fig. 4 inset).

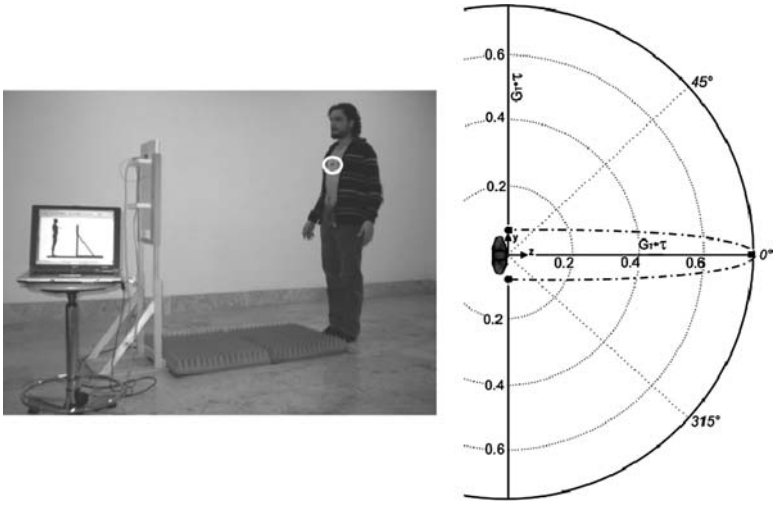


Fig. 5 (*Left*) Measurement setup comprising the short- range reader. (*Right*) Measured realized gain for the antenna placed on the torso

The measurement of impedance with the antenna attached over the leg of a volunteer gives a better than 0.9 power transmission coefficient at 869 MHz with a good agreement with simulated data (Fig. 4).

The realized gain of the tag $G_T\tau$, e.g., the radiation gain of the antenna reduced by the impedance mismatch, has been measured by using the set-up in Fig. 5 comprising a short range, remotely controlled reader CAEN A528 and a quarter-lambda patch (conventional PIFA) with maximum gain of 3.3 dB, as reader's antenna. The measured data are shown in Fig. 5 for the tag placed on the torso, evaluated along the two principals directions (y and z axis in the figure). As expected, the realized gain is maximum in front of the antenna while it is negligible in the rear side, due to the human body absorption. The experienced maximum read distance, by using the short-range reader (emitting not more than 0.5W EIRP), and tags' microchip with typical −30 dBm sensitivity, was 2.1 m. However, by using a long range reader (emitting up to 3.2 W EIRP), the maximum read distance estimated from (5) could reach 5.5 m, considerably better than those obtained in [9]. It is worth mentioning that nearly identical results are obtained when the tag is placed onto different body segments, such as the leg and the arm, thanks to the folding, which decouples the radiating part of the antenna from the body.

The integration of the motion sensor requires to properly adapt the antenna design procedures in order to achieve the modified matching condition

$$Z_A = Z^*_{chip} + Z^*_{sensor} \tag{8}$$

where $Z_{sensor} = 2.5 + j20\ \Omega$ is the sensor input impedance in the "on" state.

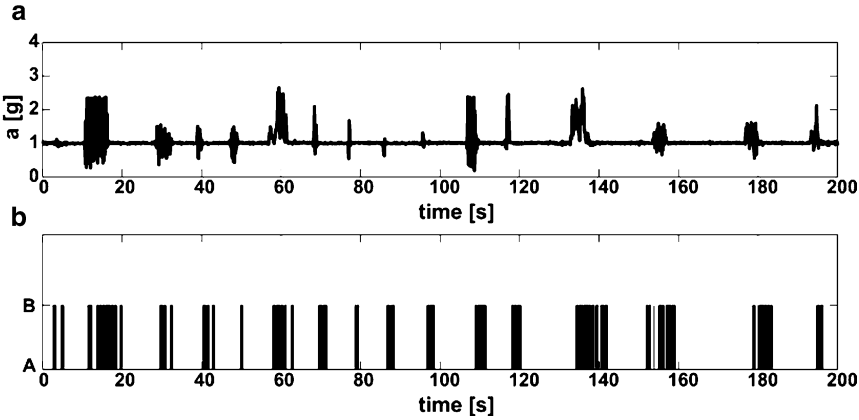


Fig. 6 Comparison of data returned by the MEMS accelerometer (**a**) with the tag response received from the RFID Motion Sensor (**b**)

A prototype of the integrated wearable sensor RFID tag has been hence experimentally evaluated in real conditions in order to verify the effective communication and sensing performances. Different accelerations, typical of common human movements, such as walk, run or downfall [10] has been applied to the tag for various periods and stored by means of the short- range RFID. The movements have been also recorded by common MEMS motion sensor (LIS302DL by STMicroelectronics). Both MEMS sensor and RFID Motion Sensor have been placed on the arm and a 16- movements sequence has been executed moving the limb randomly.

Figure 6 shows the recorded MEMS sensor data (a) and the ID-modulated data received by the reader (b), where the bars indicate the motion. A significant correlation is visible between the two motion sensors, in term of number of movements, time and duration. The RFID Motion Sensor is able to monitor every event, regardless its standing or magnitude.

5 Conclusions

Designing low-cost antennas for sensing applications is still a great challenge, especially when the human body is involved. We are just at the beginning and there are still significant possibility of methodological and technological progress to pursue in the next years. Since the power consumption of the microchip transponder is continuously reducing, according to a trend similar to the increase in the transistors' density in computer microprocessors (say the Moore Law), the concurrent research on antenna design and on smart materials, embedding also sensorial capability, will prompt new classes of distributed and massive applications, mapping the physical phenomena into a virtual reality context, accessible from anywhere.

References

1. Chong CY, Kumar S (2003) Sensor networks: evolution, opportunities and challenges. Proc IEEE 91(8):1247–1256
2. Lorincz K, Malan DJ, Fulford-Jones TRF, Nawoj A, Clavel A, Shnayder V, Mainland G, Welsh M, Moulton S (2004) Sensor networks for emergency response: challenges and opportunities. IEEE Pervasive Comput 3(4):16–23
3. Cheng-Ju L, Li L, Shi-Zong C, Chi Chen W, Chun-Huang W, Xin-Mei C (2004) Mobile healthcare service system using RFID. Proc IEEE Int Conf Networking Sensing Control 2:1014–1019
4. Reindl LM, Pohl A, Scholl G, Weigel R (2001) SAW-based radio sensor systems. IEEE Sens J 1(1):69–77
5. Sample AP, Yeager DJ, Powledge PS, Smith JR (2007) Design of a passively-powered, programmable sensing platform for UHF RFID systems. In: Proceedings of IEEE international conference on RFID, Mar 2007, pp 149–156
6. Nikitin PV, Rao KVS (2006) Theory and measurement of backscattering from RFID tags. IEEE Antennas Propag Mag 48(6):212–218
7. Smith JR, Jiang B, Roy S, Philipose M, Sundara-Rajan K, Mamishev K (2005) ID modulation: embedding sensor data in an RFID timeseries. Lect Notes Comput Sci 3727:234–246
8. Marrocco G, Mattioni L, Calabrese C (2008) Multi-port sensor RFIDs for wireless passive sensing of objects – basic theory and early results. IEEE Trans Antennas Propag 58(8)(Part 2):2691– 2702
9. Marrocco G (2007) RFID antennas for the UHF remote monitoring of Human subjects. IEEE Trans Antennas Propag 55(6):1862–1870
10. Karantonis DM, Narayanan MR, Mathie M, Lovell NH, Celler BG (2006) Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. IEEE Trans Inf Technol Biomed 10(1):156–167

Performance Evaluation of UHF RFID Tags in the Pharmaceutical Supply Chain

M. De Blasi, V. Mighali, L. Patrono, and M.L. Stefanizzi

1 Introduction

The pharmaceutical supply chain is a very complex scenario, with millions of medicines moving around the world each year. Its fragmentation, represented by the overwhelming growth of intermediate wholesalers and retailers involved in drug flow, is resulting in a decrease of transparency of the supply chain and an increase in difficulty in tracking and tracing medicines. Furthermore, the growing counterfeiting problem raises a significative threat within the supply chain system.

Of course, a serialization procedure of medicines by using standardized coding as GS1 (Global Standards 1) and EPC (Electronic Product Code), highly recommended by several international institutions (e.g., FDA, EMEA, and EFPIA), can lead to improved patient safety and to enhance the security and the efficiency of the pharmaceutical supply chain, with better traceability of medicines worldwide.

Currently, the main auto-identification technologies converge in two branches: 2-D Data Matrix [1] bar codes and Radio Frequency Identification (RFID) [2]. Furthermore, the RFID branch breaks out mainly into two segments: Ultra High Frequency (UHF) and High Frequency (HF) passive tags.

Although the Data Matrix is a very low-cost solution, there are many valid reasons for not considering it as the primary auto-identification technique. In fact, the Data Matrix is still a bar code technology, and therefore subject to the same limitations of conventional bar codes. For example, it requires line-of-sight, it cannot be written or read in bulk, it can be easily counterfeited, it can limit the speed of packaging line operations, etc. Each of these aspects leads to indirectly increase the utilization cost of this technology.

The choice between the two RFID solutions, HF or UHF, can be aided by different recent works [3], which highlight how passive UHF RFID systems provide better performance than passive HF systems.

M. De Blasi, V. Mighali, L. Patrono (✉), and M.L. Stefanizzi
Department of Innovation Engineering, University of Salento, Lecce, Italy
e-mail: mario.deblasi@unisalento.it; luigi.patrono@unisalento.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_27,

Most UHF RFID tags [4] were designed in order to exploit at one's best the characteristics of a far field coupling. In this case, the main advantages perceived are: the capability to allow multiple simultaneous reading of tags, very high read rates, and long distance between reader and tag (i.e., transponder). These characteristics lead to consider UHF band as the ideal choice for identification and tracing applications at item-level.

Unfortunately, UHF tags could occasionally encounter problems, causing performance degradation [5], when they are used in presence of materials such as liquids and metals that absorb RF energy. However, the recent introduction of near field UHF tags [6] allows mitigating these problems. These properties are making the near field RFID technology the reference technology in pharmaceutical industry, for which the item-level tagging could become of crucial interest. Nevertheless, its use is limited to the production line, where a product is scanned individually on the conveyor belt, whereas it could not be suitable in other steps of the supply chain, which require multiple readings of tags and a long-distance between tag antenna and reader antenna

The research work summarized in this paper represents a part of the results obtained by a recent research project performed at the University of Salento (Italy) in collaboration with different actors of pharmaceutical supply chain. The title of this pilot project is "Tracing and tracking pharmaceutical items using innovative EPC-aware technologies." In this paper, we analyze the advantages and disadvantages of using near field and far field UHF tags to trace pharmaceutical products at item-level in each stage of the supply chain. These tags have been tested in a controlled laboratory environment. The experimental campaigns have been performed in order to evaluate tags performance, in term of successful read rates, by varying scanning speed, tag orientation, product (i.e., item) type, package material, cases composition, and volume of purchase orders.

The rest of the paper is organized as follows. Section 2 describes the main characteristics of the test environment realized in laboratory to simulate each step of pharmaceutical supply chain. Section 3 describes main characteristics of items, cases, and purchase orders considered in the tests. In Sect. 4, experimental results are discussed. Finally, conclusions are provided in Sect. 5.

2 Description of Test Environment

The item-level traceability of drugs starts just after the packages are filled during the manufacturing process. In this step, each tagged product is scanned individually on the conveyor belt and then cased to be sent to the wholesalers. The wholesalers separate the products according to their identifiers and place them onto the shelves. Wholesalers receive orders from retailers. These orders often consist of small quantities of different products; they may contain a large number of items. The products in the orders of the retailers are picked and put into some large envelope bags that are scanned and confirmed before their distribution. Upon receipt, the retail pharmacy

scans each bag without opening it. Therefore, to simulate this scenario, a controlled laboratory environment has been created, enabling an unbiased and repeatable comparison among the technologies. Furthermore, its main components are: an items line, a cases line, and a border gate.

The items line consists of a conveyor belt whose speed can be varied in the range from 0 to 0.66 m/s, in order to guarantee real requirements of pharmaceutical manufacturing processes. For this, the following devices have been used: two Impinj Mini-Guardrail reader antennas and one Impinj Speedway UHF reader.

Similarly, the cases line consists of a conveyor belt, equipped with a line speed regulator in the range from 0 to 0.66 m/s, one Impinj Speedway UHF reader, and two roller conveyors. In the middle of the line, four small near field reader antennas (Impinj Brickyard Antenna) have been installed within a metallic tunnel.

Finally, the border gate uses a single UHF RFID reader (Impinj Speedway) and four far field UHF reader antennas.

3 Drugs Classification

The pharmaceutical market is characterized by a wide heterogeneity of drugs, which differ for several factors as medicine state (i.e., solid, liquid, gas, etc.) and material (e.g., glass, metal, plastic, etc.) of the primary package. A complete taxonomy of most popular drugs may be done taking into account these factors.

The first classification, taking into account only medicine state, splits all pharmaceutical items in four main categories:

1. Solid products: Tablets capsules, granules, etc.
2. Semi-liquid products: Creams, suppositories, etc.
3. Liquid products: Syrups, oral liquids, solutions, etc.
4. Gas products: Pressurized gasses

Another classification can be done also taking into account the material of the primary package. Plastic is the most widely used material in the pharmaceutical industry and its common applications are bottles, blister packs, and film layers. Also metal is a very popular material contained in pharmaceutical product packaging since some of its applications, such as blister packs and sachets, are more widespread than any other materials. Another common material for pharmaceutical products is glass that is very valuable especially for the liquid products. Classical applications of glass packaging in the pharmaceutical industry are bottles for liquids, ampoules, and vials. Based on the aforementioned information and discussions, Table 1 summarizes how pharmaceutical products are categorized according to their physical properties.

Note that this classification is very important to perform significative tests because different materials interact with RF waves differently. In particular, liquids cause attenuation on the RF waves by absorbing their energy, whereas metals do not let RF waves pass through by reflecting them.

Table 1 Characterization of pharmaceutical products

Product type		Package material		
		Metal	Glass	Plastic
Solid	Tablets in blister	X	–	–
	Tablets in a bottle	–	–	X
	Granules in sachets	X	–	–
	Powders in a bottle	–	X	–
Semi-liquid	Cream	X	–	X
	Syrup	–	X	–
	Single injectable solution in syringe	–	X	–
Liquid	Multiple injectable solution in syringes	–	X	–
	Oral solution	–	–	X
	Ophthalmic solution	X		–
Gas	Spray	X	–	–

4 Experimental Results

In order to carry out a performance comparison between near field and far field UHF tags applied on different drug types in each step of the supply chain, several experimental campaigns have been performed in laboratory.

In order to choose the more suitable passive UHF tags for the tests, a preliminary technological scouting has been performed. Note that for item-level tagging applications, such as the pharmaceutical one, the choice of the tags is affected by different requirements as: small size of the tag itself, compatibility with EPCglobal standard, high scanning speed, low cost, and high stress of tag label during pharmaceutical product life cycle. Three different passive UHF EPC Gen2 tags have been analyzed for the test beds:

- RSI Cube2 with a NXP Ucode G2XL chip, a passive UHF near field tag with an EPC memory equal to 96 bits
- Impinj PaperClip with a Monza2 chip, a passive UHF near field tag with an EPC memory equal to 96 bits
- Impinj ThinPropeller with a Monza2 chip, a passive UHF far field tag with an EPC memory equal to 96 bits

The antenna form and the size of these tags enable them to be easily applied on the secondary package of most pharmaceutical products. The test beds have been performed in several environment conditions in order to evaluate a performance comparison of the three different RFID tags in terms of successful read rate.

The main criteria used in the controlled test beds were the following:

- Scanning speed
- Orientation of the tag antenna with respect to the reader antenna
- Materials sensitivity
- Multiple reads of tags in presence of cases and purchase orders of the retailers

In all tests, one RFID tag has been applied on the best location of the secondary package for the production line taking into account that its position is always parallel to the reader antennas.

All test bed results, reported in the following, are accurate estimates of the successful read rate, characterized by at least a 95% confidence level whose maximum relative error is equal to 5%.

The experimental results have been split in three different subheadings, each of which takes into account a particular step of the supply chain: items test, cases test, and purchase orders test.

4.1 Items Line Testing

To evaluate the reliability of RFID on the items line, three different tests have been performed:

- Best tag test
- Scanning speed test
- Antenna orientation test

The first test aimed to discover the type of RFID tag that is characterized by the best performance for most types of drugs. In this test, optimal line conditions have been assumed: line speed equal to 0.33 m/s and tag antenna parallel to reader antenna. The experimental results obtained are reported in Fig. 1. As might be expected, under these operating conditions, the near field tag Cube2 ensures optimal performance for every drug type, even in the presence of critical materials such as liquid and metal. Likewise, the far field tag ThinPropeller obtains, in general, high performance, except in presence of considerable quantities of metal, as in bomb-spray. In this case, the estimated successful read rate is equal to 76.6%. Another interesting

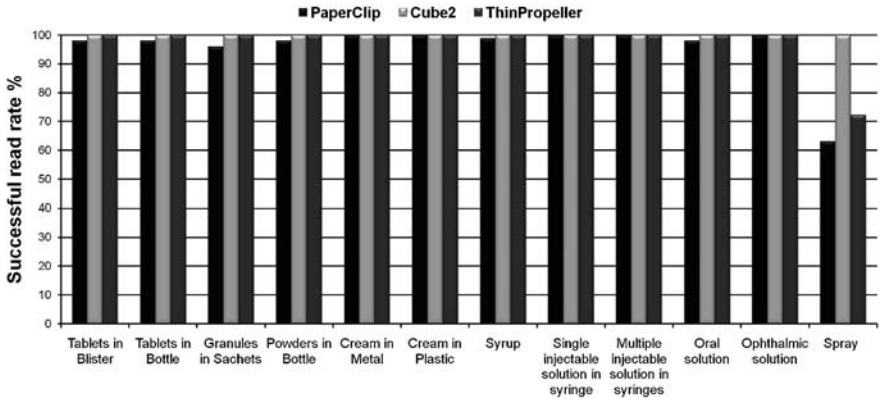


Fig. 1 Best tag test

result regards the other near field UHF tag (Paperclip): it always obtains significantly lower performance than the other tags. For the previous critical case (bomb-spray), this near field tag obtains a successful read rate equal to 63.0%, requiring, besides, an excessive closeness between tag antenna and reader antenna. The results obtained for the bomb-spray are justified by the presence of a considerable quantity of metal that, as known, do not let RF waves pass through by reflecting them.

The objective of the second test aimed to evaluate the dependence of the reading performance on the line speed. In order to simulate typical values of a real production line, the second test has been performed considering three different speeds: 0.16, 0.33, and 0.66 m/s. For this test, only the near field tag Cube2 and the far field tag ThinPropeller have been considered. The experimental results show that for both tags there is no dependence on the scanning speed.

Finally, the third test aimed to evaluate the dependence of reading performance on the orientation of the tag antenna with respect to the reader antennas. This test has been carried out by considering the worst case, i.e., a relative orientation between tag antenna and reader antenna equal to π rad. To obtain this operating condition, one of the two reader antennas was switched off and the RFID tag was applied on the face of the secondary package more far from the active reader antenna. In this test, the scanning speed has been set to 0.33 m/s and only the near field tag Cube2 and the far field tag ThinPropeller have been considered. The experimental results, reported in Fig. 2, clearly show that the performance obtained using the far field UHF tag overcomes that obtained by the best near field UHF tag. The critical conditions, according to which the far field solution is not characterized by a 100% successful read rate, have been obtained in presence of tablets in blisters (94.6%), granules in sachets (69.2%), ophthalmic solution (24.5%), and bomb-spray (7.5%). For these cases, the near field tag Cube2 has shown very poor performance, reaching values of successful read rates near to 0%.

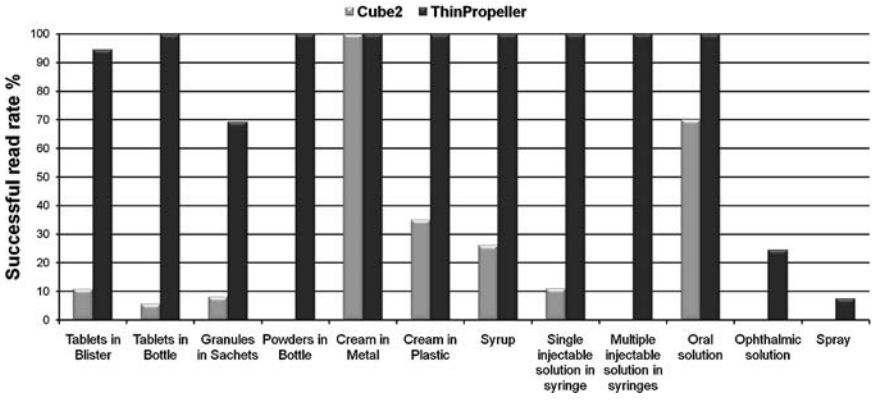


Fig. 2 Orientation test

In this case, the lower performance is due to the presence of liquid and metal that cannot be easily crossed by RF waves, inhibiting the reading of the tag antenna placed on the side of the secondary package opposite to the reader antenna.

This first set of results enables us to appreciate the effectiveness of far field UHF tags in presence of particular conditions that characterize the scanning of tags in the items line scenario. Nevertheless, performance decrease has been highlighted in some cases. In order to overcome these problems, two possible directions have been investigated: redesign of the tag antenna and strengthening of reader antennas. The first solution is a work in progress of our research group; instead, the second solution has been completely tested. In particular, two far field reader antennas have been added on the same Speedway reader obtaining a hybrid configuration of reader antennas on the items line. Then, only tests associated to previous critical conditions, i.e., presence of metal and liquid (tablets in blister, granules in sachets, ophthalmic solution, and bomb-spray), have been repeated using the new hardware setting. The experimental results reported in Fig. 3a, b, for scanning speed and antenna orientation tests respectively, have demonstrated that the use of this ad-hoc configuration of reader antennas is able to overcome the previous performance problems. In particular, Fig. 3a clearly shows that optimal performance (100% of successful read rate) can be obtained using far field UHF tags at maximum scanning speed (0.66 m/s) also in presence of considerable quantities of metal (bomb-spray). Nevertheless, Fig. 3b shows that the orientation problem is completely resolved for most drug types (e.g., tablets in blister, granules in sachets, and ophthalmic solution) and partially for the bomb-spray, obtaining a successful read rate equal to 22.0%.

4.2 Cases Line Testing

The second step of our testing campaigns aimed to evaluate the readability of multiple products at the same time, trying to analyze tags collisions problem that may occur during multiple readings of tags. Let us observe that, taking into account the variety of cases with different sizes used in the pharmaceutical sector, a square

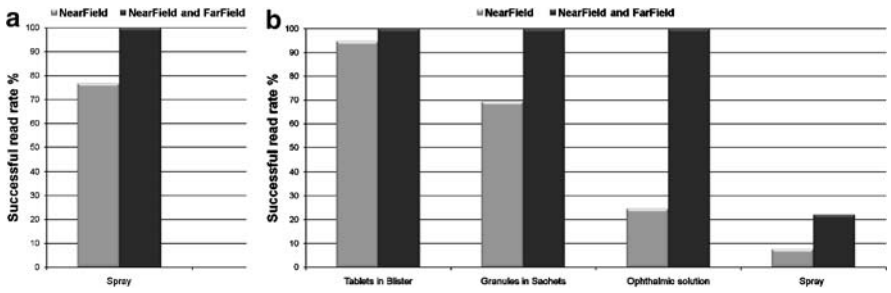


Fig. 3 (**a**) Scanning speed and (**b**) orientation test in hybrid configuration

tunnel whose width is equal to 0.60 m has been adopted. We took this choice being aware that it does not represent the optimal utilization condition for any case size since the tunnel antennas used guarantee high performance for reading distance until 0.20 m. In this test, six different product types, typical of the four categories previously defined, have been used to perform tests on homogeneous (single type of product) and heterogeneous (mix of products) cases.

Homogeneous cases have been analyzed considering two different dispositions of the items inside the case: a random disposition and an ad-hoc disposition. In the latter disposition, the items are placed in the case with their tag antennas oriented toward reader antennas, whereas in the first disposition the items are randomly positioned in the case. Figure 4a shows that the near field solution is able to guarantee a 100% successful read rate except in the presence of granules in sachets (78.8%), syrups (9%), and injectable solution (91.6%). Instead, Fig. 4b shows that the use of the far field solution is able to guarantee, for the previous critical cases, substantial performance improvements. The use of the far field tag does not reach the optimal

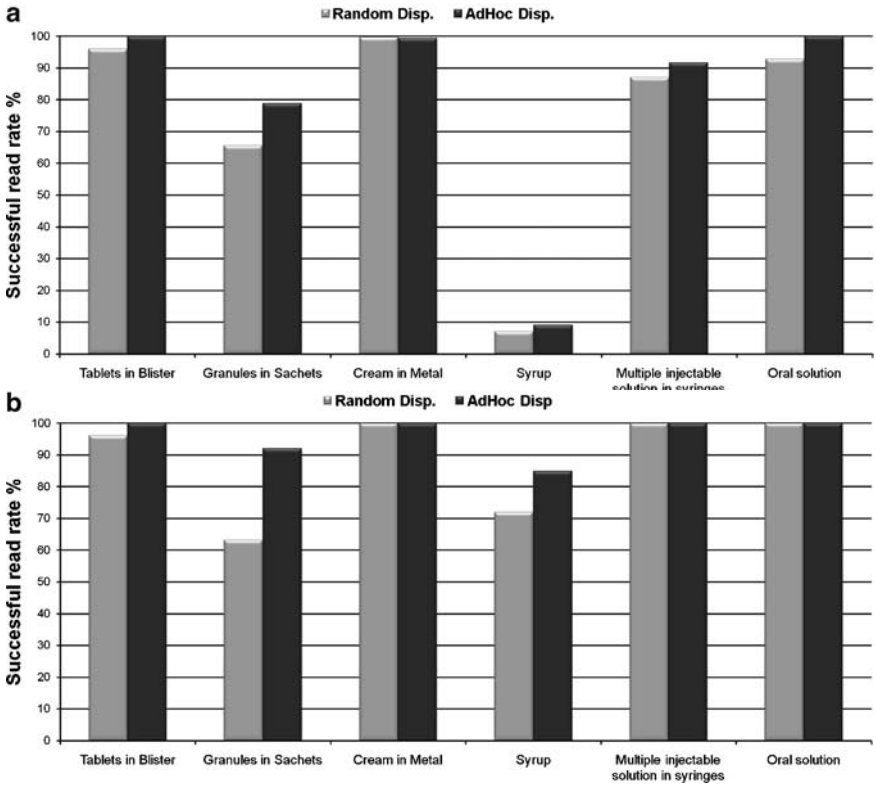


Fig. 4 Cases line test: (**a**) homogeneous cases test with the near field tag Cube2, (**b**) homogeneous cases test with the far field tag ThinPropeller

performance in presence of considerable quantities of metal (granules in aluminum sachets) and liquid (syrups). Let us observe that this performance decrease is probably caused by the nonoptimal hardware configuration.

The heterogeneous cases were prepared using 50 items and respecting three different typologies:

- MIX1: 40% solids, 40% liquids, and 20% semi-liquids
- MIX2: 50% solids, 30% liquids, and 20% semi-liquids
- MIX3: 60% solids, 20% liquids, and 20% semi-liquids

The experimental results show clearly that optimal performance (100% of successful read rate) can be obtained using both tag types.

4.3 Purchase Orders Gate Testing

To simulate the real products flows from a wholesaler to a retailer, purchase orders with different number of products have been considered. In particular, taking into account the three different typologies (MIX1, MIX2, MIX3) previously described and using only the more interesting far field tag (ThinPropeller), the performance obtained by different quantities of items in purchase orders has been evaluated. For each mix, four different tests have been performed, considering 50, 100, 150 and 200 items in each purchase order. The obtained experimental results have demonstrated that the far field solution is able to guarantee a 100% of successful read rate.

5 Conclusion

The recent introduction of near field passive UHF tags in order to improve the performance for item-level tagging in critical conditions (e.g., in presence of metals or liquids) has stimulated our research work, summarized in this chapter. In particular, several tests have been performed to analyze the advantages and disadvantage of using near field and far field UHF tags for item-level tracing of pharmaceutical products in each stage of the supply chain.

The experimental results have demonstrated that the use of far field UHF tags and the adoption of a suitable reading system configuration are able to guarantee performance better than those obtained by using near field UHF tags. In fact, the tests results have shown how a particular configuration of the reader system is able to resolve the problems of tag orientation and sensitiveness of materials, also allowing obtaining good performance with near field applications.

The conclusion of this work is that the use of passive far field UHF tags could represent the de-facto solution for item-level tracing systems in the whole supply chain. Furthermore, the obtained results lead us to assert that this solution can be easily extended to other sectors in which the item-level traceability is still an important aspect.

Acknowledgments The authors wish to thank all industrial partners (Merck Serono, FarPas, EurPack, and CTP System) of the TRUE project for their practical advices that have been fundamentals for obtaining the experimental results shown. Finally, the authors would like to thank Luca Catarinucci, researcher at the University of Salento, for his support.

References

1. International Standard ISO/IEC 16022: Information technology – Automatic identification and data capture techniques – Data Matrix bar code symbology specification
2. Finkenzeller K (2003) RFID Handbook: Fundamentals and applications in contact-less smart cards and identification, Wiley&Sons
3. Uysal DD, Emond J-P, Engels DW (2008) Evaluation of RFID performance for a pharmaceutical distribution chain: HF vs. UHF. Proceedings of 2008 IEEE international conference on RFID, The Venetian, Las Vegas, Nevada, USA, 16–17 April 2008
4. Sydanheimol L, et al. (2008) Characterization of passive UHF RFID tag performance. Antennas and Propagation Magazine, IEEE, 50(3):207–212 June 2008
5. Seshagiri Rao KV, Nikitin PV, Lam SF (2005) Antenna design for UHF RFID tags: a review and a practical application. IEEE Trans Antennas Propag 53(12):3870–3876
6. Nikitin PV, Rao KVS, Lazar S (2007) An overview of near field UHF RFID. Proceedings of IEEE international conference on RFID, Gaylord Texan Resort, Grapevine, TX, USA, 26–28 March 2007

The Benefits of RFID and EPC in the Supply Chain: Lessons from an Italian Pilot Study

Massimo Bertolini, Eleonora Bottani, Antonio Rizzi, and Andrea Volpi

1 Introduction

Radio Frequency Identification (RFID) technology and EPC Network [9] are experiencing an increasing diffusion in the logistics pipeline, where they are expected to have a major impact on the efficiency of the whole chain. Commonly quoted benefits of RFID encompass increased processes automation and labor efficiency, and better accuracy of logistics activities [1, 13].

However, improvements resulting from automated product identification may only be the tip of the overall RFID benefits, which also include new business opportunities and strategies [12]. The benefits of an extensive deployment of RFID technology within a supply chain can be represented as an iceberg (see Fig. 1). The starting point is implementing RFID technology inside a company to manage either single or multiple processes; this allows achieving, as immediate benefit, the automation of internal processes, thanks to a technology that makes possible multiple, out of sight, and completely automated identification of objects. Such benefits are also clearly tangible in terms of saved manpower for identification and checking. Conversely, several additional benefits, achievable thanks to the adoption of RFID within the entire supply chain and the sharing of related information through the Internet of things, are not immediately visible.

For instance, integration with the direct partners allows a company to automate inbound and outbound logistics (i.e., receiving and shipping operations) thanks to automated retrieval of advance shipping notifications and real-time proof of deliveries (POD) [10]. Moreover, deploying RFID along the entire supply chain and exploiting EPC Network for real-time data sharing allow increased inventory visibility. In turn, this leads to reduced inventory levels [13], availability of accurate Points of Sale (POS) data, better control of the supply chain [6], automation of administrative processes (e.g., invoicing), and improved customer service. Moreover, due to easy data capturing and real-time visibility they bring in the supply chain,

M. Bertolini, E. Bottani (✉), A. Rizzi, and A. Volpi
Department of Industrial Engineering – University of Parma,
viale G.P.Usberti 181/A, 43100 Parma, Italy
e-mail: eleonora.bottani@unipr.it; antonio.rizzi@unipr.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_28,

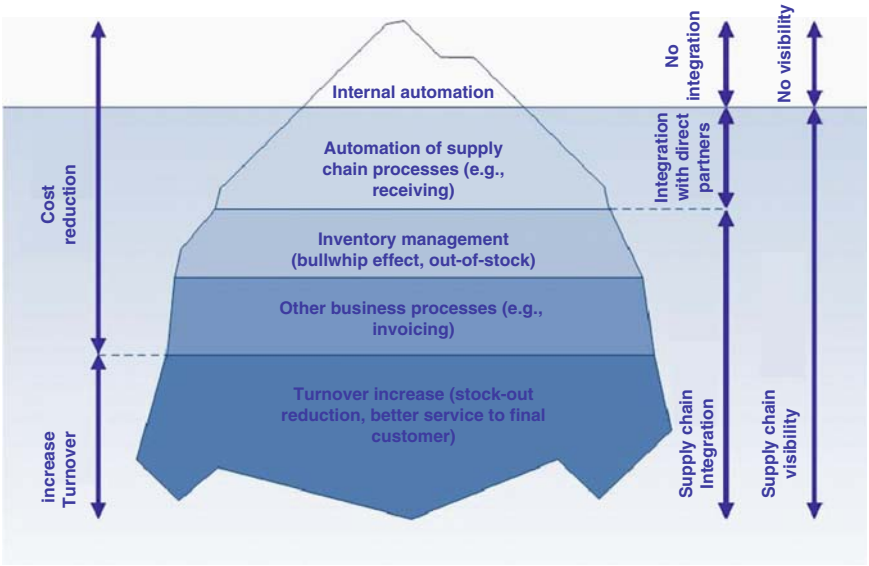


Fig. 1 The "RFID iceberg"

RFID technology and EPC Network provide companies with updated retail data, thus allowing improved sales forecasts [3]. Hence, they are also suggested as viable tools to reduce the overall supply chain inventory waste caused by the bullwhip effect (BE) [2].

The purpose of this paper is to quantify the overall benefits of RFID technology and EPC Network on the supply chain, following the concept of "RFID iceberg" previously proposed in Fig. 1. The analysis performed is grounded on the results of the RFID Logistics Pilot (RLP), a recent pilot project addressing the implementation of RFID in the Italian fast moving consumer goods (FMCG) supply chain.

The paper is organized as follows. In the next section, we provide an overview of the RLP, to explain how the results presented were derived. In Sect. 3, we describe the outcomes of the project and provide a quantitative assessment of the costs and savings resulting from RFID implementation. Section 4 proposes a feasibility study of RFID and EPC implementation. Concluding remarks and future research directions are finally presented.

2 Overview of the RFID Logistics Pilot Project

The RLP was launched at the RFID Lab of the University of Parma in June 2007, with the aim of exploiting RFID technology and innovative EPC Network tools, to enable track and trace activities and increase product flow visibility in the FMCG supply chain (www.rfidlogisticspilot.com).

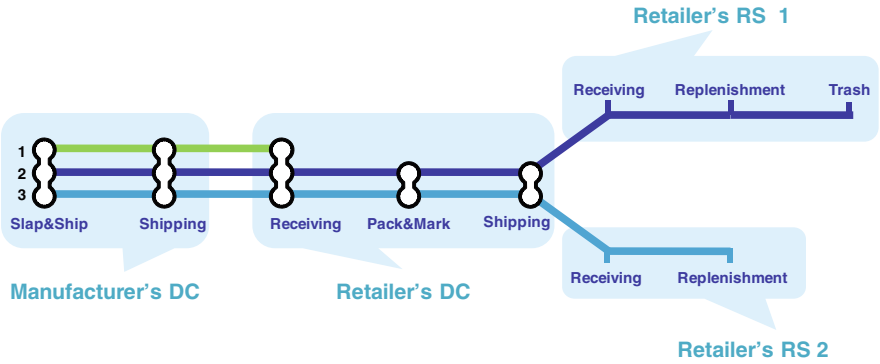


Fig. 2 The pilot supply chain and related processes (source: reprinted from [4], with permission)

The project was supported by 13 national companies and corporations, encompassing manufacturers, 3PL service providers, and distributors of FMCG, currently adhering to the RFID Lab research activities. Among others, the panel of companies include Auchan, Chiesi, Cecchi Corriere, Conad, Danone, Grandi Salumifici Italiani, Gruppo Goglio, Nestlé, Number1, Lavazza, Parmacotto, and Parmalat. While all the companies shared project costs, operational decisions, outcomes, and know-how, the deployment involved only two companies, which "lent" products and sites to the whole consortium.

The pilot supply chain involved a manufacturer's warehouse, a distributor's Distribution Centre (DC) and two distributor's Retail Stores (RS), one of which (i.e., RS1) is considered in the present analysis. Products tagged are cardboard cases manufactured by Parmacotto, an Italian market leader in pork products, included the company whole product range shipped to Auchan, a major French retailer which is particularly well grounded in Northern Italy. According to this structure, the processes investigated are shown in Fig. 2. For a detailed description of such processes and of the related EPCIS events, the reader is referred to [4]. During the RLP, more than 20,000 product cases were RFID tagged at the manufacturer's site, and coded with a SGTIN identifier. The flow of cases and pallets, these latter identified thanks to RFID tags and SSCC code aggregating SGTINs, was thus monitored throughout the supply chain, and in particular, up to the shelves of RS1. Pallets/cases data were also real-time shared between the EPCIS of supply chain partners by means of the EPC Network tools. To this extent "Accada EPCIS was adopted for capturing and querying events". All events met EPCIS standards.

3 RFID Implementation: Costs and Savings Assessment

In the following subparagraphs, we qualitatively describe the benefits resulting from RFID implementation in the pilot supply chain, following the scheme proposed in Fig. 1, and provide a detailed assessment of the corresponding costs and savings.

3.1 Internal Automation

From the internal automation perspective, it was expected that RFID technology allowed increasing efficiency and productivity of internal logistics processes of the supply chain players involved in the project. During the RLP, the improvements resulting from internal automation were assessed for: (1) shipping at the manufacturer's warehouse; and (2) replenishment of the store shelves at RS1.

As regards to process (1), savings resulting in checking for order consistency at the manufacturer shipping bays emerged as the key benefits of RFID deployment. Such savings account for about 31,000 €/year, corresponding to approx 6 h/day that can be saved during order preparation. Moreover, the use of RFID tags allows automating cases identification during picking, thus removing barcode reads. The corresponding savings account for about 11,000 €/year. The deployment of RFID tags also allows reducing mix/quantity errors during order preparation, which were estimated to account for about 23,000 €/year, corresponding to about 1 error/day that can be removed. Conversely, the removal of reads during order preparation could lead to possible errors in products prepared; such errors were monitored during the project and the corresponding costs were quantified. Moreover, from in-field measurements it emerged that the reading accuracy at case level does not reach 100% of pallets shipped, meaning that a part of the pallets shipped (approx 14.88%) still require manual checks to read missed cases. The corresponding costs, in terms of costs of employees checking for cases expected to be shipped but not read through the shipping portal, were considered as costs arising during the assessment. Finally, the implementation of RFID technology clearly involves the costs for case and pallets tagging. To quantify such cost, we hypothesized a unitary cost of tag of 0.10 €/tag. The resulting costs of pallets and case tagging account for approx 310,000 €/year. As per [5], the cost for pallets and cases tagging is totally charged to the manufacturer, while distributor's DC and RS1 are assumed to handle RFID-tagged cases and pallets exploiting all the benefits of RFID without incurring extra tagging costs. Table 1 summarizes the resulting cost/saving balance.

As far as process (2) is concerned, savings from improved efficiency of shelf replenishment operations mainly derive from automated monitoring of the quantity of items displayed on the store shelves. Items can be added to the shelf inventory when

Table 1 Costs/savings of RFID for "internal automation" – shipping at manufacturer's site

	Savings (€/year)	Costs arising (€/year)
Removal of BC reads during picking	11,229	–
Manual checks on pallets prepared	31,350	–
Removal of mix errors	23,103	–
Checks on pallet shipped	–	6,000
Increase in time required for picking	–	4,537
Pallet tagging	–	5,500
Case tagging	–	304,166.67
Total cost/saving balance	65,682	320,203.67

Table 2 Costs/savings of RFID for "internal automation" – replenishment at RS1

	Savings (€/year)	Costs arising (€/year)
Check of product availability on the shelves	19,560.00	–
Total cost/saving balance	19,560.00	–

cases of products are moved from the backroom to the shop floor, and subtracted when a bar code is scanned at the checkout counter. In the current situation, such checks approximately require 1 h/day. The corresponding savings are presented in Table 2.

3.2 Automation of Supply Chain Processes

The benefits of RFID technology become broader and involve the whole supply chain when the information about cases and pallets is shared between all players. Potential benefits on the automation of supply chain processes were quantified for: (1) receiving at the distributor's DC; (2) shipping at the distributor's DC; (3) receiving at RS1.

In general, benefits achievable during receiving/shipping refer to the possibility of reducing the manual checks to be performed on products received/shipped. Given the similarities between computational procedures followed to derive costs and benefits for the processes listed above, we describe costs and savings assessment for process (1), i.e., receiving at the distributor's DC.

Input data required are the number of cases and pallets shipped from the manufacturer's warehouse to the distributors' DC. Over a representative time, we found that 498 pallets out of 564 shipped by the manufacturer (86.54%) were fully read at case level thanks to RFID enabled receiving dock doors at the DC. EPC network made traceability data available in real-time, thus no manual operations were required. For shipments presenting pallets with missed cases, that is, cases expected to be on the shipment but whose tag was not read through the receiving gate, employees of the DC require approx only two additional minutes to complete the checks since they precisely know what cases had to be looked for; conversely, manual checks performed without the support of RFID technology and the EPC network takes about 9.5 min per shipment. Hence, considerable time savings can be achieved thanks to RFID deployment. Finally, thanks to EPC Network, transport documents could be directly retrieved, avoiding documentation errors and delivering the manufacturer automated POD. Table 3 summarizes the resulting cost/saving balance. Following a similar procedure, we also derived the cost/saving balance for processes (2) and (3). They are proposed in Table 4.

Table 3 Costs/savings of RFID for "automation of supply chain processes" – receiving at the distributor's DC

	Savings (€/year)	Costs arising (€/year)
Time required for checks during receiving	71,177	–
Uploading transport documents	85,200	–
Checking of missed RFID reads	–	15,283
Total cost/saving balance	156,377	15,283

Table 4 Costs/savings of RFID for "automation of supply chain processes" – shipping at the distributor's DC and receiving at RS1

	Savings (€/year)	Costs arising (€/year)
Shipping at the distributor's DC – total cost/saving balance	221,880.00	29,124.36
Receiving at the RS1 – total cost/saving balance	49,095.26	3,949.46

Table 5 Costs/savings of RFID for "inventory management" – safety stocks reduction for manufacturer

	Savings (€/year)	Costs arising (€/year)
Safety stocks reduction	52,790	–
Total cost/saving balance	52,790	–

3.3 Inventory Management

Thanks to complete visibility of product flow throughout the supply chain, all players involved in the project are expected to benefit from improved demand forecasting, which leads to reducing the BE. This point is particularly critical from a manufacturer perspective, since nowadays its forecasts are driven by retailer's orders rather than from POS data. From a quantitative perspective, the manufacturer could benefit from visibility in terms of reduction in demand variance and thus in safety stock levels. This effect has thus been quantified for the manufacturer.

In the pilot supply chain, the manufacturer receives orders every 2 days, which corresponds to the current visibility of product flow. The resulting average safety stock level accounts for 198 pallets. Thanks to the implementation of RFID and EPC, the manufacturer achieves complete daily visibility of flows. The resulting decrease in safety stocks, computed according to [7], accounts for approx 53,000 €/year, as reported in Table 5.

3.4 Automation of Other Processes

During the pilot study, invoicing emerged as a further process that can be automated thanks to RFID and EPC implementation. In particular, invoices can be

automatically generated based on RFID reads of product shipped, avoiding errors and misalignments. As a result, the cost of invoicing and of the errors rectification can be dramatically reduced. All supply chain players can benefit from the resulting economical savings; consequently, they were computed for manufacturer, DC and RS1.

It should be first pointed out that, to achieve the economical benefits of automated invoicing, cases/pallets shipped from the manufacturer's site should be all read in the retailer supply chain, either during receiving or at least at the store trash compactor. It can be seen from Fig. 2 that six read points are located in the retailer supply chain (i.e., from receiving at the DC to trash at RS1). In-field measurements allowed assessing that all cases/pallets shipped from the manufacturer have at least two or more reads in the retailer supply chain. In other words, sooner or later a case entering the retailer supply chain is RFID read and could be consequently invoiced. The shift from manual to automated invoicing has a dramatic impact on labor required to handle administrative flows and to manage misalignments, both for the manufacturer and the retailer.

The cost/saving balance is presented in Table 6 for all players considered. To compute savings related to misalignments, it has been taken into account that, in the pilot supply chain, mix, quantity, or document error approximately affect 12% of invoices emitted.

3.5 Stock-Out Reduction

The plain visibility of product flows throughout the supply chain allows to real-time monitor shelf availability of items at the retail store. By monitoring the throughput time of products and intervening when delays are identified, stock-out can be substantially reduced, leading to an increase in turnover. This has been quantified for the manufacturer.

On the basis of the current percentage of products experiencing a stock-out at retail stores (i.e., approx 7%, according to [8]), on the turnover of RS1, and on the improvements observed by Wal Mart and the university of Arkansas in shelf availability thanks to RFID deployment at case level during the project [11], the resulting economical benefits were estimated to account for approx 295,000 €/year, as reported in Table 7.

Table 6 Costs/savings of RFID for "automation of other processes" – invoicing for manufacturer, distributor's DC and RS1

	Savings (€/year)	Costs arising (€/year)
Manufacturer – total cost/saving balance	64,875.00	7,500.00
Distributor's DC – total cost/saving balance	135,219.74	29,000
RS1 – total cost/saving balance	23,124.35	3,000.00

Table 7 Costs/savings of RFID for "turnover increase" for manufacturer

	Savings (€/year)	Costs arising (€/year)
Stock-out reduction	295,312.50	–
Total cost/saving balance	295,312.50	–

Table 8 Feasibility study of RFID investment for manufacturer, distributor's DC and RS1

	Cost/saving balance (€/year)	Investments (€)	NPV (€)	PBP (years)	IRR (%)	ROI (%)
Manufacturer	150,955	226,600	274,684	1.9	48.00%	39.53%
Distributor's DC	440,070	225,500	1,101,576	0.7	137.60%	150.83%
Retail store	84,830	113,300	164,067	1.8	51.20%	46.68%

4 Feasibility Study and Discussion

On the basis of costs and savings computed in the previous section, a detailed feasibility study was performed. To this extent, the investments required in each of the above processes were also computed, taking into account the investments made in RFID hardware, networking servers, software and integration services, also encompassing EPCglobal subscription fees. Specific amortization periods and rates were hypothesized for the investment categories considered.

As a final result, net present value (NPV), payback period (PBP), internal rate of return (IRR), and return on investment (ROI) of the RFID implementation were quantified over a 5-year period, and considering a 5% interest rate. The corresponding results are proposed in Table 8 for all echelons considered.

As a first outcome, it can be seen from Table 8 that all supply chain players benefit from positive revenues of the RFID investment over the time horizon considered. In the case of the manufacturer, the NPV accounts for about 270,000 €, and the investment is paid back in approx 2 years. However, a more detailed analysis has shown that this result is substantially affected by the possibility of reducing product stock-outs at store level, which leads to considerable savings (about 295,000 €/year). In the case such savings are not taken into account, the RFID investment for the manufacturer turns out to be highly unprofitable, as the NPV at the end of the 5-year period accounts for −811,000 € and the investment is not paid back over that time horizon. This result suggests that, in the case of the manufacturer, the deployment of RFID technology for internal efficiency, and automation of supply chain processes does not allow repaying the required investments.

In other words, from a manufacturer perspective, RFID costs cannot be balanced only by labor and accuracy improvements in internal processes. Conversely, RFID at case level becomes economically sustainable only when it enables on shelf availability, thus impacting on manufacturer revenues.

From outcomes in Table 8, the retailer DC emerges as the supply chain echelon which benefits from the highest cost/saving balance, leading to a very high NPV of the RFID investment (approx 1 million €). For that player, receiving, shipping,

and invoicing are the processes that benefits from the highest savings from RFID implementation. In the case of the RS, the profitability of RFID investment is clearly lower, accounting for about 160,000 €, due to the lower flow of items handled by the RS compared with the DC. About half the annual cost/saving balance of the RS is generated by the automation of receiving operations, and again from visibility due to the Internet of things.

5 Conclusions

It is recognized in literature that the implementation of RFID technology and EPC system in the FMCG supply chain leads to significant economical advantages, ranging from reducing labor costs, up to reducing stock-out at retail store and safety stock levels, according to the concept of "RFID iceberg." This paper has provided a detailed investigation of such aspects, which a particular focus on "invisible" benefits (i.e., the basis of the iceberg), which result from availability of real-time product data throughout the supply chain, as a consequence of a wide deployment of RFID technology.

It emerges from the results of our study that the "invisible" savings from RFID technology allow all supply chain players to benefit from positive revenues from the RFID investment. Conversely, without such benefits, the RFID investment of the manufacturer generates a negative cost/saving balance, and has no potential to be profitable. The above result also confirms that most of the benefits of RFID derive from a wide deployment of the technology along the supply chain and from the complete integration of supply chain partners.

Starting from the results of this study, our ongoing research is devoted to extend the analysis carried out in two main ways. First, the study will be extended by examining a supply network, including several manufacturers and retailers, and also involving 3PL service providers, to assess the shortcomings of RFID implementation in that context. Outcomes of this project are expected to provide the basis for a wide use of RFID technology in the Italian FMCG supply chain. As a second point, some aspects of this study, which have potential to substantially affect the profitability of RFID investment, will be investigated in greater detail in further works. Such aspects include: (i) the use of RFID technology to monitor products availability during sales promotion, (ii) the causes of out-of-stocks of products on the store shelves and the impact of RFID to rectify them; and (iii) the impact of RFID on some specific processes (namely, shelf replenishment and receiving at the DC), for which a great potential for cost saving emerged in this study.

References

1. Agarwal V (2001) Assessing the benefits of Auto-ID technology in the consumer goods industry. Cambridge University, Auto-ID Center. http://www.autoidlabs.org.uk. Accessed July 2005
2. Alinean (2006) Shrinking the supply chain expands the return: the ROI of RFID in the supply chain. http://www.alinean.com. Accessed January 2007
3. Asif Z, Mandviwalla M (2005) Integrating the supply chain with RFID: a technical and business analysis. Commun Assoc Infor Sys 15:393–427
4. Bottani E, Bertolini M, Montanari R, Volpi A (2009) RFID enabled business intelligence modules for supply chain optimization. International Journal of RF Technologies: Research and Applications 1(4):253–278
5. Bottani E, Rizzi A (2008) Economical assessment of the impact of RFID technology and EPC system on the Fast Moving Consumer Goods supply chain. Int J Prod Econ 112(2):548–569
6. Bushnell R (2000) RFID's wide range of possibilities. Mod Mat Hand 55(1):37
7. Chen F, Drezner Z, Ryan JK, Simchi-Levi D (2000) Quantifying the bullwhip effect in a simple supply chain: the impact of forecasting, lead time, and information. Manage Sci 46(3):436–443
8. ECR (2004) Optimal shelf availability. http://www.ecrnet.org/04-publications/blue_books/pub_2003_osa_blue_book.pdf. Accessed July 2009
9. EPC Global (2004) The EPCglobal Network™: overview of design, benefits & security. http://www.epcglobalinc.org. Accessed July 2006
10. Fernie J (1994) Quick response: an international perspective. Int J Phys Distr Log Manage 24(6):38–46
11. Hardgrave BC, Waller M, Miller R (2007) Does RFID reduce out of stocks? A preliminary analysis. http://itrc.uark.edu. Accessed June 2007
12. Krotov V (2008) RFID: thinking outside of the supply chain. http://www.cio.com/article/174108/RFID_Thinking_Outside_of_the_ Supply_Chain. Accessed June 2009
13. Prater E, Frazier GV, Reyes, PM (2005) Future impacts of RFID on e-supply chains in grocery retailing. Supply Chain Manage 10(2):134–142

RFID Data Analytics in Apparel Retail

Frédéric Thiesse and Jasser Al-Kassab

1 Introduction

The growing interest in the use and application of Radio Frequency Identification (RFID) on the part of the retail industry in recent years has sparked an intensive debate in academia and practice on the benefits to be expected. A large number of white papers, articles in trade journals, and research contributions have discussed the impact of RFID on supply chain performance (Ngai et al., 2008) [3,4]. The majority of prior works has so far concentrated on operational efficiency gains from RFID-based process automation. In contrast, our research is motivated by the question, to what extent retail companies can draw benefits beyond simple efficiency gains from the analysis of large amounts of RFID data.

We consider a European retailer, who operates more than 100 department stores. The company offers a wide assortment of products with apparel and footwear, accounting for the majority of the turnover. In order to evaluate the technology's potential in the apparel retail industry, the retailer implemented an RFID infrastructure in the menswear department in one of his department stores with a total of 22,000 ft^2 in size.

Figure 1 gives a schematic overview of the RFID installation. All products to be sold in the store are shipped from a nearby distribution center (DC). Items are tagged manually on the item-level in the DC and transported in trucks to the store, where they are read at the incoming gates. Goods intended for immediate sale are sent via the freight elevators to the menswear department, where they leave the back store and enter the front store via a transition gate. On the average, 30,000 individual items equipped with EPC compatible transponder labels are constantly available to customers. RFID readers are installed at the escalators and elevators, on the

F. Thiesse (✉)
Institute of Technology Management, University of St. Gallen, St. Gallen, Switzerland
e-mail: frederic.thiesse@unisg.ch

J. Al-Kassab
SAP Research CEC St. Gallen, SAP (Switzerland) Inc. & Institute of Technology Management, University of St. Gallen, St. Gallen, Switzerland
e-mail: jasser.al-kassab@sap.com, jasser.al-kassab@unisg.ch

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_29,

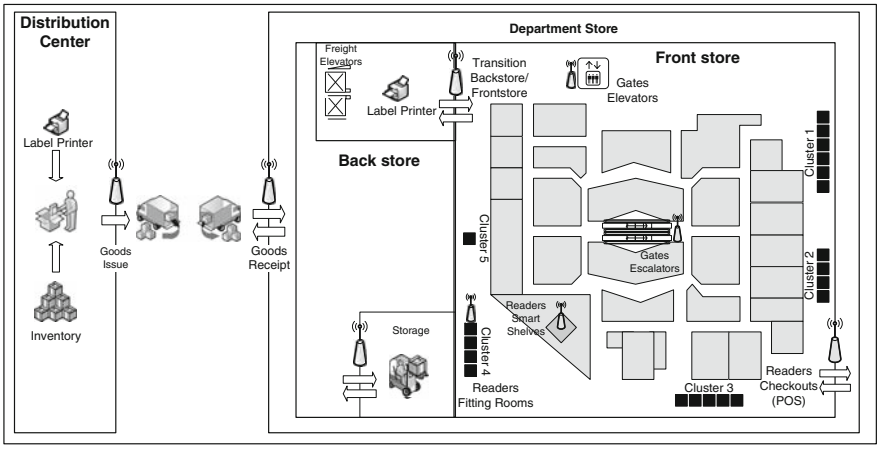


Fig. 1 Schematic overview of the RFID installation

gateways between the sales floor and the backroom, on several shelves, and in all 20 fitting rooms. In total, the infrastructure includes more than 50 RFID readers and more than 200 antennae.

2 Data Analyses

The objective of our study was to investigate the value of RFID data in retail beyond efficiency gains by means of simple process acceleration. With the help of several illustrative examples, we show how a company can combine RFID data with traditional data sources in order to generate many novel performance metrics that create unprecedented insights into in-store process execution and customer behavior on the sales floor. A summary of the necessary data analyses procedures and the resulting metrics is given in Fig. 2.

In the following paragraphs we will briefly describe theses analyses. We illustrate how the retailer under study made use of RFID data to calculate several performance metrics based on raw, filtered, and enriched RFID data to gain insights into his internal processes and customer perceptions of his store assortment.

2.1 RFID Reader Performance

Analyses of RFID raw data alone do not provide much business value but help the company to monitor the performance of their data collection infrastructure and the quality of the generated data. By comparing POS data with RFID data from the

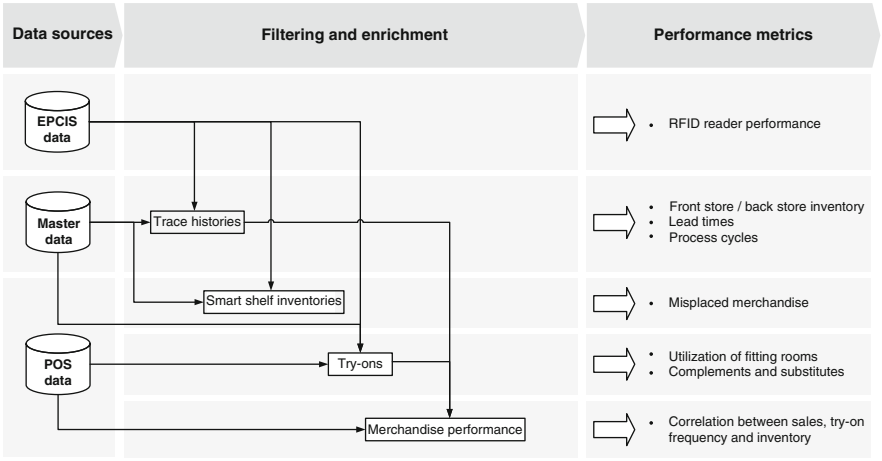


Fig. 2 Overview of RFID data analyses procedures

check-out reader, for instance, it is possible to investigate the difference between the two data collection procedures, which allows for drawing conclusions on the fraction of undetected products. Other examples are periodic tests of the amount of data generated by individual readers and antennae.

2.2 Front Store/Back Store Inventory

Using trace histories, the front store and back store inventory of items can be estimated in real-time or reconstructed for a given time and date. For example, a list can be generated for items on product group level, for each day of the year and plotted as a graph. By providing more visibility regarding the distribution of the department's inventory stock between front store and back store, this analysis supports the retailer in reducing the number of out-of-shelf but in stock situations, which are responsible for lost sales and reduced customer satisfaction.

2.3 Lead Times

Trace histories can also be aggregated on the product and on the category level, to investigate lead times with regard to different lifecycle stages of a given product. Lead time analyses support the retailer in improving his category management and his process efficiency. Articles that spend lots of time on the sales floor until they are sold, for example, could be removed from the store's assortment.

2.4 Process Cycles

In regards to process execution, the retail company wanted to assess the efficiency of their in-store processes. They therefore analyzed the event data and looked for certain patterns, such as loops, or products that were taken very often to the fitting room or into the storage, or which moved between different floors. For that purpose, trace histories were aggregated in order to count the number of an item's appearances at different read points, such as fitting room readers, readers between front store and back store, elevator and escalator exits, storage, and checkouts.

2.5 Misplaced Merchandise

Misplaced merchandise situations are mainly caused by customers not returning the items they tried on to the original merchandise fixture. They would either leave the items in the fitting rooms or on nearby shelves. In order to detect misplacements, inventory data from the smart shelves were searched for items that were not supposed to be on the shelf. In a second step, these misplaced items were grouped according to their affiliation with their department and the duration of the apparent misplacements. The retailer thus gains a picture of the cleaning and tidying processes on the sales floor.

2.6 Utilization of Fitting Rooms

In order to analyze the utilization of fitting rooms, try-on data were counted and grouped into clusters. With the information provided by this analysis, the retailer under study could reassess the trade-off between the total number of fitting rooms on the shop floor, and within one cluster – with a surface of 16 ft^2 each – against more sales floor space.

2.7 Complements and Substitutes

In order to optimize category management and to learn about customer's fitting room visits, the retailer under study was interested in the article groups that the customer's perceived as substitutes or complements. For this purpose, try-on data were grouped and joined with themselves to investigate items from two particular product categories were taken together to the fitting room. With this information, the company can systematically measure the consumer perception of its products, which can help to improve the category management and to identify cross-selling potential, for example.

2.8 Correlation Between Sales, Try-on Frequency, and Inventory

During the data filtering and enrichment process, sales information, try-on information, and inventory level information for each article and read event was merged into one data table. As such, these data can be analyzed using logistical regression models in order to identify a level of inventory on the front store that would be optimal in terms of number of try-ons and sales, and eventually customer satisfaction. With these results, the retailer under study can systematically approach the optimal level of articles on the front store, thus increasing sales and customer satisfaction, and at the same time decreasing the capital costs of items on the sales floor. For the optimization of his category management, the number of try-ons and sales for each article group were correlated, resulting in a try-ons/sales ratio. With this information, the retailer could compare articles within one article group by brand, supplier, color, and size, for example, thus optimizing his product range by removing articles from the assortment that were often tried on but rarely bought.

3 Outlook

The case example indicates how retailers as well as other companies might use RFID in the not-too-distant future beyond simple process acceleration. The previously described analyses allow for cost savings through process improvements (e.g., labor costs), the specific benefits of the apparel retail industry, such as additional revenue from RFID consumer applications (e.g., recommender systems, installed in fitting rooms), improved shop floor layout, optimized product assortment, inventory levels, management of sales personnel, category management, reordering, purchasing, shelf replenishment, and inventory level controls.

While the value of these data analyses seems evident and compelling at first, the construction of a cause-and-effect chain between RFID investments and an increase in process performance is not a trivial task. In our case example, the value of continuously analyzing data collected by the RFID infrastructure was affirmed by virtually all involved retail managers. However, a general answer on how to make use of this information to generate quantifiable impacts on performance could not be given so far [1]. This finding leads us to the conjecture that RFID impacts on the level of management processes depend on the existence of a company's individual capabilities to translate RFID data into value.

References

1. Al-Kassab J, Mahmoud N, Thiesse F, Fleisch E (2009) A cost-benefit calculator for RFID implementations in the apparel retail industry. 15th Americas Conference on Information Systems, San Francisco, CA
2. Ngai EWT, Moon KKL, Riggins FJ, Candace YY (2008) RFID research: An academic literature review (1995–2005) and future research directions. International Journal of Production Economics, 112(2):510–520

3. Sellitto C, Burgess S, Hawking P (2007) Information quality attributes associated with RFID-derived benefits in the retail supply chain. Int J Retail Distrib Manag 35(1):69–87
4. Thiesse F, Condea C (2009) RFID data sharing in supply chains: what is the value of the EPC network? Int J Electron Bus 7(1):21–43

Towards the Future Internet of Sensors

Olivier Alphand, Andrzej Duda, Martin Heusse, Benoît Ponsard, Franck Rousseau, and Fabrice Theoleyre

1 Introduction

In this position paper, we propose a new view on the integration of wireless *sensor and actuator networks* (SANET) in the Internet. Such networks become increasingly important for gathering various physical measures and acting upon objects in the physical world. Usually, they support a single application and use sophisticated protocols and wireless technologies optimized for energy consumption. Significant recent research effort aims at making them more generic and interconnecting with end-hosts on the Internet.

One approach to achieve this objective is to provide end-to-end IP connectivity to all sensor/actuator nodes. IETF promotes the idea of using the standard Internet protocols: the 6LowPAN working group proposes to adapt IPv6 to operate in a sensor network [7]. As the IPv6 addresses are long, the group proposes to set up a gateway that translates 128-bit addresses into 16 bits. Moreover, header compression techniques are further used to reduce the overhead. In addition to these techniques, the IETF ROLL working group proposes to develop a routing protocol over low power and lossy networks [5]. Several industrial developments (ArchRock, Dust Networks, Cisco) and alliances (IPSO Alliance – IP for Smart Objects) follow the same track. However, we think that it is also interesting to explore other approaches. As SANET nodes are very different from IP routers, we would like to find another paradigm that provides a higher-level, data-oriented view on the physical world in which this kind of networks operate.

At the same time, we observe that such a data-centric view perfectly corresponds to the new approach of the Future Global Internet: several initiatives proposed to base the new generation of the communication infrastructure on the *data or content dissemination* paradigm. We believe that such a model is particularly suitable for SANETs and allows us to begin experimentation with the new vision of the Future Internet in the context not so ossified as the current Internet. At the same time, it may

O. Alphand (✉), A. Duda, M. Heusse, B. Ponsard, F. Rousseau, and F. Theoleyre
Grenoble Informatics Laboratory, University of Grenoble, Grenoble, France
e-mail: Olivier.Alphand@imag.fr

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_30,

help gaining insight into new solutions to specific problems of SANETs. Our paper presents the initial ideas on the data-centric view for integrating SANETs into the Internet, discusses several related issues, and proposes a research agenda for future activities.

The rest of the paper is organized as follows. We discuss SANETs and their integration in the current Internet in Sect. 2. Then, we present our view of the data-centric approach to the future Internet of Sensors in Sect. 3. Section 4 briefly discusses more detailed issues of a research agenda for achieving this objective. Section 5 summarizes the related work and Sect. 6 concludes the paper.

2 SANET Networks and the Internet

Sensors and actuators are increasingly used in many applications for gathering information about the physical world or controlling it. We can cite industrial applications in which sensors measure operation parameters for detecting abnormal conditions and controlling affected systems. Smart buildings use sensors for controlling temperature, light intensity, humidity, or detecting a fire. Multimedia sensors are able to capture images or video streams for intrusion or feature detection, tracking objects, etc. Sensor networks communicate by means of cheap radio, which significantly reduces the deployment cost. In SANETs, some nodes have also the ability to act on their environments, for instance, they can move objects as well as activate or control some devices. Sensor nodes face many challenges in networking, embedded systems, databases, and hardware design because they are highly constrained: their processing power is limited, they have small memory, and they are battery operated with the requirement of a long life. Their internal design, protocols at network, and MAC layers, as well as application operation need to satisfy all these stringent constraints and provide a suitable tradeoff between small imprint, efficiency, and energy consumption.

Sensor networks have become a challenging hot research topic and one aspect of the current research is their integration with the Internet. IETF adopted a conservative view of applying the standard Internet protocols to SANETs thus providing the end-to-end IP connectivity to all sensor nodes. Pushing IPv6 to sensor nodes seen as end systems raises several issues. First, using the unchanged IPv6 results in too much overhead, so the version for sensor networks needs to use a shortened header as mentioned above and this modification means that there is no pure end-to-end connectivity, because packets need to be translated by a gateway acting as a sort of a network address translation (NAT) box. Second, short 16 bits addresses may be not sufficient for large-scale sensor networks (if sensor nodes form a “dust,” thousands of nodes need to be addressed). Furthermore, perhaps there is no need for assigning an IP address to each sensor node, because better views on the information provided by sensors are possible. The analogy that we take to explain this issue is a computer composed of a CPU, memory, network card, an external disk, and a graphic card. Why do not we assign an IP address to the graphic card? It contains a processor,

some memory and we could provide IP connectivity to each computer component. Simply, we do not need such a feature, because the view provided by the computer to other end-systems is different and does not include communicating with a graphic card. The internal bus interconnects these components, they have various functions, and you do not need to extend IP connectivity to them. At the same time, the computer provides a kind of virtual view on all its components and it is viewed by other end-systems connected to the Internet as a unique entity.

Sensor networks are somehow similar – they need to provide the information gathered by sensors, but they can do it independently of considering each sensor node as an IP end-system. Furthermore, the stringent constraints of sensor nodes call for optimized MAC and routing protocols, as well as a cross layer approach for efficient interaction of their operation. The ROLL working group that proposes a "routing over low power and lossy networks protocol" has begun to address this issue [5]. In fact, SANETs are intrinsically different from the global Internet, because they operate according to a different paradigm. SANET nodes sleep almost all the time for saving energy and wake up to transmit low volumes of data, while the Internet supports high bandwidth planetary communication infrastructure. Imagine that many Internet hosts want to communicate with a single sensor node, this may lead to the performance bottleneck in the SANET. The situation of SANETs is similar to residential networks in which people do not run Web servers, but rather upload content to Web servers in ISP networks – content producers are not the same as content serving hosts. A special gateway between a SANET and the Internet may provide an abstract view on data gathered by sensor nodes, serve them for many destinations, and eliminate possible performance bottlenecks.

Note also that the sensor network community has already considered this architectural issue and converged to the conclusion that SANETs need to be seen as a Layer 2. Polastre et al. state that *"wireless sensor networks would benefit from a unifying abstraction (or "narrow waist" in architectural terms), and that this abstraction should be closer to the link level than the network level"* [10].

3 Future Internet of Sensors

We start by considering a new vision of a communication substrate that a SANET network needs to provide to Internet end-systems. We can adopt a similar approach as the Internet at its beginning that aimed at providing a minimal scalable interconnection layer with a horizontal interface over different transmission technologies and enabling new communication applications. The socket layer to TCP/IP contributed to the initial success of the Internet by offering the right interface to applications. We thus would like to provide similar features in SANETs – interconnect them on a large scale and provide applications in the Internet with the ability to interact with the physical world.

We think that SANETs are very different from traditional end-points in the current Internet, because they provide the interface to the physical world.

In the Internet, IP addresses belong to a structured virtual space unrelated to the physical world, while sensor nodes are placed at a given geographical location. For most applications, the information on the place and the time of a sensor reading is almost as important as the reading itself. For instance, it would be important to know that a node has detected a fire in a given office at a given time. At the same time, much information gathered by sensors is asynchronous, for instance a node that detects a threshold crossing of a measured quantity needs to notify a base station for further processing. All these examples lead to the conclusion that the focus of communication in SANETs is on data, which can include other information such as place and time in addition to a simple reading, rather than on IP addressable end-points, so the current communication model of the Internet is not suitable for SANETs. Thus, we can imagine a layer that offers the view of typed data chunks coming from and going to some nodes instead of the traditional vision of addressable end-points. Briscoe has expressed a similar view on the interest of a data centric approach applied to sensor networks and pervasive computing [2].

Similarly, one would be interested in communicating with a SANET network through primitives that allow expressing the fact that a given piece of information comes from a given place, at a given instant, and provides a measurement of a given quantity. Current communication primitives such as sockets do not support this data-centric view.

To define the right place of SANETs in the Internet, we consider below the following most relevant issues:

- Interconnection architecture: what is a suitable architecture that provides minimal scalable interconnection?
- Communication interface: what is the right abstraction for representing a SANET network?
- Protocols: what are the right protocols to support the communication interface?

3.1 Interconnection Architecture

Figure 1 illustrates the proposed interconnection architecture: a *data router* interconnects SANETs and offers a data centric view on the physical world to the rest of the Internet. As in the data centric view for the Future Internet, the data router operates on *data chunks* that are *typed, signed*, and possibly *encrypted*. Instead of communication end-points with IP addresses, *data publishers* representing sensor sources of information provide data chunks to *data subscribers*. Actuators can correspond to data subscribers willing to receive some data of special type, e.g., control commands. The data router acts as a data forwarder and a cache so that a subscriber can obtain a data chunk published before a subscription. It can interconnect several heterogeneous SANETs as well as offer data chunk publishing and subscription mechanisms to end-hosts in the Internet. To discover potential data publishers and subscribers, the data router offers a function that advertizes data types and metadata related to each publisher or subscriber. In this way, an application can discover how

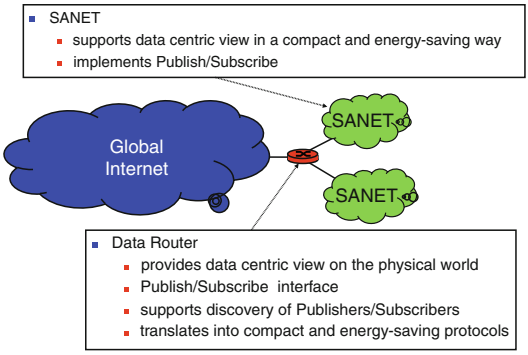


Fig. 1 Vision of the Internet of Sensors – integration of SANETs within the Internet

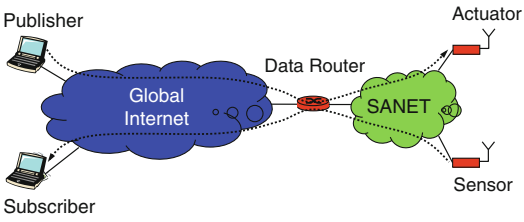


Fig. 2 Data publishers and subscribers

a given data chunk can be consumed when received or what data chunk can be sent to a subscriber. The data router can also provide a directory service for discovering potential data publishers and subscribers. Note that the main difference between this view and existing Publish/Subscribe proposals for sensor networks [3,6,9,13] is the placement of the functionalities in the protocol stack: in our vision, Publish/Subscribe protocols belong to lower layers and they are closely coupled with MAC access methods rather than provided by a middleware.

Figure 2 further illustrates the operation of the proposed interconnection architecture. We distinguish four entities: Sensors, Actuators, Publishers, and Subscribers (note that Sensor are also Publishers and Actuators are also Subscribers). Sensors publish their data for consumption by Subscribers and Publishers provide some data such as commands to Actuators. In the figure, we have placed Publishers and Subscribers in the Internet part, but they can also connect to a SANET. There can be several Data Routers and SANETs.

To explain the operation of the architecture, let us consider a sensor that publishes a data chunk: it propagates in the SANET (at the beginning it may propagate as a flooding message across the SANET) and reaches the Data Router that can store it in a cache and register in a directory of published content. The Data Router replies with a message that follows the inverse path to the Sensor thus fixing the direct route in intermediate nodes (note that this way of operation is inverse to Directed Diffusion [8]). A Subscriber sends request for a data chunk to the Data Router that can provide the data. If the operation starts with a Subscriber, it is up to the Data Router to propagate the request for data in the SANET so that the Sensor can publish its data chunks (this behavior is similar to Directed Diffusion). In a similar way, an

Actuator can send a subscription for commands that the Data Router can match to the data provided by a Publisher.

Propagation of subscriptions in the Internet may rely on the standard IP multicast, whereas routing in the SANET requires exploring new approaches, for instance based on establishing routing indices in intermediate nodes in function of received subscriptions and published data chunks. Routes need to take into account the quality of radio links and energy of nodes. Propagation of subscriptions can also be related with setting up common time schedules for wake up at MAC layer. As we can imagine Data Routers that change their positions, routing in the SANET needs to dynamically adapt to their possible different locations.

3.2 DATA Layer and its Interface

Figure 3 shows the structuring of the protocol layers. The *DATA* layer implements the Publish/Subscribe protocol for discovering and matching producers and consumers of typed data chunks. It offers the interface that we call *content* to emphasize on its role – enable communication via typed data content. Above the DATA layer, application protocols may provide some other useful high-level functions such as a database interface for *querying* information retrieved from a SANET network, receiving *asynchronous notifications* or *pushing* some information to a SANET network. We can observe that Demmer et al. have defined a similar communication API [4], however it is more oriented toward the standard networks as a kind of a data-centric replacement to the socket layer.

Data chunks in our model are vectors of typed values: $\bar{V} : \{v_1, v_2, \ldots, v_n\}$, where each component v_i is a value of a given type. For instance, a data chunk published by sensor measuring a temperature at a given place at a given instant could be the following: {`Temp: int`, `Place: geo-coordinates`, `Time: timestamp`}; an actuator that controls a camera can be seen as a subscriber that can accept the following chunks {`Tilt: tilt-units`, `Pan: pan-units`, `Zoom: zoom-units`, `Place: geo-coordinates`, `Time: timestamp`}.

A flow of data chunks is uniquely identified by a content identifier (CID), a short value derived either from the identifier of a sensor (e.g., public key) or from some values of the data chunks that remain constant. The idea is to attach a small identifier to data chunks that the network could use for forwarding and dissemination. De-

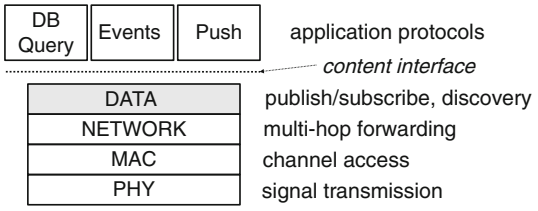


Fig. 3 DATA layer in the SANET protocol stack

pending on different types of data, CID can support different communication primitives:

- *Geocast*. CID specifies geographic coordinates (e.g., camera at Tour Eiffel)
- *Timecast*. CID specifies a given time instant (e.g., sensor data generated at midnight)
- *Subjectcast*. CID specifies a name of an object or class of devices (e.g., a light switch)
- *Controlcast*. CID specifies an operation to perform (e.g., camera tilt)
- *Objectcast*. CID specifies the object of an operation (e.g., zoom on a car)
- *Modecast*. CID specifies the mode of performing an operation. (e.g., wide pan).

The Data Router maintains the mapping between a CID and typed values of data chunks $\bar{V}$, so that subscribers can discover the format of the information and learn how to interpret them. Another possibility to explore is to build a distributed service based on a DHT to store the mapping.

We can think of CIDs as a kind of generalized label used for forwarding data chunks. Consider the example of CID referring to geographical coordinates – they allow to exploit geographical routing that presents many advantages with respect to traditional network addressing. For instance, we can apply the approach based on Waypoint Routing [12] to SANETs. Similarly, we can explore how to set up efficient and energy aware routing based on other types of information such as device names or operations to perform. Note that CIDs are intended to be only used within a SANET, so that their scope is limited and particular routing protocols can be adopted [1].

To be able to subscribe to a given flow of data chunks, subscribers need to first discover them. Data Routers can run a Discovery service as a centralized directory or as a distributed service based on a DHT approach. Matching types advertized by Data Routers may require the use of *ontologies* to set formal basis on the terms used in type definitions and avoid problems with not exactly the same meaning of data chunks.

The main functions of the DATA layer protocol are twofold:

1. A Data Router needs to advertize flows of data chunks, their data types, and metadata describing the contents. Consider for example a sensor that measures a temperature at a given place and an actuator that adjusts heating to obtain a given temperature. A control application needs to discover the possibility offered by the sensor of providing the temperature reading and by the actuator of accepting to set a heating level.
2. Publishers and subscribers need to declare themselves through a Publish/Subscribe mechanism. For instance, a control application may subscribe to the data published by a sensor. Similarly, an actuator may need to subscribe to a possible source of heating control commands and a control application publishes such commands.

4 Research Agenda

Here, we point out some issues of a research agenda to achieve the objectives of the Internet of Sensors.

- Design and develop a prototype Data Router for interconnecting SANETs and representing them in the Internet. As most of SANETs we consider are wireless, this means that the router is in fact a radio gateway between different technologies used in sensor networks (e.g., 802.15.4, Coronis Wavenis) and 802.11 WLAN connected to the Internet. We plan to develop such a prototype based on a software/reconfigurable radio able to operate as a 802.11 and a 802.15.4 entity. The card will use an FPGA software radio prototype and enable experiments with interfacing SANETs with the Internet through an 802.11 access network.
- Prototype the DATA layer for interconnecting with Internet applications. This will require an efficient implementation of the Publish/Subscribe protocols in the Internet part of the whole interconnection architecture. A natural support for this kind of protocols is IP multicast. The data router would need to provide support for advertizing flows of data chunks, informing about the types of their data, and supporting the data-oriented communication primitives discussed above such as geocast, timecast etc.
- Define and develop the internal routing and MAC protocols for supporting the data centric view inside SANETs. We plan to work on efficient and energy limited mechanisms for the Publish/Subscribe protocols based on CID forwarding. In particular, we think that forwarding published data chunks and subscribe messages may require setting up routing indices that dynamically provide the information about the neighbor, to which a given message needs to go in the function of its CID. Propagation of data chunks will also require efficient dissemination protocols such as probabilistic flooding or other optimizations that take into account energy. The internal SANET protocols would also require exploring new cross-layer approaches (PHY/MAC/NET) for finding the right trade-off between high performance and energy savings in the implementation of the discovery protocol, subscription propagation, and publishing data chunks. Supporting data-oriented communication primitives like geocast, timecast etc. inside SANETs may require providing geographical routing, localization schemes, time synchronization, and efficient packet forwarding.
- Specify security mechanisms for authenticating all entities, digitally signing and encrypting, if needed, data chunks, managing access authorizations in the Discover/Publish/Subscribe protocols.

5 Related Work

Several authors have considered the Publish/Subscribe model for sensor networks. Briscoe has discussed many issues related to this view in the context of ubiquitous networks and proposed several original ideas on how to integrate small sensor

nodes within the large scale Internet [2]. Several authors have proposed to develop a Publish/Subscribe middleware for managing sensor networks [3, 6, 9, 13]. This approach integrates the Publish/Subscribe paradigm at upper layers, so that it is difficult to achieve sufficient efficiency and energy aware operation in SANETs. RTFM (Rendezvous, Topology, Forwarding, and physical Media architecture) is the only architecture to our knowledge that puts the Publish/Subscribe into lower layers (on top of the link abstraction) [11], however this proposal concerns the Future Global Internet and not sensor networks.

6 Conclusion

In this position paper, we have proposed a new view on the integration of wireless *Sensor and Actuator Networks* in the Internet. We believe that SANETs may benefit from the *data dissemination* paradigm that some researchers suggest for the Future Internet. In this approach, the network conveys typed data chunks while applications organize communication according to the Publish/Subscribe model: data consumers subscribe to chunks advertized by producers. We think that such a model is particularly suitable for representing the operation of SANETs and will foster new interesting research activities required for integrating them within the Internet.

In our view, the central element of the integrated architecture is a *data router* that interconnects SANETs and offers a data centric view on the physical world to the rest of the Internet. To discover potential data publishers and subscribers, the Data Router offers a function that advertizes data types and metadata related to each publishers or subscribers. At the same time, it needs to interact with sensor and actuators according to a lightweight Publish/Subscribe protocol tailored to the energy and computing constraints.

Acknowledgment This work was partially supported by the French Ministry of Research project ARESA under contract ANR-05-RNRT-01703.

References

1. Bachir A, Barthel D (2005) Localized max–min remaining energy routing for WSN using delay control. In: Proceeedings of the ICC, IEEE international conference on communications, vol. 5. pp 3302–3306
2. Briscoe R (2004) The implications of pervasive computing on network design. BT Technol J 22(3):170–190
3. Costa P, Picco GP, Rossetto S (2005) Publish-subscribe on sensor networks: A semi-probabilistic Approach. In: Proceedings of the 2nd IEEE international conference on mobile ad-hoc and sensor systems (MASS05). Washington, DC, USA
4. Demmer M, Fall K, Koponen T, Shenker S (2007) Towards a modern communications API. In: Proceedings of HotNets-VI
5. Dohler M, Watteyne T, Winter T (2008) Urban WSNs routing requirements in low power and lossy networks. Technical report, IETF ROLL workgroup

6. Hauer JH, Handziski V, Köpke A, Willig A, Wolisz A (2008) A component framework for content-based publish/subscribe in sensor networks. In: Proceedings of 5th European workshop on wireless sensor networks (EWSN), Bologna, Italy
7. Hui JW, Culler DE (2008) Extending IP to low-power, wireless personal area networks. Internet Comp IEEE 12(4):37–45
8. Intagonwiwat C et al (2003) Directed diffusion for wireless sensor networking. IEEE/ACM Trans Netw 11(1):2–16
9. Nam CS, Jeong HJ, Shin DR (2008) Design and implementation of the publish/subscribe middleware for wireless sensor networks. In: International conference on networked computing and advanced information management, vol 1. pp 270–273
10. Polastre J, Hui J, Levis P, Zhao J, Culler D, Shenker S, Stoica I (2005) A unifying link abstraction for wireless sensor networks. In: ACM SenSys '05: proceedings of the 3rd international conference on embedded networked sensor systems, pp 76–89
11. Sarela M, Rinta-aho T, Tarkoma S (2008) RTFM: publish/subscribe internetworking architecture. In: Proceedings of the IST Mobile Summit. Stockholm, Sweden
12. Schiller E, Starzetz P, Rousseau F, Duda A (2008) Binary waypoint geographical routing in wireless mesh networks. In: Proceedings of the 11th international workshop on modeling analysis and simulation of wireless and mobile systems (MSWiM 2008), ACM, Vancouver, BC, Canada, pp 252–259
13. Souto E, Guimaraes G, Vasconcelos G, Vieira M, Rosa N, Ferraz C (2004) A message-oriented middleware for sensor networks. In: MPAC '04: proceedings of the 2nd workshop on middleware for pervasive and ad-hoc computing, pp 127–134

Energy and Distortion Minimization in "Refining" and "Expanding" Sensor Networks

Franco Davoli, Mario Marchese, and Maurizio Mongelli

1 Introduction

The deployment of sensor networks is often such that measurements acquired by the sensor nodes are conveyed toward a sink, where they need to be processed and analyzed. Recently, there has been a growing interest in understanding the peculiarities of wireless sensor networks (WSN) from both an information theoretic and decision theoretic point of view, particularly when the measured quantities can be represented as Gaussian random variables (see, e.g., [1–5]). In particular, when the measured variables are analog quantities, they can either be transmitted as such, or they can be quantized and transmitted according to a digital scheme. Given a distortion measure for the reconstruction of the variables at the sink, and the statistical characteristics of the communication channel and of the sources, it is not straightforward to determine whether joint source-channel coding (with analog transmission) can outperform the separation that is typical of digital communications [1]. The optimality of uncoded transmission – well known in the case of a scalar Gaussian source over a Gaussian channel with quadratic distortion – has been proved in one such network configuration [2]. However, even small changes in the structure of the channels (e.g., working with an inhomogeneous network with different channel attenuations due to path loss and fading) can produce different results [4].

In [1], two paradigms are identified, named "refining" and "expanding" sensor network, respectively. In the first case, the communication infrastructure is relatively "poor" (e.g., with a high level of interference), and the sensors pick up measurements from a limited number of sources. The situation considered in [2] is a special case of this. In the second network type the sensors measure each relatively independent sources and the communication infrastructure is "rich" (in the extreme case, each sensor has an independent channel).

In the present paper, we consider both of the above problems from a decision theoretic perspective, in the setting of team theory [6]. Specifically, we expand

F. Davoli (✉), M. Marchese, and M. Mongelli
DIST-University of Genoa, Via Opera Pia 13, 16145 Genoa, Italy
e-mail: franco@dist.unige.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_31,

with further results the considerations reported in [5], where we have introduced decentralized decision models of this type, based on the approximation of the optimal decision strategies by means of fixed-structure parametrized nonlinear functions, by applying the Extended Ritz Method (ERIM) [7]. The reason for seeking numerical approximations to the optimal coding/decoding strategies stems from the fact that a team optimization approach to these problems presents formidable analytical difficulties (originally pointed out in [8], even in a scalar source-channel model).

The paper will be organized as follows. Our main emphasis is in the "refining" WSN model, where we use the neural approximating functions to derive a nonuniform power distribution among the encoders at each sensor node. We show that, in the presence of correlated measurements from the same source representing a physical phenomenon, a large reduction in the overall transmission power can be obtained, at the expense of very little increase in distortion, with respect to linear encoding strategies. This is achieved by selecting a few "representative" sensors from the total available pool, and by distributing the available power nonuniformly among the sensors.

As regards the "expanding" WSN, we aim at minimizing again (under an overall power constraint) a quadratic distortion function that, however, now weighs the sum of the square errors of each individual measured variable. The optimum (centralized) encoder–decoder pair in the class of linear strategies for this problem was originally derived in [9]. We define and briefly discuss two other decentralized team optimization problems in this setting, whose numerical solution and comparison will be the subject of future work.

2 Problem Statement for the "Refining" WSN

We consider a number N of sensors deployed over a geographical area, each one observing a realization of some physical phenomenon described by a random variable (r.v.) S (the source). We adopt the model of [4], which we describe in the following section. We suppose the observations to take place at discrete time instants, but, since we are interested in real-time, single-letter coding, we do not introduce the time index in the following for simplicity of notation. Successive source outputs are uncorrelated; however, there is a spatial correlation between the source and the event observed by sensor i, represented by the r.v. S_i. As a consequence, the r.v.'s S_i and S_j are also mutually correlated. We indicate by $\rho_{s,i}$ and $\rho_{i,j}$ the correlation coefficients between S and S_i, and between S_i and S_j, respectively. Moreover, we suppose $S \sim \mathcal{N}\left(0, \sigma^2\right)$, and that all the other variables $S_1, \ldots, S_N$ are jointly Gaussian, with 0 mean, the same variance σ^2, and covariance matrix Σ_S. Measurements are corrupted by observation noise, so that sensor i observes a realization of the r.v.

$$X_i = S_i + N_i \tag{1}$$

with $N_i \sim \mathcal{N}\left(0, \sigma_N^2\right), \forall i$. The measurements are encoded at each sensor according to some real-time coding strategy

$$Z_i = f_i(X_i) \tag{2}$$

and the sink receives a channel output of the type

$$\boldsymbol{Y} = \text{col}[Y_1 \dots Y_N], Y_i = Z_i + W_i \tag{3}$$

with $W_i \sim \mathcal{N}\left(0, \sigma_W^2\right), \forall i$. The sink's decoding strategy is also real-time and given by

$$\hat{S} = g(\boldsymbol{Y}) \tag{4}$$

Functions $f_i(\cdot), i = 1, \dots, N$ and $g(\cdot)$ should be chosen to minimize the quadratic distortion measure

$$D = E\{(S - \hat{S})^2\} \tag{5}$$

under the overall power constraint

$$\sum_{i=1}^{N} E\{Z_i^2\} \leq \Gamma \tag{6}$$

This problem, which will be referred to as *Problem 1*, assumes the presence of multiple receiving antennas at the sink (i.e., of an additive Gaussian noise MIMO channel, characterized by an identity matrix).

To simplify the analysis, the topology of the network is considered fixed, namely, no faults or movements of the sensors are possible. Thus, a static covariance matrix describes the mutual correlation among the input of the sensors. It depends on the distance between each pair of source-sensor and sensor–sensor; we refer to it as *topological covariance matrix*. To operate in the same setting as [4] for comparison, the noise W_i in (3) is ignored from now on. Transmission noises could anyway be included straightforwardly in our treatment.

In [4], uncoded transmission is adopted for the sensors; then, by exploiting the fact that the source observations are correlated, the minimum number of sensors that need to be activated to achieve nearly optimal distortion is sought, out of the total number of deployed sensors. An example may help understand. We consider both source S and noise in (1) having standard normal distributions ($\sigma^2 = \sigma_N^2 = 1$). Figure 1 represents a possible deployment of 30 sensors over a 50×50 grid (the other elements in the figure will be used in later comments); each element of the topological covariance matrix, with indexes i, j, is given by $\sigma^2 \cdot \mathrm{e}^{-d_{ij}/10}$, according to a power exponential covariance model, d_{ij} being the distance between nodes i, j. The choice of the subset of sensors to activate that is operated in [4] is based on the rationale that one should activate no more than a certain number of sensors (e.g., $N/10$, out of the N available), and they should be at the same time sufficiently

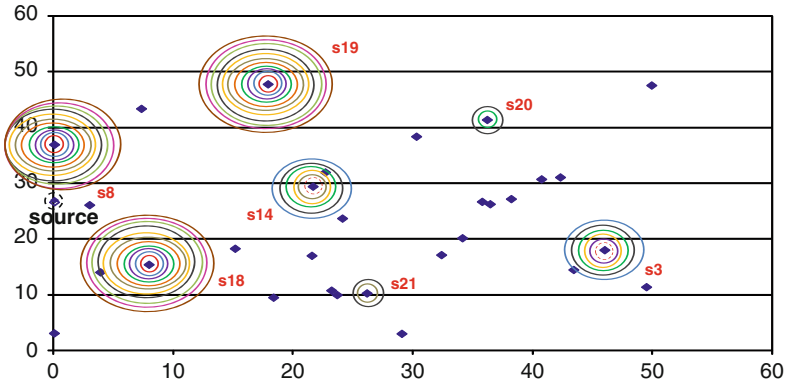


Fig. 1 Sensors' deployment (distances in $[m]$ on x and y axes) and activations after training using different combinations of source and noise variances

close to the source and sufficiently far away from one another in order to minimize mutual interference. However, though it is known that the optimal subset depends on some geometric property of reciprocal sensors' positions, how to let the network self-learn this subset is still left as an open issue, and the subset is determined by means of successive trials in [4]. A method for this determination, different from the one we derive here, is provided in reference [10]; an advantage of the neural strategies we apply in the following consists in the simultaneous assignment of power to the selected sensors.

3 Nonlinear Parametric Approximation of the Optimal Strategies

In this perspective, we reformulated Problem 1 in [5], by letting the coding and decoding strategies (2) and (4) depend on some nonlinear approximation scheme. We remark again that the coding–decoding strategies are derived for a static topological covariance matrix. This assumption will be relaxed later. Introducing nonlinear approximators in (2) and (4) means replacing them with:

$$Z_i = \widehat{f}_i(X_i, \boldsymbol{w}_{f_i}) \tag{7}$$

$$\hat{S} = \widehat{g}(\boldsymbol{Y}, \boldsymbol{w}_g) \tag{8}$$

where $\widehat{f}(\cdot)$ and $\widehat{g}(\cdot)$ are neural networks depending on the choice of the basis functions (e.g., sigmoidal) of each layer, and $\boldsymbol{w}_{f_i}$ and $\boldsymbol{w}_g$ are vectors of parameters activating the basis functions. Let $\boldsymbol{w}_f = \mathrm{col}\left[\boldsymbol{w}_{f_1}, \ldots, \boldsymbol{w}_{f_N}\right]$. The application of the entire vector $\boldsymbol{Y} = \mathrm{col}\,[Z_1, \ldots, Z_N]$ in (8) helps highlight the performance gain induced by taking into account the topological structure (through explicit consideration of the cross-correlation in the strategies). Equations (7)

and (8) are called *neural coding* and *decoding strategies*. Replacing (2) and (4) in the cost (5) with the neural strategies leads to the following parametric optimization problem (*Problem 2*):

$$\begin{aligned} \boldsymbol{w}_f^{\mathrm{o}}, \boldsymbol{w}_g^{\mathrm{o}} &= \arg \min_{\boldsymbol{w}_f, \boldsymbol{w}_g} J(\boldsymbol{w}_f, \boldsymbol{w}_g); J(\boldsymbol{w}_f, \boldsymbol{w}_g) = E\left\{\left(S - \hat{S}\right)^2\right\}; \\ \hat{S} &= \widehat{g}\left(\left[\widehat{f}_1(X_1, \boldsymbol{w}_{f_1}), \ldots, \widehat{f}_N(X_N, \boldsymbol{w}_{f_N})\right], \boldsymbol{w}_g\right); \; i = 1, \ldots, N \end{aligned} \tag{9}$$

in order to find out the optimal neural strategies $\widehat{f}_i^{\mathrm{o}}(\cdot) = \widehat{f}_i(\cdot, \boldsymbol{w}_{f_i}^{\mathrm{o}})$ and $\widehat{g}^{\mathrm{o}}(\cdot) = \widehat{g}(\cdot, \boldsymbol{w}_g^{\mathrm{o}})$ under the power constraint (6): $\boldsymbol{w}_{f_i}, i = 1, \ldots, N$: $\sum_{i=1}^{N} E\left\{\widehat{f}_i^2(X_i, \boldsymbol{w}_{f_i})\right\} \leq \Gamma$. How the optimal neural strategies are capable to introduce a performance gain in the distortion and to distribute the power among the sensors better than the linear strategies is studied in the next section.

Some technical details about the solution of Problem 2 can be found in [5]. Resorting from the team decision formulation of the functional optimization Problem 1 to the parametric approximation of Problem 2 is just an application of the methodology known as *Extended Ritz* [7]. Since a closed-form expression of the expected cost $J(\cdot)$ in (9) is not easily available, $J(\cdot)$ is substituted by its Montecarlo estimation $\tilde{J}(\cdot)$, $\tilde{J}(\cdot)$ being an arithmetic average over a given number Ξ of realizations of the random variables. More specifically, Ξ different samples of X_i and $N_i, i = 1, \ldots, N$, are generated on the basis of the topological covariance matrix and the distortion is computed under a given structure of the neural strategies (i.e., $\boldsymbol{w}_{f_i}$ and $\boldsymbol{w}_g$ are fixed). In doing this, the cost is also "augmented," by adding a penalty term, to ensure the satisfaction of the power constraint. Then, a gradient descent procedure belonging to the family of *stochastic approximation* algorithms [11] is applied, where the gradient at each descent step, after proper initialization, is computed by means of the back-propagation algorithm of neural networks (see [7] and references therein).

4 Performance Evaluation and Discussion

The performance evaluation is related to the network of Fig. 1. Source S and noise in (1) have normal distributions (with unitary variances). We suppose that for linear strategies, each sensor cannot transmit with more than one unit of power. The constraint over the overall power consumption used in Problem 2 is therefore $\Gamma = 30$. Neural strategies are based on one-hidden layer neural networks with hyperbolic tangent neural units (one unit for each sensor and 5 units for the sink). Gradient stepsizes (both for encoders and for the decoder) have the form $1/500 + k$; the penalty cost function parameter K_{p} (used to weigh a quadratic penalty term) is 0.25 and $\Xi = 10^5$. The training phase took 18 min over an Intel

processor@1.73 GHz. The distortion during the training phase decreased from a value of about 30 to almost 1 in slightly more than 100 training steps [5], whereas the power allocation at the end of training is exactly Γ. All this is comparable to what can be achieved by the linear strategies with uniform power allocation. However, the most important information that can be derived from the neural analysis is that power is allocated unequally, and a subset of "good" sensors clearly emerges, whose information would give the main contribution to decreasing distortion. In the case of our example, it is shown in [5] that the best subset is $\{8, 18, 19\}$ (with reference to Fig. 1), with a total power of 3.72 units. Basically, neural strategies learn the principle of activating sensors close to the source and sufficiently separated to avoid reciprocal interference.

In order to validate the robustness of the neural power allocation, we consider now the repetition of several training phases with respect to different combinations of source and noise variances. Variances in play are in the range $5 \div 30$, for both source and noise. A training phase is started from the beginning for each combination of variances. The qualitative result is depicted in Fig. 1, where a circle is marked around a sensor each time that sensor is chosen after training in virtue of its final power allocation (a simple comparison is made at the end of training, by setting a threshold on the assigned power for activation). The position of the circles in Fig. 1 means that the discovered optimal choice for sensor activation (for most of the times, $\{8, 18, 19\}$) is invariant to the source and noise statistical behaviors. The rationale of this effect can be explained as follows. It is arguable that the final result in the power allocation is mainly influenced by the reciprocal correlation of each couple of sensors and source-sensor (the spatial correlation coefficients $\rho_{s,i}$ and $\rho_{i,j}$ introduced at the beginning of Sect. 2). In this view, changing the specific statistical behavior of source and noise does not introduce a significant effect. The neural scheme is capable to capture the reciprocal influence coming from the spatial correlation coefficients to discover an optimal power allocation scheme, which is thus robust to possible changes in the statistical environment. However, it is remarkable that a new training phase would be needed when a significant topology change takes place: the results in Fig. 1 significantly change if we generate a topological permutation over a large subset of nodes. How the power allocation may be invariant to the "size" of such a permutation is an intriguing question, as well. The metric itself of the permutation should be studied; this means looking for a synthetic representation of the mapping between nodes' positions and optimal power allocation. More specifically, it may be arguable that the power allocation does not significantly change until some synthetic representation of the topology remains stable. A possible metric is the size of the areas of the underlying Voronoi diagram, as highlighted in [10]. A Voronoi diagram measures how the nodes in a graph are far away from each other and synthesizes their reciprocal influence. A large set of applications is covered by the Voronoi principle, also including wireless networks [12]. The robustness of the approach with respect to source mobility is currently under investigation, too.

One final remark is worth to be mentioned concerning the robustness of the neural scheme. It is related to the distortion obtained by the strategies that allocate the same power (1 unit each) to the chosen sensors (termed "Linear Sensor – Neural Sink $\{8, 18, 19\}$" technique in [5], and which represent the best compromise between distortion and power consumption, according to the results therein) in comparison with those of [4] applied to the same sensors (termed "Linear Strategies $\{8, 18, 19\}$" technique in [5], and not taking into account mutual correlations in the decoder's strategy), in the presence of variances different from the ones used during training. This means that we test the two techniques above with some "test sets," which are different from the ones used for training the neural networks. More specifically, these sets collect the sequences of samples (10^5 in our case) produced by the source and noise probability distributions during a simulation that calculates the distortion (and the power allocation) via a Montecarlo approximation, as explained in Sect. 3. The training set exhibits fixed normal distributions and the test sets derive from normal distributions with variable variances. We denote by δ_{d} the quadratic difference between the distortion obtained at the end of training (i.e., over the set extracted during the last training step) and the one obtained with a specific test set. Figure 2 shows δ_{d} as a function of increasing source variance (σ^2) and noise variance (σ_N^2) used in the test sets ($\sigma^2 = \sigma_N^2 = 1$, over the training set). In the linear approach, the linear coding formula ((6) in [4]) is used with $\sigma^2 = \sigma_N^2 = 1$ and an "ideal" distortion d^* is calculated using samples generated from normal distributions (coherently with the expected $\sigma^2 = \sigma_N^2 = 1$); δ_{d} is thus computed in the linear case as the quadratic difference between d^* and the distortion obtained by using samples generated from Gaussian distributions with specific variances, namely, with the coding formula not updated with respect to the real variances used to generate the samples. From Fig. 2, it is quite evident that the neural approach outperforms the linear one, as it limits the error (δ_{d}) introduced by the application of test sets, whose samples are progressively different from the expected ones.

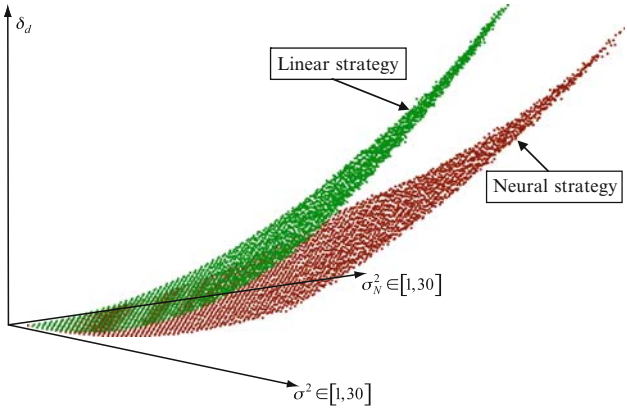


Fig. 2 δ_{d} with different test sets

5 Considerations on the "Expanding" Case

We are here in almost the same setting as in the previous case, with the exception of two notable differences: (i) we consider a number M of sensors deployed over a geographical area, where the ith sensor observes a realization of some physical phenomenon described by a random variable S_i (the ith source), and we suppose the source variables to be jointly Gaussian, with zero mean and covariance matrix $\Sigma_S(\boldsymbol{S} = \mathrm{col}\,[S_1, \ldots, S_M]; \boldsymbol{S} \sim \mathcal{N}(0, \Sigma_S))$; (ii) the distortion is now given by $D = E\left\{\left\|\boldsymbol{S} - \hat{\boldsymbol{S}}\right\|^2\right\} = E\left\{\sum_{i=1}^{M}\left(S_i - \hat{S}_i\right)^2\right\}$.

We are interested in investigating and comparing the following problems, which will be the subject of future work.

Problem 3: Find the neural (decentralized) encoders and the decoder, along the same lines as outlined in Sect. 3.

Problem 4 (Centralized Linear), used for comparison: This is the Gaussian vector source and channel problem treated in reference [9], under the constraint that the encoder's and decoder's strategies be linear. The encoder is centralized, i.e., it has access to all sensors' measurements; this case might correspond to that of a sink, whose task is simply retransmitting all collected measurements toward a processing center.

Problem 5 (LQG): Suppose that the decoding strategy is constrained to be a linear operator. Then, the decentralized encoding team problem is a static LQG (Linear-Quadratic-Gaussian) one. However, the problem is nonclassical, owing to the presence of the power constraint and to the fact that the coefficients of the decoding matrix are dependent on the encoding strategy. If the latter is linear (as in the classical LQG static team optimization [13,14]), the determination of its coefficients entails a nonlinear parametric optimization problem, rather than a linear one as in the classical case. These points deserve further investigation, since the global optimality of the linear solution in the LQG static team is related to its unicity, which may be no longer guaranteed.

6 Conclusions

The collection of measurements in wireless sensor networks raises intriguing and interesting problems, both from the information theoretic and the decision theoretic point of view. The latter is characterized by the difficulty of dealing with decentralized team optimization problems, whose solution generally entails formidable analytical and computational difficulties. We have further investigated the effectiveness of the application of parametric approximations of the optimal strategies, based on back-propagation neural networks, which we adopted in [5] for the case of multiple correlated observations of a single source representing a physical phenomenon

("refining" WSN). The additional results described in this paper highlight a certain degree of robustness of the neural strategies in determining the most effective power allocation even in the presence of mismatching between the data used for training the approximators and the real ones.

As regards the case of multiple (possibly correlated) sources, whose measurements are again transmitted over multiple channels ("expanding" WSN), we have defined a number of problems, whose investigation will be the subject of our future work.

References

1. Gastpar M, Vetterli M, Dragotti PL (2006) Sensing reality and communicating bits: a dangerous liaison. IEEE Signal Process Mag 23(4):70–83
2. Gastpar M (2008) Uncoded transmission is exactly optimal for a simple Gaussian 'sensor' network. IEEE Trans Inf Theory 54(11):5247–5251
3. Wei S, Kannan R, Iyengar SS, Rao NS (2008) Energy efficient estimation of Gaussian sources over inhomogenous Gaussian MAC channels. Proceedings of IEEE Globecom 2008, New Orleans, LA, Nov–Dec 2008, pp 1–5
4. Vuran MC, Akan ÖB, Akyildiz IF (2004) Spatio-temporal correlation: theory and applications for wireless sensor networks. Comput Netw 45:245–259
5. Davoli F, Marchese M, Mongelli M (2009) A decision theoretic approach to Gaussian sensor networks. Proceedings of adhoc and sensor networking symposium, IEEE international conference on commununications 2009 (ICC 2009), Dresden, Germany, June 2009
6. Ho YC, Kastner MP, Wong E (1987) Teams, signaling, and information theory. IEEE Trans Automat Contr AC-23:305–311
7. Zoppoli R, Sanguineti M, Parisini T (2002) Approximating networks and extended Ritz method for the solution of functional optimization problems. J Optim Theory Appl 112(2):403–439
8. Witsenhausen HS (1968) A counterexample in stochastic optimum control. SIAM J Control 6:131–147
9. Pilc RJ (1969) The optimum linear modulator for a Gaussian source used with a Gaussian channel. Bell Syst Tech J 48:3075–3089
10. Vuran MC, Akyildiz IF (2006) Spatial correlation-based collaborative medium access control in wireless sensor networks. IEEE/ACM Trans Networking 14 (2):316–329
11. Kushner HJ, Yin GG (1997) Stochastic approximation algorithms and applications. Springer-Verlag, New York, NY
12. Portela JN, Alencar MS (2005) Spatial analysis of the overlapping cell area using Voronoi diagrams. Proceedings of IEEE microwave and optoelectronics, Brasilia, Brazil, 25–28 July 2005, pp 643–646
13. Ho Y-C, Chu K-C (1972) Team decision theory and information structures in optimal control problems – Part I. IEEE Trans Automat Contr AC-17(1):15–22
14. Marschak J, Radner R (1972) Economic theory of teams. Yale University Press, New Haven, CT

An IEEE 802.15.4 Wireless Sensor Network for Energy Efficient Buildings

Chiara Buratti, Alberto Ferri, and Roberto Verdone

1 Introduction

In the recent years, the problem of energy saving has attracted the attention of many researches in a plethora of fields. In Europe, 40% of the energy is consumed in buildings, more than by industry or transport. The tendency shows that the total energy consumption has been rising since 1990 and will continue if strong actions are not taken.

The eDIANA (Embedded Systems for Energy Efficient Buildings) project [1], funded by the European Commission within FP7 through the ARTEMIS framework, addresses the need of achieving energy efficiency in buildings through innovative solutions based on networked embedded systems.

The main goal of eDIANA is to achieve greater efficiency in the use of resources, prioritizing energy as scarce resource, more flexibility in the provision of resources, and better situation awareness for the citizen and for service and infrastructure owners. This will be achieved through the deployment of embedded systems in the eDIANA scenario. eDIANA is a strong application-oriented initiative that is focused on the design, development, and validation of the eDIANA platform and which will integrate intelligent embedded devices installed in residential and non residential buildings to improve energy efficiency and optimize overall energy consumption, production, and storage.

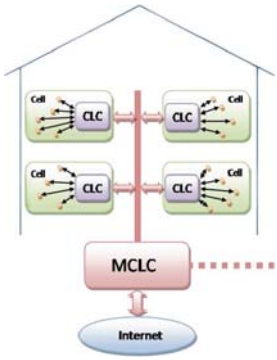
The main elements of the eDIANA scenario are the cells, that could be single houses, apartments, or working units, and the macrocells, which are in general groups of cells. Each macrocell identifies the contract with the energy service provider. Therefore, in case of apartments, the cell will coincide with the macrocell; whereas in management buildings, the macrocell will be composed of different cells, the working units using the same contract.

To handle and optimize energy use in cells and macrocells, the knowledge, in real-time, of the power consumed or produced (in case, for example, of the presence

C. Buratti (✉), A. Ferri, and R. Verdone
WiLAB, DEIS, University of Bologna, Bologna, Italy
e-mail: c.buratti@unibo.it; albertfe83@gmail.com; roberto.verdone@unibo.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_32,

Fig. 1 The eDIANA scenario



of photovoltaic panels) by every electrical appliance is fundamental. To such aim, wireless sensors could be distributed in the environment to forward the monitored data to a control unit, denoted as concentrator, incharge of optimally managing energy consumptions. The control is performed at the cell (cell-level concentrator, CLC) and at the macrocell level (MCLC). In Fig. 1, an example of management building scenario is shown.

Sensors could also be used to detect an event [2, 3]. As an example, in case a person approaches a washing machine, the concentrator could start providing energy to the machine; in this case, the sensor is used to detect the arrival of a person in a room.

Such wireless sensor network (WSN) must be able to work in a fully unplanned context; all the cells need to work under a self-paradigm, and the issue of interference between separate, uncoordinated cells is one of the most relevant of all. The WSN could also be complemented by a network using power line communication modems.

The eDIANA project started in February 2009. At the time when this paper was written, the scenario and system requirements were not fully described, and the technical solutions (e.g., air interfaces) were still to be selected. However, at the University of Bologna (one of the 22 partners and leader of the task related to the communication network), simulation activities to determine candidate radio technologies already started. This chapter reports on the current state of simulations, based on a scenario compliant with eDIANA and the assumption that IEEE 802.15.4 is used as air interface technology for the WSN component. In fact, such technology is very suitable for WSN [8, 11] and is one of the first candidate for the platform. The integration with the power line communication network will be considered at a later stage.

In this paper, we consider an apartment (a cell) and a building, composed of a number of cells (a macrocell), where a number of sensors (hereafter denoted as nodes) are distributed in given positions. IEEE 802.15.4 standard-compliant nodes [4] are distributed, and we assume that one personal area network (PAN) is formed in each cell and that the PAN coordinator is located at the CLC.

The application requires that the CLC receives data measured by sensors with a given periodicity. To this aim, the CLC, which is the coordinator of the PAN,

periodically sends queries and waits for replies. The data measured by sensors could be a sample of power consumed by an electrical appliance, or a sample of temperature, etc.. We assume that nodes use the beacon-enabled mode defined by the standard (i.e., the query coincides with the beacon packet). Nodes transmit their data via a direct link (star topology) or through a two-hop communication network (tree rooted at the coordinator). Finally, according to the standard, we assume that the different PANs work at different frequencies selected among the 16 carriers made available by the standard.

Owing to the complexity of the application, different performance should be studied. As an example, if a person approaches a washing machine for using it, the data transmitted by the sensor to allow the providing of energy to the machine should be received by the CLC with a certain reliability and with a limited delay. Also, energy consumption issues are fundamental to avoid a frequent recharge of sensor batteries. Therefore, to evaluate the applicability of 802.15.4 to the eDIANA scenario, we evaluate performance in terms of packet error rate, average delays, and energy consumption. The impact of interference on network performance is evaluated and also suitable comparison between different network topologies (star and tree) is accounted for.

2 The Reference Scenario and Channel Model

We consider an apartment and a building, composed of different apartments per floor, and possibly having different floors. In Fig. 2, the map of a single apartment is shown. One coordinator (denoted as CLC in the figure) and 36 nodes are deployed in different rooms. Sensors placed on the walls of the rooms are set at 0.4 m from the floor, since we assume they are located in the plugs for monitoring energy consumption. The sensors in the center of the rooms are on the ceiling, at 2.70 m. They can be used to control the intensity of the light that could change depending on the presence or absence of people in the room. Finally, nodes placed nearby the windows are at

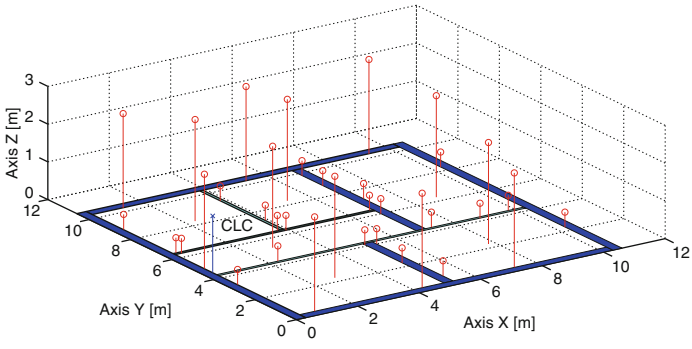


Fig. 2 The eDIANA apartment scenario considered

2.50 m. The PAN coordinator is placed at 1.5 m into the electric panel of the house. Its task is to control the WSN and exchange data via powerline communication with the power meter and the cell-level concentrator of the macrocell.

Two different kinds of walls are considered: thin (0.1 m) and thick (0.3 m). Also, the presence of the ceiling is accounted for.

The multiwall channel model described in [5] is used. According to this model, the loss in dB between two nodes at a distance d is given by

$$L = k_0 + k_1 \ln(d) + N_{\mathrm{wt}} L_{\mathrm{wt}} + N_{\mathrm{wl}} L_{\mathrm{wl}} + N_{\mathrm{c}} L_{\mathrm{c}} \tag{1}$$

where k_0 and k_1 are two constants; L_{wt}, L_{wl}, and L_{c} are the losses introduced by the thin and thick walls and the ceiling, respectively. N_{wt}, N_{wl} and N_{c} are the number of thin and thick walls and the number of ceilings between the two communicating nodes, respectively. We set $L_{\mathrm{wt}} = 5.9\,\mathrm{dB}$ $L_{\mathrm{wl}} = 8\,\mathrm{dB}$ (see results in [5] related to 2.4 GHz frequency), and $L_{\mathrm{c}} = 14\,\mathrm{dB}$ [6].

For what concerns the packet capture model, we use a threshold model. We assume that a packet is correctly received when both the following conditions are satisfied: (i) $P_{\mathrm{r}} > P_{\mathrm{r_{min}}}$, where $P_{\mathrm{r_{min}}} = -85\,\mathrm{dBm}$ is the receiver sensitivity and P_{r} is the received power given by $P_{\mathrm{r}}[\mathrm{dBm}] = P_{\mathrm{t}}[\mathrm{dBm}] - L[\mathrm{dB}]$, where P_{t} is transmit power and L is given by (1); (ii) $\frac{C}{I} \geq \alpha$, where C is the power received from the useful signal and I is the sum of the interference powers. We distinguish between co-channel (I_{co}), adjacent (I_{ad}), and alternate channel (I_{al}) interferences. I is given by $I = I_{\mathrm{co}} + w_{\mathrm{ad}} I_{\mathrm{ad}} + w_{\mathrm{al}} I_{\mathrm{al}}$, where the two weights are set according to the standard [4], therefore $w_{\mathrm{ad}} = 0.44$ and $w_{\mathrm{al}} = 0.44{\times}10^{-3}$; finally, we set $\alpha = 3.5\,\mathrm{dB}$.

3 The IEEE 802.15.4

The beacon-enabled mode of the 802.15.4 is used [4]. According to the standard, time is organized in a superframe structure, managed by the coordinator, composed of three parts: an inactive part, the contention access period (CAP), where the access to the channel is managed through a slotted carrier sense multiple access with collision avoidance (CSMA/CA) algorithm, and the contention free period (CFP), where a maximum number of seven guaranteed time slots (GTSs) can be allocated by the coordinator to specific nodes. Each superframe starts with a packet denoted as beacon, transmitted by the coordinator, which coincides with the query.

The duration of the active part and of the whole superframe depends on the value of two integer parameters ranging from 0 to 14: the Superframe Order, denoted as SO, and the Beacon Order, denoted as BO. In particular, the duration of the whole superframe (i.e., the interval of time between two successive beacons), denoted as T_{q}, is given by $T_{\mathrm{q}} = 16 \times 60 \times 2^{\mathrm{BO}} \times T_{\mathrm{s}}$, where T_{s} is the symbol time, equal to 16 μs; whereas the duration of the active part of the superframe (composed of CAP and CFP), denoted as T_{a}, is given by $T_{\mathrm{a}} = 16 \times 60 \times 2^{\mathrm{SO}} \times T_{\mathrm{s}}$.

For the sake of conciseness, we do not report the details of the CSMA/CA algorithm, but we refer to the standard [4].

An acknowledge mechanism is performed: each node, after the transmission of a packet, waits for the acknowledge packet for an interval of time equal to 54 T_s. In case the acknowledge is not received, the packet is retransmitted till the maximum number of retries is reached, or the superframe ends. We assume in fact that each node has one packet to be transmitted per superframe, and in case it does not succeed in transmitting it correctly by the end of the current superframe, the packet will be lost.

As stated above, two topologies are accounted for: stars and trees. In case of star topologies, nodes transmit their packets via a direct link to the coordinator by using the active part (CAP or CFP) of the superframe defined by the coordinator. In case of tree, the Zigbee-compliant tree-based topology is realized [7]. A tree rooted at the coordinator is formed and the inactive part of the superframe is used to allow children nodes in the tree to transmit toward the respective parents [8]. The tree-based topology and the access to the channel used in this case is described in the following.

3.1 The Tree-Based Topology

When the number of nodes in the PAN gets larger, star topologies are not suitable and peer-to-peer or tree-based topologies should be used [7, 10]. A three-level tree rooted at the PAN coordinator (namely, at level zero) is considered. Level 1 nodes receive data from level 2 nodes and forward them to the PAN coordinator. The tree-based topology defined by the Zigbee Alliance [7] is accounted for.

The tree is formed according to the following procedure. The PAN coordinator sends the beacon, and nodes that receive this packet could become level 1 nodes, which in turn transmit beacon packets to allow other nodes (level 2 nodes) to join the network. To balance the number of level 1 and level 2 nodes, we impose a maximum number of level 1 nodes. The maximum is set to a given percentage, denoted as p_1, of nodes in the network. This means that, being N the number of nodes in the network, if the number of nodes triggered by the PAN coordinator is larger than $N \times p_1$, $N \times p_1$ nodes will be randomly selected as level 1 nodes, whereas the remaining nodes will become level 2 nodes.

According to the Zigbee specifications, nodes work in beacon-enabled mode: each child node tracks the beacon of its parent and transmits its own beacon at a predefined offset with respect to the beginning of its parent beacon. The offset must always be larger than the parent superframe duration and smaller than beacon interval. This implies that the beacon and the active part of child superframe reside in the inactive period of the parent superframe: no overlap between the active portions of the superframes of child and parent is present. This concept can be expanded to cover more than two nodes: the selected offset must not result in beacon collisions with neighbouring nodes. Obviously, a child will transmit a beacon packet only in

case it is a router. Each child will transmit its packet to the parent in the active part (CAP or CFP) of the parent superframe.

We assume that all the active parts of the superframes generated by the routers and by the coordinator have the same duration (i.e., we set a unique value of SO). In these conditions, assuming to allocate the first part of the superframe to the PAN coordinator (for receiving data from level 1 nodes) once we set BO, the number of level 1 routers that will have a portion of superframe available for receiving data from their children will be equal to $2^{\mathrm{BO}-\mathrm{SO}} - 1$ [8]. If a larger number of level 1 routers is present, some of them will not have a portion of superframe available, their children cannot access the channel, and their packets will be lost.

4 Numerical Results

In this section, some examples of results that could be achieved through the simulation platform are shown. The simulator used is written in C and results are achieved by simulating 10,000 superframes (meaning 10,000 transmissions from sensors to the PAN coordinators).

Results are obtained by setting, if not otherwise specified, $k_0 = 40\,\mathrm{dB}$, $k_1 = 13.03$, $\mathrm{SO} = \mathrm{BO} = 2$ and $P_\mathrm{t} = 0\,\mathrm{dBm}$. No GTSs are used here. The scenario simulated is that of Fig. 2 when 36 nodes are present. The cases of lower number of nodes in the apartment are obtained by eliminating one or more nodes per room. For example, in case of 30 nodes, we have five nodes per room and we average results obtained by randomly changing the node eliminated from the different rooms.

The performance metrics considered are (i) the average delay affecting a transmission from a node to the coordinator; (ii) the packet error rate (PER), which is the probability that a packet transmitted by whatever node in the cell (or in the macrocell) is correctly received by the coordinator; and (iii) the average energy spent by a node per received packet.

The PER takes into consideration losses due to MAC and connectivity: a node is isolated if it does not receive any beacon coming from the PAN coordinator or level 1 nodes (in the tree topology case).

For what concerns the evaluation of the energy consumed by nodes, we assume that nodes spend energy when they receive, sense the channel, transmit, and do back-off. We set the energy spent to transmit a bit equal to $0.324\,\mu\mathrm{J/bit}$, the energy spent to receive or sense a bit equal to $0.39\,\mu\mathrm{J/bit}$, and the energy spent in back-off equal to $0.195\,\mu\mathrm{J/bit}$. These data are taken from Freescale devices data sheets [9].

The first four figures are related to the star topology case.

In Fig. 3, the PER as a function of the number of nodes per cell for different values of the packet size is shown. Here, we assume that nodes can perform three retransmissions of the same packet within the same superframe. We consider two different scenarios: one cell and two identical cells (the same of Fig. 2) put side-by-side. As we can see, an increment of the number of nodes competing for the channel and also of the packet size results in an increase of the PER. Two channel

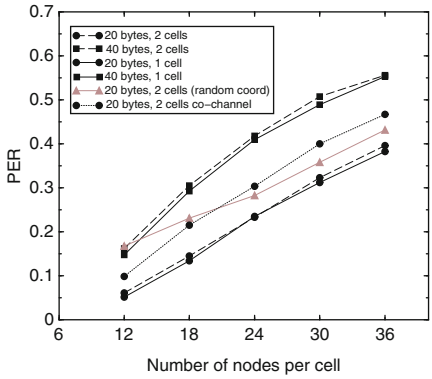


Fig. 3 The PER as a function of the number of nodes per cell, when one or two cells are present

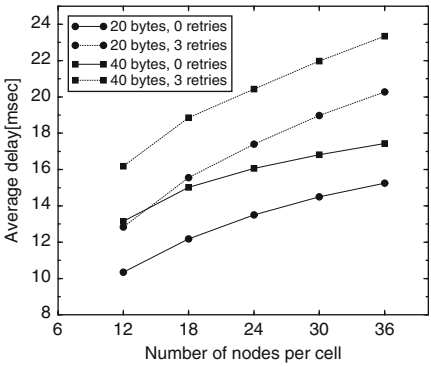


Fig. 4 The average delays as a function of the number of nodes in the single cell case

selection strategies are considered: (i) each coordinator randomly selects one of the 16 available channels and results are obtained by averaging over a large number of different realizations of the frequencies choice; (ii) the two cells work on the same channel (co-channel case). As we can see, in the first case the curves for one and two cells are approximatively overlapped. Whereas, in the second case, the PER notably increases in the two-cells case. In the figure, we also show results achieved by considering a random location of the PAN coordinators in the cells (see the curve: 20 bytes, two cells (random coord)). By averaging results over different positions of the coordinators, the PER increases, since in the previous case the coordinators were at large distance and less interferences were produced.

In Fig. 4, the average delay with which a packet coming from whatever a node in the PAN is received by the coordinator as a function of the number of nodes in the cell (single cell case) is shown. By increasing the packet size and the number of retries, the delay increases, as expected.

In Fig. 5 we show the average energy consumed by a node in the cell per correctly received packet. This means that we average the energy spent by nodes in the cell over the number of packets correctly received. As we can see, by increasing the number of retries and the packet size, the energy consumed increases. Moreover, the energy consumed increased by increasing the number of sensors in the cell, since

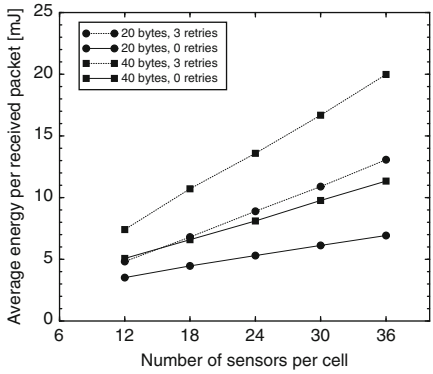


Fig. 5 The average energy spent per received packet as a function of the number of nodes in the single cell case

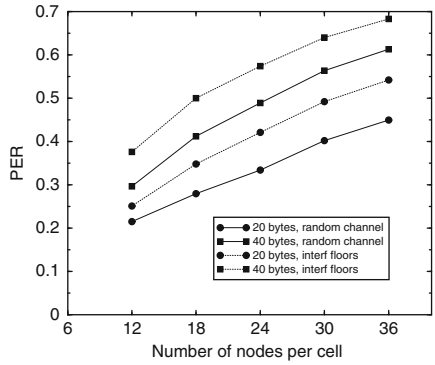


Fig. 6 The PER as a function of the number of nodes in the building case

the losses increase and on average, the network spend more energy per received packet.

In Fig. 6, a building formed by eight cells, four cells per floor (equal to that shown in Fig. 2), is considered. Two cases of channel selection are accounted for: (i) random selection of channels for all the eight cells; (ii) random selection of channels for the cells at the first floor and co-channel interference between cells at the two floor (i.e., a cell at the second floor uses the same channel used by the below cell). The figure shows the increase of the PER due to the presence of the interference coming from the second floor.

Finally, the three-level tree-based topology is compared with the star topology. Different setting of the parameters SO and BO and p_1 are considered. In all the cases shown, the tree topology improves performance in terms of PER. This is due to an improvement of the connectivity and also to the decreasing of the number of nodes competing for the channel. However, this improvement is achieved at the cost of larger delays. As far as the tree results, we can note that by increasing p_1 the PER gets larger, since increases the number of level 1 nodes competing for the channel. Moreover, performance improves by increasing BO since more level 1 routers have a part of the superframe allocated to receive data from their children.

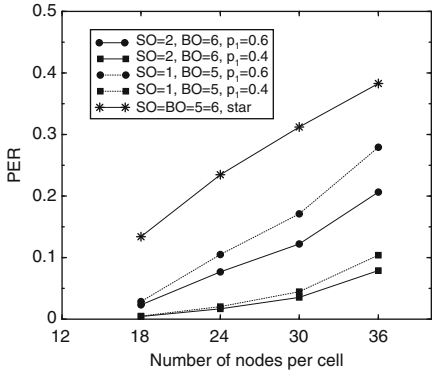


Fig. 7 The PER as a function of the number of nodes in the cell for the tree and star topology cases

5 Conclusions

In this chapter, we investigated a new and challenging application scenario for WSNs: energy efficient buildings realization. The reference scenario of the eDIANA project is reproduced and studied through simulation analysis. Results, in terms of packet error rate, average delay, and energy consumption are achieved through the developed tool. These results are the first tests on the applicability of the 802.15.4 technology to the eDIANA scenario, and they represent a first step toward the implementation of the wireless communication part of the eDIANA platform.

Acknowledgment This work was supported by the Artemis project eDIANA (contract no. 100012).

References

1. eDIANA, Artemis Project. http://www.artemis-ediana.eu/
2. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) A survey on sensor networks. IEEE Commun Mag 40(8):102–114
3. Rajaravivarma V, Yang Y, Teng Y (2003) An overview of wireless sensor network and applications. In: Proceedings of 35th Southeastern symposium on system theory, Mar 2003, pp 432–436
4. IEEE 802.15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs). IEEE, 2003
5. Zvanovec S, Pechac P, Klepal M (2003) Wireless LAN networks design: site survey or propagation modelling? Radioengineering 12(4):42–49
6. Katulski RJ, Lipka A (2007) Methodology of radio signal power distribution modeling for WLAN networks. In: Proceedings of IEEE international conference on computer as a tool, EUROCON 2007, 9–12 Sept 2007, Warsaw, pp 864–868
7. The ZigBee Alliance web site: http://www.zigbee.org/en/index.asp
8. Buratti C, Verdone R (2008) A hybrid hierarchical architecture: from a wireless sensor network to the fixed infrastructure. In: Proceedings of IEEE European wireless, EW2008, June 2008, Prague, Czech Republic

9. Freescale Semiconductor's MC13192 Developer's Kit
10. Alliance Z, Zigbee specifications. Zigbee Standard Organisation, 2008
11. Gutierrez J, Callaway E, Barret R (2003) Low-rate wireless personal area networks – enabling wireless sensors with IEEE 802.15.4. IEEE Press, New York

A Real Implementation and Deployment for Wine Production Management Based on Wireless Sensor Network Technology

Luca Bencini, Giovanni Collodi, Davide Di Palma, Antonio Manes, and Gianfranco Manes

1 Introduction

In the last decade, the advancements in communication technologies have contributed to establish a wide wired and wireless network to get people in touch all over the world.

Information can be spread everywhere on Earth in real-time, files can be shared from ones own desk, simply connecting to the World Wide Web, the "Internet of People": the great bet of connecting people was won.

Nevertheless, it was only the first important step toward a new era of communications: the new and higher stake problem of integrating different networks between smart system and sub system arises; in the next future, human-to-human interactions will be only a small part of a greater communication infrastructure that will comprise exchanges between man and things and between things and things (and system and among different system).

The increasing request to live in a smart environment, where everything is under control and also self-controlling, will lead to the development of smart systems that can communicate, self-organize, analyse specific environmental parameters, and to actuate self-made decisions onto the surrounding environment.

A self-controlled environment is still far from the realization, but the Ambient Intelligent (AmI) Paradigm takes a step toward it and toward the "Internet of Things" Concept: the main idea is to get simple information distributed in the surrounding environment, using smart sensors, fusing and mining data to allow the final user to have a wrapped but reliable and detailed view of changing phenomena and to let the end-user make decisions and control the actuators, simply connecting to the Web, wherever in the world.

L. Bencini (✉), G. Collodi, D. Di Palma, A. Manes, and G. Manes
Department of Electronics and Telecommunications, University of Florence,
via di Santa Marta 3, 50139 Firenze, Italy
e-mail: luca.bencini@unifi.it; giovanni.collodi@unifi.it; davide.dipalma@unifi.it;
antonio.manes@unifi.it; gianfranco.manes@unifi.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_33,

This chapter aims at describing the implementation of the AmI Paradigm within the EU Integrated Project FP6-IST-1-508774-IP "GoodFood" based on the successful integration and smart cooperation of different systems.

The chapter provides comments and results coming from the five deployed pilot sites, about two million data stored, and 2 years of experience in the real case study such as the wine chain production.

The chapter is organized as follows: in Sect. 2, the AmI Paradigm and the Precision Agriculture Concept are described; Sect. 3 proposes a real application (the wine production chain) that joins the AmI theory to a true case study, where distributed sensing can become a keystone to improve quality in wine production.

In Sect. 4, the Wireless Sensor Network System (WSNS) developed to carry into effect the AmI paradigm is presented.

Section 5 describes the developed and installed pilot sites in the vineyards and provides the results obtained from a detailed and distributed monitoring system, showing the great chances that a Wireless Sensor Network System can give toward the "Internet of Things" era.

2 Ambient Intelligence and Precision Agriculture

Ambient intelligence involves the convergence of several technology areas: the ubiquitous or pervasive communications, computing, and intelligence interfaces, the last being combined with intelligent equipment provides intelligence interaction with their surroundings and with other equipment.

Its major contribution is the development of various ad hoc networking capabilities that exploit highly portable or else numerous, very-low-cost computing devices. Another key area is intelligent systems research, which provides learning algorithms, pattern matchers and situation assessment.

A third essential element is the interaction of objects in the environment.

There are many scenarios that attempt to give substance to what this research might produce: they span from the application of travel and tourism, and the in-home or office automation application to the environmental and structural monitoring. The AmI Concept applied to a selected scenario foresees that data collected by sensors are locally pre-processed and analyzed, and then remotely stored in a database, using the network gateway unit. Local user, through web-based graphical interfaces, can then query the remote database gathering either general information or single sensor parameters. The basic idea of the data fusion and the data aggregation represents a network level of intelligence that allows a more effective fruition of the information and allows a first degree of knowledge discovery. User-friendly tools for executing complex operations, such as comparing cross-related parameters and historical values, are also available. Moreover, the user interfaces are specifically arranged and presented according to the user skills and requirements, i.e. presenting raw data to the agronomist or biologist expert, or graphical interactive interface for the non-skilled end user. Data access could be made using a 2-D representation

of the controlled area and sensors installed, providing direct information on the network state and monitored values, highlighting anomalous detections.

User can also directly interact with the sensor network by using an handheld device (or similar), implementing local network specific communication protocols and networking interface. In this way he can easily access the monitored parameters.

The Ambient Intelligence Paradigm is fully oriented to provide the end user a new approach toward the environment: in accordance with AmI Paradigm the environment provides the end user distributed and detailed information to continuously monitor and rapidly react to the changing phenomena, thus preventing harmful situations.

In this vision, Precision Agriculture is entirely integrated with the so-called emerging concept of "zonation": "zonation" is a key goal in seeking quality improvement; it was developed by producers as a fulfilment of their needs to operate on a known area with defined pedological and climatic characteristics. This kind of "zonation" is called agro-ecological zonation, since it determines areas with specific soil and climate peculiarities. Agro-ecological "zonation" can be used as a starting point for a methodology of land inventory and monitoring, as it shows uniform portions of earths surface as another synonym of quality in agriculture: proving that certain products have undergone a well-defined production protocol is one guarantor of quality for the consumer.

In order to reach these two irremissible targets, agriculture needs to combine tradition and innovation, introducing systems that are able to give substance to "zonation" and to guarantee traceability, furnishing distributed, reliable and timely data, and friendly end-user tools to understand the phenomena and allow immediate and appropriate reactions.

3 A Killer Application Example: Wine Production Monitoring

There are many interesting areas in agriculture where the quality of the final product and its traceability are becoming a key issue, both for the producer and for the consumer; one of the most attractive areas is the wine chain, since deep research in high quality and in traceable production procedures is carried on, and is also related to conspicuous capital investments by many producers all over Europe (and world), and this is why it is was chosen as a "belonging to realities" case study. The wine production chain can represent the first example of tradition and innovation conjugation, where the irreplaceable experience of the farmer can be supported by indispensable information provided by a distributed, high precision monitoring system, in other words an "IoT System."

One of the most important steps to guarantee the quality of wine is grape production, which depends on many different factors (environmental, meteorological,...) and is not related to objective measurements and parameters but more often simply related to past experience.

It follows that a scientific method to analyse vineyard status and to foresee possible long-term problems could be a valid solution to drive grapes growth toward high quality and to create an objective production protocol to replicate high quality wine year after year.

It means that a vineyard could be analyzed in a continuous and detailed way during all the farming seasons, to recognize and to highlight differences ("micro-zonation") in order to fit the treatments (water and pesticides) by applying them only when it is necessary; as a result, together with quality, there are cost benefits due to a decrease in input treatments and work.

The quality of the entire production will be higher, even if differences persist between zones, due to different soil conditions and sun exposure, but this diversity could be pointed out and taken into consideration during harvesting to divide grapes into different fermentation tanks, thus obtaining various certified quality wines.

It follows that even traceability can be guaranteed, since a deep knowledge of the growth and harvest of grapes can be an important instrument to prove that a certain wine is the result of a high profile production, respecting all protocols and regulations for denomination of origin.

4 The Wireless Sensor Network System

If the aim of the "Internet of Things concept" is to "*connect everyday objects*," the RFID sensors can represent a valid solution to spread the investigation at a very deep level; nevertheless to allow the analysis to cover wide areas, RFID sensors must join other communication technologies, thus implementing Ambient Intelligence, but not all the available technologies are really capable of putting it into reality, since the following inevitable features must be satisfied:

- A capillary but distributed, continuous monitoring of the parameters of interest must be guaranteed.
- The acquisition system must be reliable and provide data for a long term.
- Data should be available to the end user always and everywhere.
- A Smart Web Interface should provide raw and detailed data to expert users, and furnish an easy and informing aggregation of them to the standard user.

These four features are the pillars of an Ambient Intelligence System.

This means that identifying a technology to get data from somewhere or something is important, but it is not enough; the acquiring system must be part of a complete chain, from the sensor to the web tools interface, thus building a system close to the "Internet of Things" Concept.

At this aim, the Wireless Sensor Network technology has revealed to be a wise and suitable choice to fulfill the above mentioned requirements, since it is an accurate, reliable, space-based solution, combining wide monitored area coverage [1].

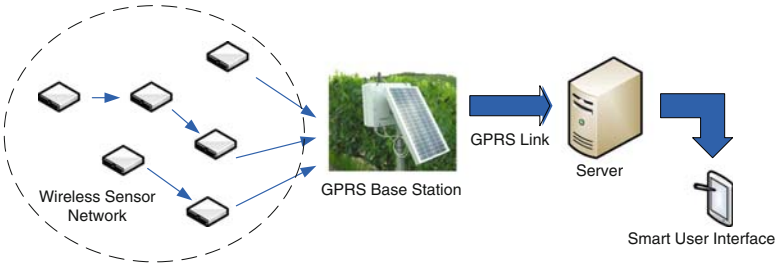


Fig. 1 The wireless sensor network system: block diagram

At present, the communication system that we have developed is based on a Wireless Sensor Network using wired sensors, but in the next future the further step of introducing wireless, RFID-Driven sensors will give the entire system a greater flexibility, physically splitting the communication network and the sensing system.

Nevertheless, the implemented Wireless Sensor Network System (WSNS) allowed us to put into practice the AmI Paradigm: it is comprised of Wireless Sensor Network technology, a GPRS Base Station, a Server, and a Smart User Interface [2] (Fig. 1).

Data collected every 15 min from sensors, placed on wireless nodes, are transmitted through the wireless network, following a weighted routing algorithm, to a master node, located on the GPRS Base Station and suddenly forwarded through the GPRS link to a server, using the TCP/IP protocol and stored.

Data can then be queried by the end user and analyzed using the Smart User Interface.

The commercially available sensors can monitor all the interested parameters of a growing plant:

- Soil moisture
- Soil temperature
- Trunk diametric growth
- Differential leaf temperature
- Air temperature
- Air humidity

and each wireless node can manage up to 16 different sensors, providing the possibility to deeply investigating the surrounding environment.

Each wireless node, protected in a hard mechanic and waterproof box, is battery powered and with a consumption of 30 μA in power saving mode can work for more than 16 months.

The energy management is optimized by a custom MAC layer protocol, called STAR MAC, that guarantees both energy saving and a high network reactivity and synchronization between nodes [3, 4]. In particular, it fulfills power saving requirements, due to the introduction of a duty cycle, along with the advantages provided by the offset scheduling, while avoiding penalties in signaling overhead. According to

the STAR MAC protocol, each node can be set either into an idle mode, or in an energy saving sleeping state. The transitions between states are synchronous; accordingly, a duty cycle function can be introduced. To provide full communication capabilities, all the nodes need to be weakly synchronized, i.e. they are aware at least of the awakening time of all their neighbors. According to this scheme, a node sends, frame by frame, one synchronization message to each of its neighbor nodes known to be in the listening mode (*Synchronous Transmission*). During the set up phase, when each node is discovering the network topology, the control messages are asynchronously broadcasted, while its neighbors periodically awake and enter the listening state independently (*Asynchronous Reception*).

The improved 868 MHz RF section in the proprietary hardware platform adopted (MIDRA Mote) [5] provides a wide area coverage, up to 170 m in open field.

The multihop algorithm running on nodes allows to extend the total network area coverage, minimizing data losses thanks to the weighted choice of the best path to reach the destination. In particular, we refer to a proactive algorithm belonging to the class of link-state protocol that enhances the capabilities of the Link Estimation Parent Selection (LEPS) protocol. It is based on periodically sending a control message to neighbour nodes to carry on information needed for building and maintaining the local routing table.

Data forwarded to the master node, on the GPRS Base Station, are temporarily stored in a local queue and then transmitted to the GPRS board by means of a serial bus and forwarded through GPRS link to a server, placed at CSIAF (Centro Servizi Informatici Ateneo Fiorentino).

The GPRS Base Station is powered by a 12 V battery, recharged by a 20 W solar panel, and can run unattended for years.

At the end of the chain, the Smart User Interface provides end user an ensemble of tools to best appreciate information coming from sensors: data can be shown in a raw way, or plotted (i.e. Fig. 2), and exported in different kinds of files, or aggregated, using high qualified mathematic models that sum up various data into few but important macroparameters (Fig. 3), such as physiological activity of plants, field water management, and pest management.

Even if a very complex and heterogeneous system could appear fault intolerant, the strict anti-blocking measurements adopted will make the Wireless Sensor Network System extremely reliable: the most exposed parts to problems and faults such as nodes and the GPRS Base Station improved with recovery strategies algorithms, in order to restart operations in case of deadlocks.

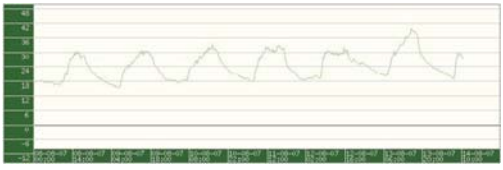


Fig. 2 Plotted data in a chosen period (temperature sensor)

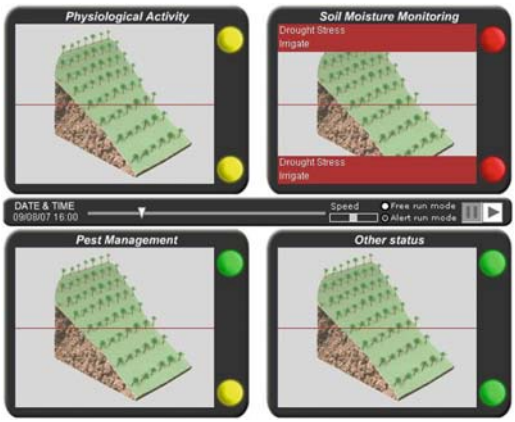


Fig. 3 Aggregate data models presenting few macroparameters

In addition, the wireless sensor network results very scalable and flexible, since the removal of a node does not affect other nodes that, in case, have only to change paths to reach the master node, based on the multihop algorithm.

5 Pilot Sites

The pilot sites developed and installed within the EU Integrated Project FP6-IST-1-508774-IP "GoodFood" can be considered in every respect as one of the first examples of Ambient Intelligence Paradigm declination into the real world.

The idea of collecting simple data (from commercially available sensors) and using them to analyze more complex parameters, using mathematic models resulted successful and can now be appreciated on the web: http://www.unfi.it/midra/goodfood.

The first pilot site was deployed in November 2005 in a hill vineyard of the Montepaldi farm, in Chianti Area Tuscany (Italy), owned by the University of Florence: with 11 nodes and 35 sensors distributed on 1 ha area and a GPRS Base Station running unattended since 21 months, this pilot site represents the first successful attempt to densely monitor a vineyard and its environment.

More than two million data were collected since the deployment, during all seasons, providing a distributed but a detailed knowledge of changing phenomena in the vineyard.

Soil moisture positioned at different depths in the vineyard allows to verify if summer rains run off on the soil surface or seep into the earth and provoke beneficial effects on the plants: this can be appreciated with a rapid view by using the soil moisture aggregate report, representing moisture sensors at two depths and coloring in green tones the moisture differences (Fig. 4).

Stress conditions on plants, due to dry soil and/or to hot weather can be highlighted by the accurate trunk diametric growth sensor, that with microresolution can

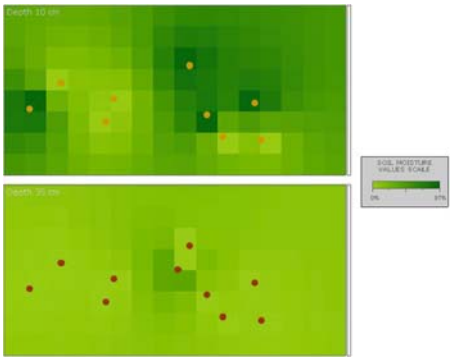


Fig. 4 Soil moisture aggregation report: the *upper map* represents soil moisture at 10 cm in the soil and the *lower map* represents soil moisture at 35 cm in the vineyard of Montepaldi on the 4th of August 2007, after a slipping rain

Fig. 5 Diametric growth diagram

follow each minimal variation of the trunk, giving important information on plant living activity (Fig. 5).

The fine and detailed tools to study single data from the vineyard are supplemented with more friendly instruments based on mathematical models analyse the entire amount of data in a selectable period and provide ready-to-use information on three macroparameters:

- Vineyard water management
- Plants physiological activity
- Pest management

A Flash Player application applies models to incoming data using cross light colors for each parameter in different parts of the vineyard (currently two), which points out normal (green), mild stress (yellow), or heavy stress (red) (Fig. 6) conditions, and it provides suggestions to the farmer to apply pesticides or water in certain parts of the vineyard.

Nowadays five pilot sites are running unattended in different vineyards in Europe, and they are collecting data creating a wide database, useful to investigate differences between cultivation procedures, environments, and treatments.

Several researchers in agronomics all over Europe are allowed to query whether database can run experiments and apply new mathematic models to real collected data.

One pilot site is deployed in a farm in the Chianti Classico area, where at about 500 m above sea level in a stony hill area of 2.5 ha, ten nodes and 50 sensors are monitoring the environmental variations in "terroir," since July 2007 producing one of the most appreciated wines in the world.

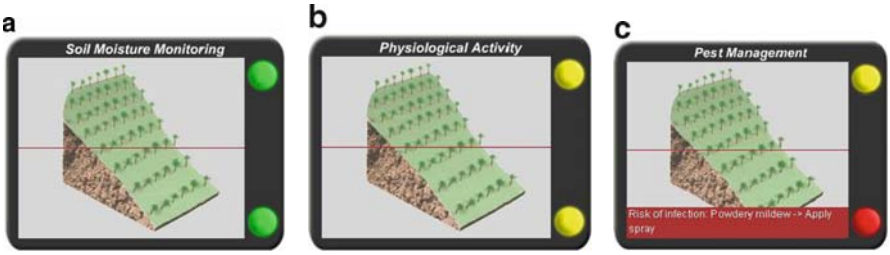


Fig. 6 (**a**) Normal conditions, (**b**) mild stress conditions, (**c**) heavy stress conditions

Another pilot site is deployed in a prestigious farm of the Chianti area near Florence, where the soil is more plane and clayey, and data are coming from 50 sensors distributed in 4.5 ha.

The fourth Wireless Sensor Network System is installed in South France, in the vineyard of Pech Rouge, in Gruissan: the hill vineyard is close to the seaside and in one rocky hectare 68 sensors are collecting data: this high sensor density was chosen to guarantee redundancy to measure and to provide a deep knowledge of phenomena variation, in a experimental vineyard where "microzonation" is applied and where water management experiments are performed to study the reactions of plants and quality of grapes.

The fifth pilot site is a small sensor network in the cellar of the Montepaldi farm: it is aimed at monitoring cellar conditions, where wine is stored in barriques and bottles, providing data on air temperature and humidity and on light radiation, in order to complete the wine production chain monitoring and to demonstrate the versatility of the Wireless Sensor Network System adopted.

Summing up, since July 2007 three new pilot sites were added at the existing one in the Montepaldi vineyard and cellar, providing a total monitored area of about 9 ha in different soil conditions, with 49 nodes and more 200 sensors deployed.

6 Conclusions and Further Developments

According to the International Telecommunication Unit (ITU Report 2005) Internet of Things can be defined as a vision "*... to connect everyday objects and devices to large databases and networks (using) a simple, unobtrusive and cost-effective system of item identification....*"

The shown Wireless Sensor Network System is entirely set into this vision, since it is a valid solution to monitor common parameters using simple, unobtrusive, commercial and cheap sensors, forwarding their measurements by the means of a heterogeneous infrastructure, consisting of wireless sensor network technology, GPRS communication and ordinary internet data transfer (TCP-IP protocol).

Data coming from sensors are stored in a database that can be queried by users everywhere in the world, only using a laptop or a PDA: the smart user interface also allows to read and analyze data in an easy way.

The adopted system revealed to be capable of putting the Ambient Intelligence Concept into reality since it guarantees capillary but distributed continuous monitoring of the interested parameters, it proves to be reliable and able to provide data for long term and it provides raw and detailed data representation to the final users.

The presented pilot sites are a first, important example of the deep and detailed information provided by such a system and they show the great chances that the declination of the "Internet of Things" Concept to a real application can give to human life and the added value to environmental understanding.

Acknowledgement This work was supported in part by the EU Integrated Project FP6-IST-1-508774-IP "GoodFood."

References

1. Chiti F, Fantacci R (2006) Wireless sensor network paradigm: overview on communication protocols design and application to practical scenarios. EURASIP Newslett 17(4):6–27
2. Report on existing test beds and platforms. CRUISE WP122, Deliverable D122.1, Jan 2008
3. Chiti F, Ciabatti M, Collodi G, Di Palma D, Fantacci R, Manes A (2006) Design and application of enhanced communication protocols for wireless sensor networks operating in environmental monitoring. In: IEEE ICC06, Istanbul, Turkey, June 2006, vol 8, pp 3390–3395
4. Manes G, Fantacci R, Chiti F, Ciabatti M, Collodi G, Di Palma D, Manes A, Nelli I (2007) Efficient MAC protocols for wireless sensor networks endowed with directive antennas: a cross layer solution. In: EURASIP JWCN special issue on cross-layer optimized wireless sensor networks, vol 2007, Article ID 37910
5. Mattoli V, Mondini A, Razeeb KM, OFlynn B, Murphy F, Bellis S, Collodi G, Manes A, Pennacchia P, Mazzolai B, Dario P (2005) Development of a programmable sensor interface for wireless network nodes for intelligent agricultural applications. In: Proceedings of IE05, June 2005

Performance Evaluation of an IEEE802.15.4 Standard Based Wireless Sensor Network in Mars Exploration Scenario

R. Pucci, E. Del Re, D. Boschetti, and L. Ronga

1 Introduction

The interest in wireless communication in space missions has been growing up in the last years, driven by the complexity of new space systems and supported by the technological trend for ground commercial utilization that makes available standards and products based on RF. Potential space applications for RF wireless systems are numerous, such as planetary surface exploration, intra-satellite devices communication, and extra-vehicular operations.

About planetary exploration missions, after the failure of the Beagle II Mars mission, the research community has focused its attention on redundant systems, such as WSN, in which many small sensors are deployed, where, even if some sensors should fail, the mission would not. The IEEE 802.15.4 standard (ZigBee) provides low-cost and low-power connectivity for WSN devices that need a month's or a year's duration of battery, with low data rate and small dimensions. In order to valuate ZigBee usability in planetary exploration context, a set of possible scenario is studied, referring to a realistic scenario of Mars. Therefore, thanks also to the recent information given by Phoenix Mission, in this paper different Martian radio frequency channel models are used; such models take into account all Martian features such as Tropospheric effects, clouds, wind, snow, gaseous attenuation, and dust storm effect. The knowledge about such channels allows investigating system transmission and network performance, thanks to simulation process.

2 Related Work

In order to obtain a characterization of the channels used by WSN and its performance, we follow the block diagram shown in Fig. 1.

R. Pucci (✉), E.D. Re, D. Boschetti, and L. Ronga
Thales Alenia Space Italia, University of Florence, Florence, Italy
e-mail: renato.pucci@unifi.it; puccirenato@gmail.com

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_34,

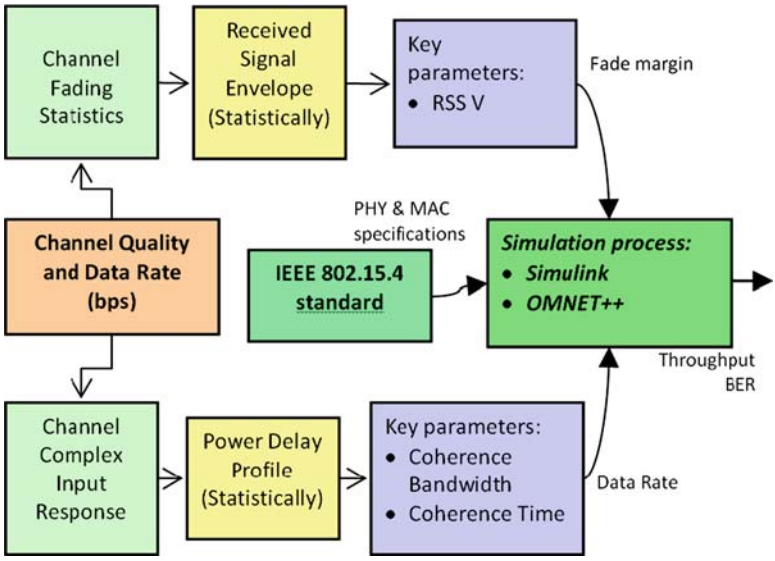


Fig. 1 Analysis process used to obtain WSN performance in planetary exploration context

Thanks to the analysis of the complex input response and fading statistics, we obtain a quality and data rate estimation of the channel; therefore, such parameters are used together with PHY and MAC specifications of the standard to execute simulation processes in Simulink® (to obtain Bit Error Rate (BER) performance of an OQPSK transmission system) and OMNET++ (to obtain an estimate of the network performances).

3 Martian Channels Characterization

Considering the planetary exploration scenario of Mars, there are several features affecting the propagation of radio waves such as the presence of clouds, snowfalls, temperatures, gaseous composition of the Martian atmosphere (different from that of Earth's atmosphere), presence of obstacles (rocks and craters), and possible occurrence of dust storms. Among them, the last two features offer major contributions to the radio wave attenuation and, furthermore, such contributions vary according to the Martian region of interest.

Taking into account different surface morphology scenarios (see Fig. 2) in terms of rock dimension and density, we consider two different channel modes:

- "Normal" channel (Low-medium rock density and dimension),
- "Rocky" channel (High rock density and dimension).

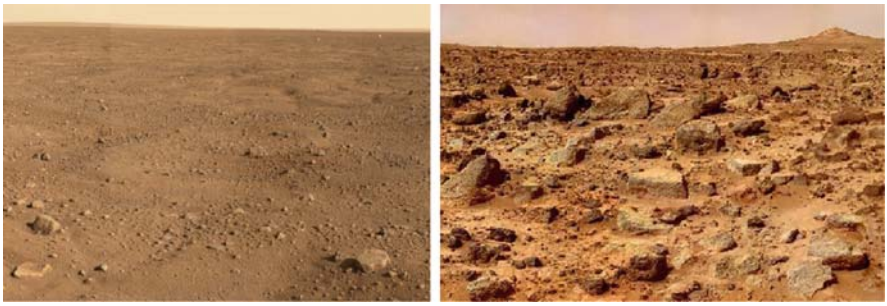


Fig. 2 Different rock density regions of Mars, made by "Phoenix" and "Viking1" landers

In particular, considering the possible wave propagation paths (LOS and NLOS) and the multipath components, "Normal" channel can be described statistically as a Ricean channel, with a Rice Factor $k = 10$ (ratio between LOS and NLOS component powers) and a probability density function given by:

$$p(E) = \frac{E}{\sigma^2} \mathrm{e}^{-\frac{E^2+A^2}{2\sigma^2}} I_0\left(\frac{EA}{\sigma^2}\right) \tag{1}$$

where the parameter A denotes the peak amplitude of the dominant signal, $I_0(\bullet)$ is the modified Bessel function of the first kind and zero-order, and σ^2 is the time-average power of the received signal.

"Rocky" terrain scenario is characterized, thanks to the frequency distributions obtained from data collected by Viking 1 and 2 landers; such distributions give information about dimensions and the number of rocks per m^2 of the Martian rocky landing sites. Taking into account such information and comparing sensor and rock dimensions, "Rocky" channel can be described statistically as a Rayleigh channel, with a probability density function given by:

$$p(E) = \frac{E}{\sigma^2} \mathrm{e}^{-\frac{E}{2\sigma^2}} \quad E \geq 0 \tag{2}$$

Dust storms occur primarily in the south hemisphere of Mars, but sometimes they can cover the whole planet. Since they differ in dust particle density (from $N_\mathrm{T} = 1 \times 10^7 \mathrm{m}^{-3}$ to $N_\mathrm{T} = 8 \times 10^7 \mathrm{m}^{-3}$) [1] and wind strength (from 2 to 28 m/s), we consider three different channels:

- "Faint Dust Storm" channel (Low particle density, faint wind, $\overline{r} = 1\,\mu\mathrm{m}$)
- "Strong Dust Storm" channel (Medium particle density, strong wind, $\overline{r} = 10\,\mu\mathrm{m}$)
- "Heavy Dust Storm" channel (High particle density, heavy wind, $\overline{r} = 20\,\mu\mathrm{m}$)

Referring to [2], the attenuation caused by a sand storm could be obtained by:

$$A(\lambda) = \frac{1.029 \times 10^6 \varepsilon''}{\lambda[(\varepsilon' + 2)^2 + \varepsilon''^2]} N_T \bar{r}^3 \quad [\text{dB/km}] \tag{3}$$

where λ is wavelength in meters, N_T is the total particle density in $\#/m^3$, ε' and ε'' are the real and imaginary part of the dielectric permittivity index, and $\bar{r}$ is the mean particle radius in meters, obtained through an integration over all sizes of particles in the normalized particle number density $N(r)$. At the present time, there are no accurate measurements of the mean particle radius, but it can be considered included in $[1\,\mu\text{m}; 20\,\mu\text{m}]$.

4 BER Performances

For every five channel model above described, we obtain the BER performances of OQPSK modulation (used in IEEE 802.15.4 standard) for different distances of transmission. The results shown in Fig. 3 and Fig. 4 are obtained simulating the transmission system with a Simulink® based model, varying parameters of "Channel" block according to different channel simulation.

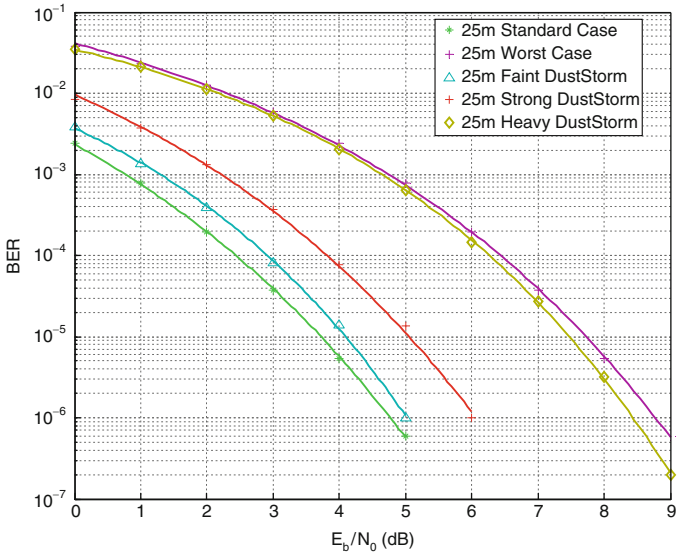


Fig. 3 BER trends comparison among different channels, considering a 25 m distance transmission and OQPSK modulation

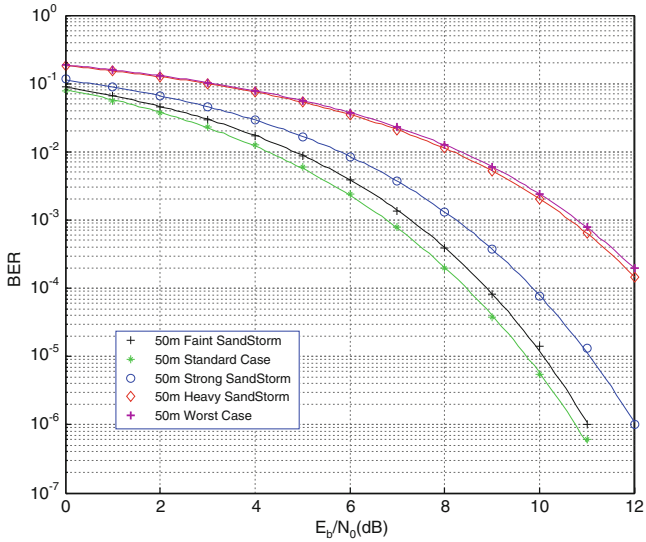


Fig. 4 BER trends comparison among different channels, considering a 50 m distance transmission and OQPSK modulation

5 SER Performances

According to the IEEE 802.15.4 standard for transmission at 2.4 GHz, we include in the Simulink model a DSSS spreading technique; thanks to it, we obtain a performance improvement in terms of Symbol Error Rate (SER), as shown in Fig. 5.

6 Packet Level Coding Evaluation

In case of degraded or fading channel, the utilization of packet level coding techniques could appreciably improve the network performance. Taking into account the limited computational capability and the low power consumption requirements for every sensor, we consider the implementation of cyclic Bose-Chaudhuri-Hocquenghem (BCH) coding technique with a limited number of redundant bits. In Table 1 [3] are reported the most suitable BCH-code parameters for the context we are considering. The utilization of transmission coding could be optional: whenever the "link quality indicator" (LQI, see [4]) detects a degradation of the transmission channel characteristics, the WSN coordinator informs the other WSN sensors to switch in encoded transmission mode. Such command could be simply transmitted in the beacon; in this way, retransmission could be quite reduced.

In Fig. 6 are shown the BER performance of a ZigBee network implementing BCH coding technique for different coding rates. The results are obtained, thanks to Simulink simulation processes.

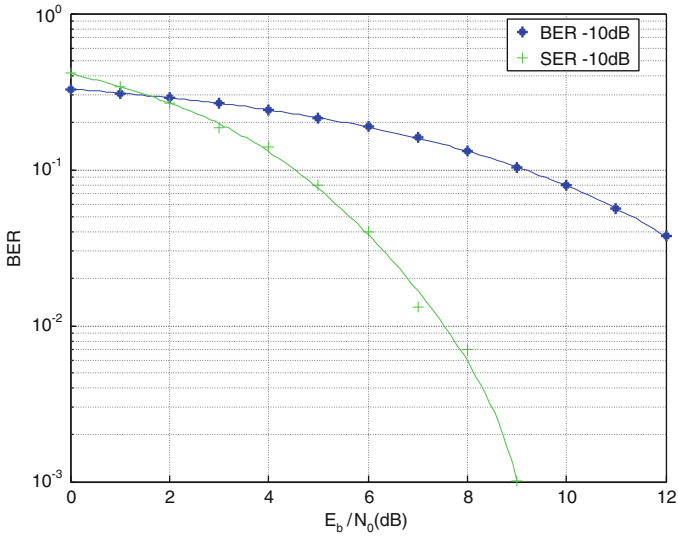


Fig. 5 SER vs. BER trends for a Normal channel considering a 25 m distance transmission

Table 1 Some coefficients of generator polynomials for BCH codes (n, k)

n	k	t
7	4	1
15	11	1
	7	2
	5	3
31	26	1
	21	2
	16	3
	11	5

7 OMNET++ Simulations

7.1 Used Model

Thanks to OMNET++, an open source discrete event simulation system, we realized a transmission system of a IEEE 802.15.4 network, formed by a mobile rover and 40 sensors, randomly deployed in a 100 m^2 area. Every sensor is defined as a compound element, since the architecture of every sensor is composed by modules (see Fig. 7).

"Physic" block is used to create, maintain, and delete the transmission links, "MAC" block checks the MAC header of every packet, "AODV" block represent the routing algorithm, "Application" is used to describe the type of the collected data (i.e. temperature, pressure, seismic activity, etc), "Battery" block simulate the

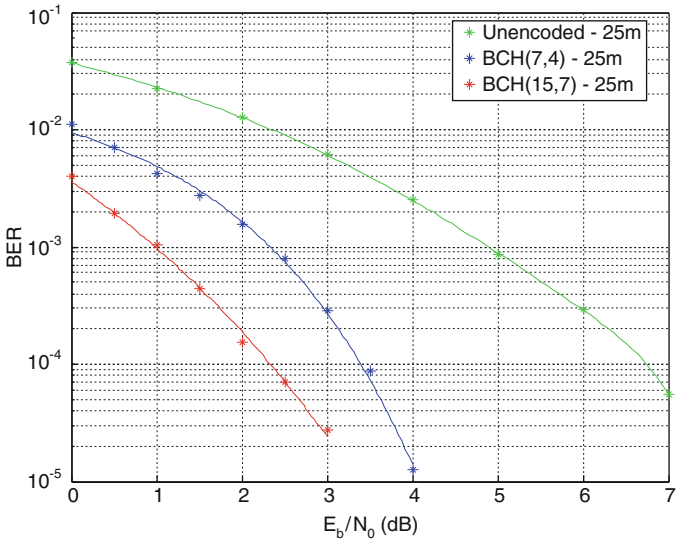


Fig. 6 Comparison among BER trends of uncoded and BCH-coded transmissions for a "Heavy" dust-stormy channel, considering a 25 m distance transmission

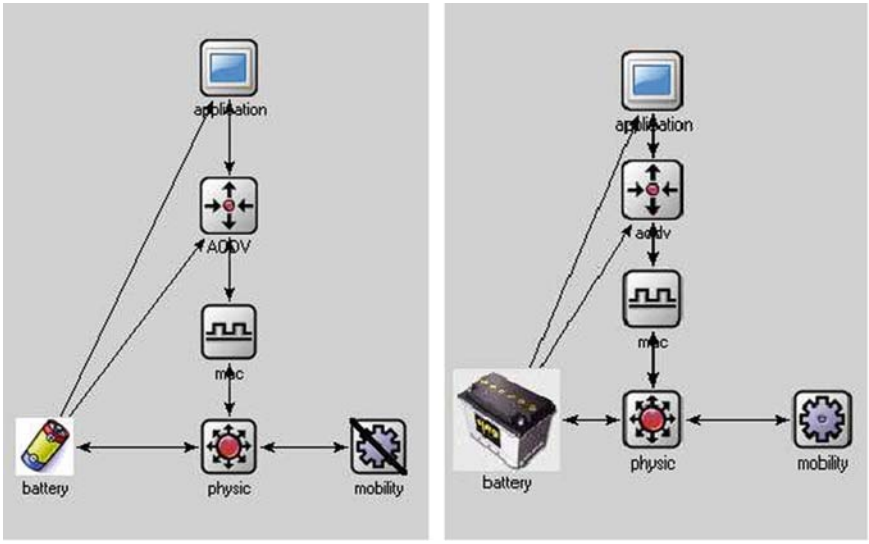


Fig. 7 Sensor (*left side*) and rover (*right side*) architecture in OMNET++

battery life of the sensor (it decreases for every sent message), and "Mobility" block defines the direction and the velocity of the rover. Taking into account the possibility of sensor damage or transmission inability caused by a bad deployment, we consider a sensor loss equal to 10%.

Referring to IEEE 802.15.4 physical specifications for OQPSK modulated transmission at 2.4 GHz, we consider an uncoded Bit Error Probability given by:

$$P_{\mathrm{e}}\left(\mathrm{SNR}_{20\mathrm{m,dB}}, d\right) = \frac{1}{2} erfc\left(\sqrt{[\mathrm{SNR}_{20\mathrm{m,dB}} + L_{20\mathrm{m,dB}} - L(d)_{\mathrm{dB}}]_{\mathrm{lin}}}\right) \quad (4)$$

where $\mathrm{SNR}_{20\mathrm{m,dB}}$ is the reference SNR term at the receiver placed at 20 m from the transmitter (in dB), and the $[]_{\mathrm{lin}}$ operator represents the linear conversion of the argument. The path loss terms are obtained considering:

$$P_{\mathrm{RX}} = \frac{P_{\mathrm{TX}}}{L(d)} \quad (5)$$

where

$$L(d) = \left(\frac{4\pi d}{\lambda}\right)^3 \quad (6)$$

$$L_{20\mathrm{m}} = L(20) \quad (7)$$

Since the reference scenario is a terrain with a medium/high density of rocks, we consider a third order exponent for the Path Loss as experimental approximation for terrains with numerous scatterers. Imposing the Bit Error Probability maximum equal to 4.8×10^{-5}, we obtain that 20 m transmissions are permitted at least.

7.2 Simulation Results

In order to evaluate the variation of the network performance during a 24-h transmission, we establish that on time $t = 6$ h of the simulation, and a 12-h sandstorm occurs, as shown in Fig. 8. The intensity of the sandstorm depends on $\mathrm{N_T}$ and the attenuation is calculated with (5).

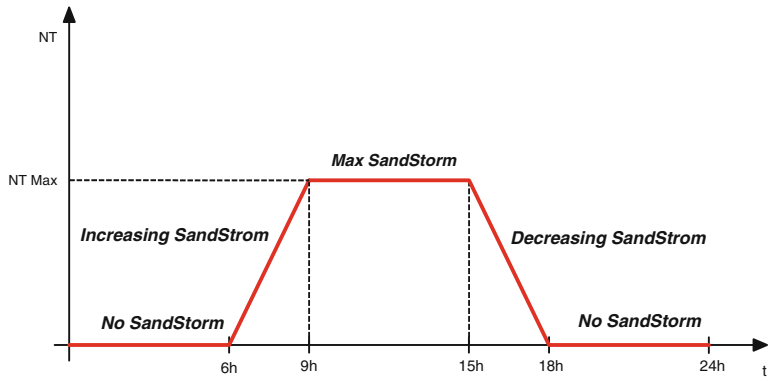


Fig. 8 Simulation time sequence: occurrence of a sandstorm

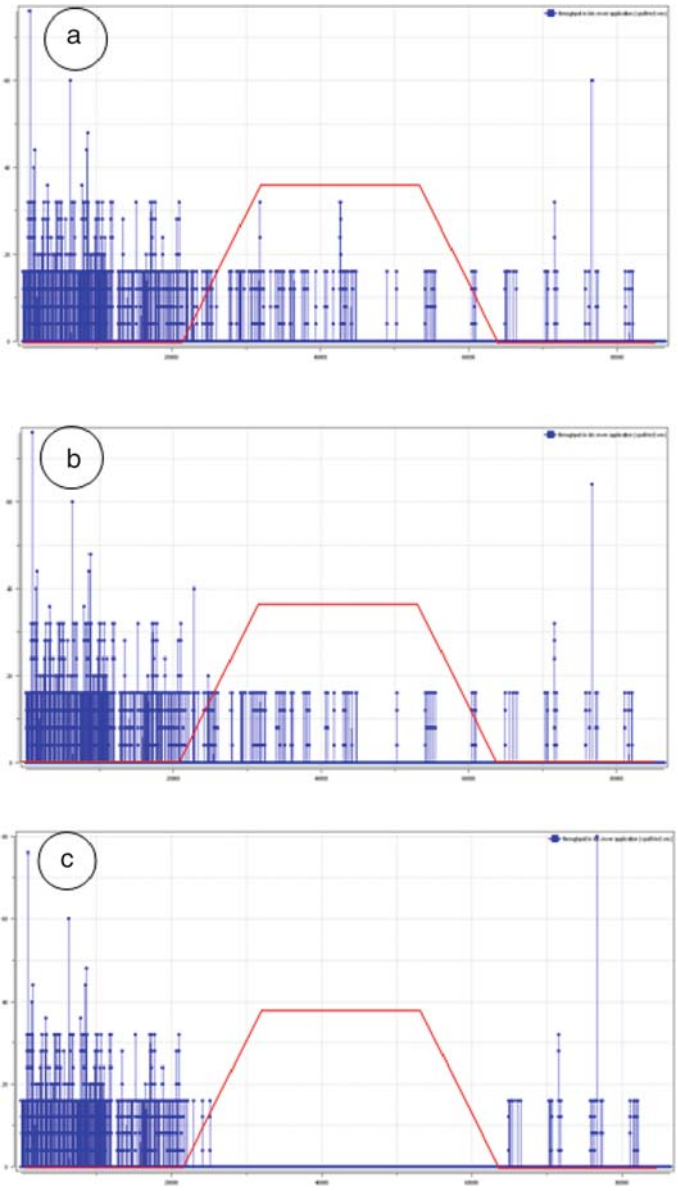


Fig. 9 Throughput measured by the Rover in case of: "Faint" sandstorm (1), "Strong" sandstorm (2), "Heavy" sandstorm (3)

In line with the above-mentioned Martian channel in the presence of a dust storm, we consider three different cases of simulation. Figure 9 shows the simulation results; a clear throughput performance degradation is appreciable, according to the increase of the sandstorm intensity. In case of "faint" and "strong" sandstorm

throughput performance makes lower, but network still works. Nevertheless, for a "heavy" sandstorm, no transmission is possible and network turns off.

8 Conclusions

In this chapter, we consider the opportunity to use an IEEE 802.15.4 standard-based network in Martian planetary exploration context. Thanks to an evaluation of the main features of Mars, we define five different channels that model the five most common propagation contexts. In order to obtain the channel data rate and availability, a capability analysis is performed, considering the power delay profile, the received signal envelope, and the measure of channel RF traffic. For the mentioned channels, we evaluate BER and throughput performances of a WSN, performing 24 h-long simulation campaigns.

Considering a network formed by a mobile rover and 40 sensors, the results obtained from simulations demonstrate that an IEEE802.15.4 based WSN can be used in planetary exploration context. Such WSN works pretty well, also in case of transmission within terrains with high density of rocks.

In the case of sandstorm occurrence, network performances degrade proportionally with the storm intensity and dust particle dimensions. For faint dust storm, WSN works well; in case of heavy dust storm, WSN performances become critical. Nevertheless, the WSN is able to react positively after the cessation of the perturbation.

The results shown in this chapter demonstrate that WSN should be used in future missions of planetary exploration; a test campaign should follow in order to validate simulated and predicted data.

Acknowledgments We would like to thank Francesca Paradiso and Jacopo Pesci for OMNETT++ simulations results.

References

1. Goldhirsh J. (1982) A parameter review and assessment of attenuation and backscatter properties associated with dust storms over desert regions in the frequency range of 1 to 10 GHz. IEEE Trans Ant Propaga AP-30:1121.
2. Chu TS. (1979) Effects of sandstorms on microwave propagation. Bell Syst Tech J 58(2):549.
3. ANSI/IEEE Std 802.11 Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 1999 Edition (R2003).
4. IEEE 802.15.4-2006, "Part15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs)", Sept. 2006.

The PECES Project: Ubiquous Transport Information Systems

Antonio Marqués and Manuel Serrano

1 The Challenge

The dramatic growth of the amount of information that is made available through computer systems and the increasing need to access relevant information anywhere at any time are more and more overwhelming the cognitive capacity of human users. This is an immediate result of the design goal of providing transparent access to all available information that guides the development of today's information and communication technology. Thus, instead of providing the right information at the right time, current computer systems are geared toward providing all information at any time. This requires humans to explicitly and repeatedly specify the context of the required information in great detail.

The overall problems resulting from this type of information access are amplified by the fact that an ever-increasing number of users are accessing information on the move through portable computer systems such as PDAs and cellular phones. Due to their form factor, such systems cannot be equipped with input devices such as keyboards that are suitable to manually enter large amounts of context information. Hence, these systems are becoming increasingly ill suited to provide efficient mobile access to relevant information.

The vision of Pervasive Computing aims at solving these problems by providing seamless and distraction-free support for user tasks with devices that are invisibly embedded into the environment. In order to provide task support in an unobtrusive and intuitive way, the devices are equipped with wireless communication and sensing technology. This allows them to cooperate with each other autonomously, i.e., without manual intervention, and it enables them to perceive relevant parts of the physical world surrounding their human users. Together with the richer input and output capabilities realizable by the joint utilization of these embedded devices, this can greatly reduce the cognitive load that is put on users when they need to access information.

A. Marqués (✉) and M. Serrano
ETRA I+D, Tres Forques 147, 46014 Valencia, Spain
e-mail: amarques.etra-id@grupoetra.com; mserrano.etra-id@grupoetra.es

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_35,

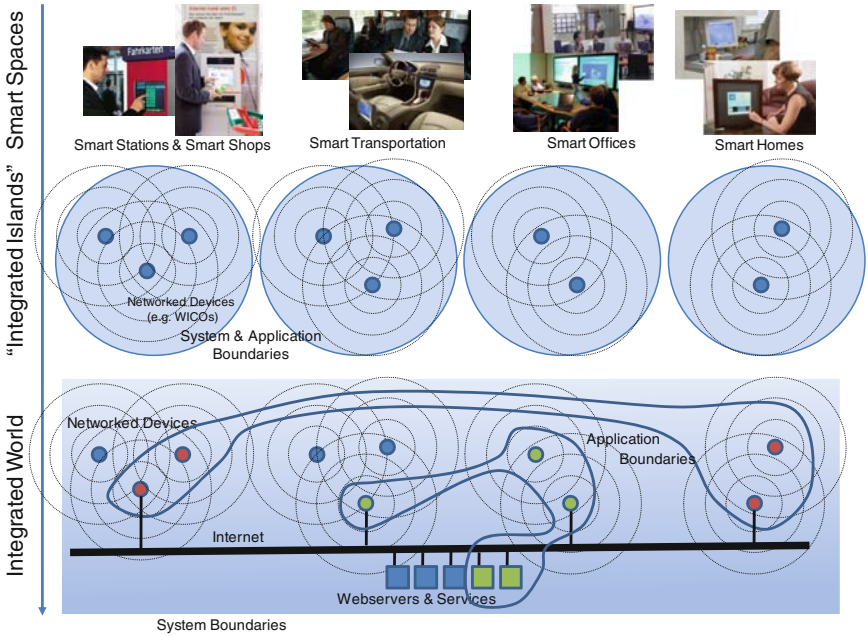


Fig. 1 Pervasive computing vision

While there are various approaches toward enabling the vision of Pervasive Computing, existing approaches are mostly focusing on concepts to realize smart spaces such as smart meeting rooms or offices. However, truly seamless support for user tasks requires the development of one system that exposes a single and unifying image to its human users. This requires the integration of multiple smart spaces with each other and with information system infrastructure that exists today as shown in Fig. 1.

1.1 The Objectives of the Project

The PECES project is geared toward addressing the previously described challenges in order to provide a truly integrated solution. However, the most substantial innovations can be expected from solving the challenge of developing new coordination mechanisms to enable the automated formation of dynamic groups of cooperative devices that are secure and trustworthy. In conjunction with communication mechanisms that have been developed already in the context of previous European research projects, PECES provides a flexible and efficient solution to this challenge, establishing the basis for the development of a solid cooperation layer that can support a broad spectrum of pervasive traveller support services and applications.

The main objective of PECES is the development of a solution to enable the communication among heterogeneous devices across multiple smart spaces, breaking the traditional barrier of "smart islands" where only the services offered in a nearby spatial area can be actually reached.

1.2 Smart Access Control

One of the key challenges PECES technology addresses is providing the traveller with a seamless experience when he/she moves through different smart spaces, being physical or virtual. A delicate balance between usefulness, security, and non-intrusiveness must be kept. Technology must be there all the time, but the user must not see it, he/she has to perceive just the benefits brought to him/her by the applications enabled by PECES technology (Fig. 2).

In this context, the Smart Access Control prototype helps validate some of PECES' main features. To get an idea of the scenario, imagine the user, John Smith, travelling in his car. He has a PDA where he planned his trip – a visit to one of his main customers to hold an important meeting. The moment he got in the vehicle, all smart devices on board – from the PDA to the in-car satellite navigator – became aware of each other's presence. PECES enabled their mutual discovery and their dynamic interaction. Based on the interests of the user, the devices present the possible

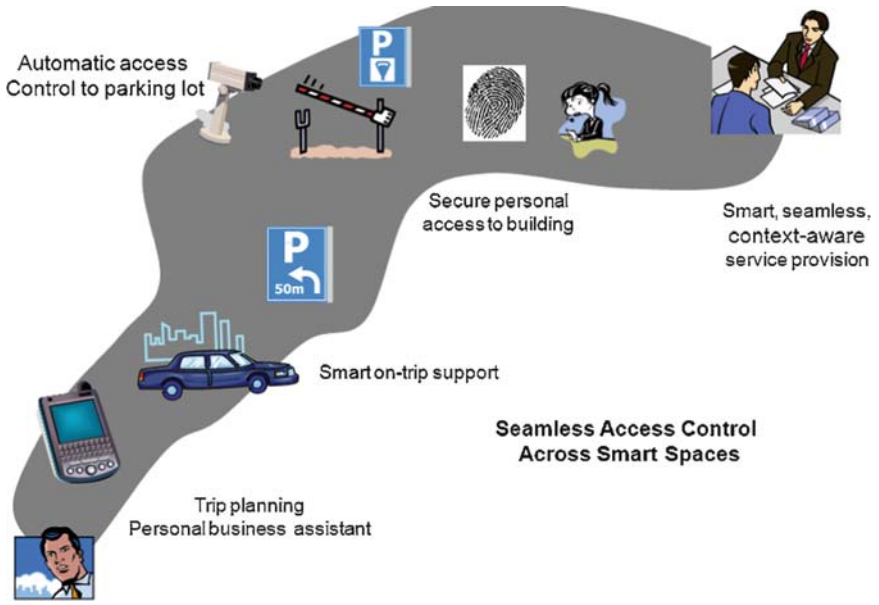


Fig. 2 Smart access control

functionalities available and offer the user a number of services. These services may include the computation of the optimal route to the customer's office, taking into account the real-time restrictions imposed by weather conditions and the traffic jams in the city. Furthermore, they may include the tuning of John's favourite radio station, and a number of other convenience-related functionalities.

When he approaches his destination, his personal device makes contact with the system managing access to the building where his customer's offices are located. They negotiate the access of John's car to the parking facilities of the building. Once there, John's car number plate is recognized by a CCTV camera and his car is granted access to the car park. Within the car park, the in-car navigation system leads John to the parking space, which has been allocated to him by the system.

While he parks, the reception management system of the building negotiates with John's personal device his personal access to the building. He leaves the car, follows the directions displayed on his PDA's screen and reaches reception, where his fingerprint is recognized and the gates are opened for him. Once he is in the building, he will get access to all the locations and services that the system assigns to users with a "guest" profile.

To implement the smart access scenario in the context of the PECES project, three different smart spaces have been defined: the smart car, the smart access control, and the smart office spaces.

1.3 Smart Car

The smart car space is formed by the integration and cooperation of the elements deployed within a car with a user interface device (i.e., the handheld). The most interesting feature arising from such a smart space is the customization of as much parameters of the vehicle as possible upon the entrance of the driver. The vehicle notices who is going to drive, retrieves all the available information about the programmed trip and, according to it, configures itself to the driver contextual preferences. The user interface, on the other hand, is also able to sense the elements within the vehicle and offer a centralized and even automatic interaction with them, thus enabling a series of seamless services enhancing the comfort of the driver.

Within the smart car space, three different elements are considered: a PDA or mobile phone providing an interface between the services and the user, a satellite navigator offering routing services and an on-car radio offering different radio stations and/or CDs.

These elements will form a PAN based on short-range radio communication channels. A suitable and broadly deployed technology fulfilling these requirements is Bluetooth. Nevertheless, other technologies (such as Zigbee, for example) can be explored depending on the specifics of the involved hardware.

Once the PAN is formed, the PECES middleware is able to work. Its first task is to identify the devices present in the PAN, construct the logical group – which we call "smart car" – and present such group to the application layer. The application

grouping units are smart spaces, not PAN members. This feature breaks the current geographical constraints present in most cooperative systems. In order to successfully construct the smart space, and due to the nature of the smart space, the devices need to authenticate themselves and check the driver's permissions prior to allow the interaction (e.g., only allowed drivers will be permitted to take part of the smart space).

1.4 Smart Access Control

The second smart space used in the smart access control scenario focuses on the design of an advanced access control application, enhanced by the new functionalities that the PECES middleware provides. The final objective is to achieve all the tasks necessary to plan a trip (route guidance, ePayment of urban tolling and parking lot reservation) in a mostly transparent way to the user, by enabling the interaction of several smart spaces across heterogeneous networks, avoiding location-based limitations.

The main enhancement of this application, when compared with the traditional approaches, is the seamless way in which all the operations are completed. Traditionally, the user must stop at the entrance of controlled areas, where physical barriers are installed, in order to pay the toll and get access. This system has been proved to be not suitable in areas with a high density of traffic and a small amount of lanes due to the required waiting time and the size of the queues formed in front of the entrances. The automated solutions available nowadays, on the other hand, do need the installation of new hardware in the vehicle. The interaction with the users' handhelds enabled by the PECES middleware enables a low-cost system for the inclusion of fast lanes where users who have paid in advance can drive through, only using their already available handheld devices (Fig. 3).

The possibility to reserve a parking lot as part of the trip design does also greatly enhance the potential of current satellite navigators, enabling a comfortable experience from origin to destination.

In this application, the "smart car space" interacts with up to other two smart spaces: a "smart parking" and a "smart city centre access control".

The smart car space is the same described in the previous section. Regarding the smart access application, the most important element inside the car is the PDA/mobile phone, since it is the element that initiates the trip planning mechanism and interacts with the elements in the other smart spaces.

On the other side, there is a "smart parking space" that implements a communication module built over the PECES middleware and connected to the Internet – the "Remote Reservation System." That system consists of a software module running on the industrial PC where the rest of the parking management modules are deployed.

If the route goes through a restricted area or an urban area with urban tolling control, the PECES middleware can be used to automatically register the vehicle in such

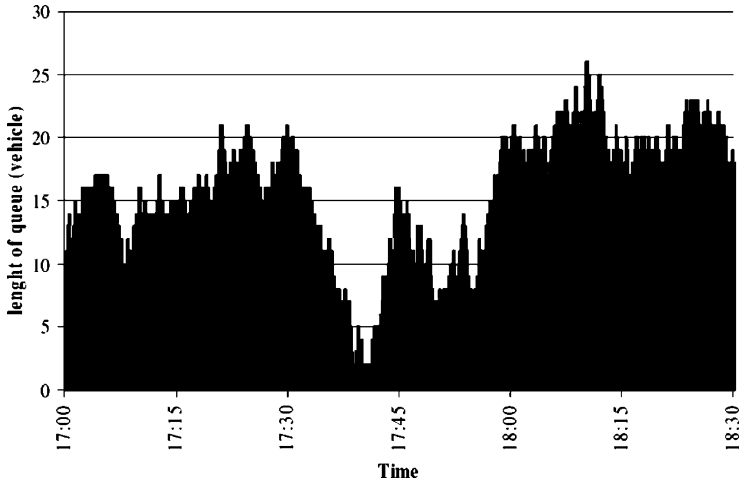


Fig. 3 Queue length at a tollbooth on the Bosphorus bridge (Istanbul) [1]

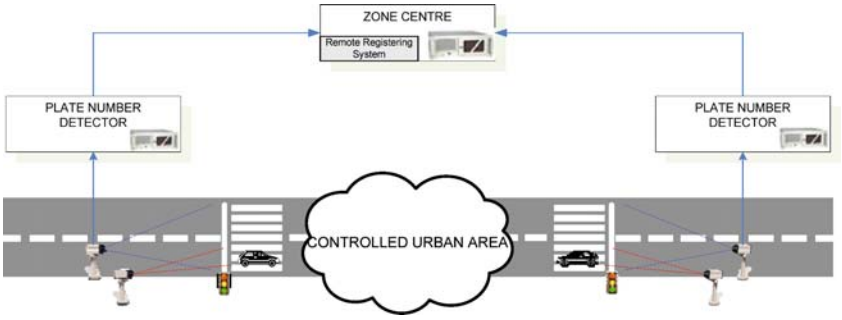


Fig. 4 Smart city centre access control

system and perform the payment. This system is composed of a series of cameras that are installed over each of the lanes of the streets entering the controlled area. When a vehicle enters the area, it is identified and its plate number is crosschecked with the database of allowed vehicles. If there is no match, the violation is communicated to the proper authorities (Fig. 4).

Once the trip destination is known by the PDA (by consulting the detailed agenda or directly prompting the user), and the PDA reaches the smart car space, the PDA can initiate the reservation of a parking lot in one of the parking lots nearby the destination point.

To do this, the PDA trip management application performs a query to the PECES middleware, asking for a list of parking spaces in the nearby of the destination, together with some of their basic details. The PECES middleware uses the available communication channels (GPRS or UMTS in this case) to transparently get the information for the application.

A list with information about these parking lots is presented to the user or automatically crosschecked with his/her preference list. Once the most appropriate parking is selected, the PDA application is able to communicate directly with the "parking smart space" and provide the information necessary to reserve and pay for the parking lot. When the reservation gets completed, the identifier of the granted lot is sent back to the user.

In addition, if the route goes through an urban area that requires a toll payment, the PDA can perform the payment in advance and register the vehicle's plate at the database of the "smart city centre access control". This allows the car to legally perform the trip without any disturbance to the driver. The trip management application performs a query to the PECES middleware looking for the suitable "remote registering system" (e.g., the "remote registering system" of the destination city) once the trip is planned. As soon as the communication is established, the trip management application sends the payment details and the plate number to the access control system, and it waits for an acknowledgment of the operation.

1.5 Smart Office

In the last part of the scenario, the user needs to access some services once he arrives at the destination office in order to, for example, confirm his presence and find the room where the meeting will take place.

The smart office environment offers several services to employees and guests in a seamless manner, using their PDAs and mobile phones as main user interfaces. Therefore, the first identified element is the personal device, which is common to all the three smart spaces of the scenario.

The operation of this third smart space is similar to the operation described in the smart car section. Once the user enters the office, his/her handheld device takes integrates into the office LAN. Afterward, the PECES middleware performs the necessary checks in all the involved devices in order to include the PDA itself in the office smart space (specifically, in the subset required for guests).

2 Conclusions

The first results of the project are available in the last quarter of 2009. A set of context ontologies focused on the mobility scenario described above will be published in a first phase. These ontologies will be used as first approach to develop a number of software libraries that will finally support the deployment of new transport ubiquitous services.

The smart access control described in this chapter will be one of the three domains used to test the validity of PECES approach: a trade show guide assistance system and e-health monitoring application will complete the set of use cases to test and demonstrate under real-life conditions the new features provided by PECES middleware.

References

1. Öğüt KS. (2004) Toll plazas at the bosphorus bridge. 6th International Congress on Advances in Civil Engineering, Bogazici University, Istanbul, Turkey

HYDRA: A Development Platform for Integrating Wireless Devices and Sensors into Ambient Intelligence Systems

Markus Eisenhauer, Peter Rosengren, and Pablo Antolin

1 Developing Middleware to Enable Software Developers to Create Services for Embedded Systems

Embedded Systems are everywhere, built into healthcare devices, into building automation, heating systems and home appliances, into mobile phones and communication, into cars, roads, bridges, and tunnels, and even into our clothes. They are interconnected into networks of many devices and they form the building blocks of the future Internet of Things.

Embedded Systems technologies are deployed in all relevant market sectors and have a major impact on the way these sectors work and collaborate, how they will develop, and how successful their products will be on the world market.

Manufacturers are thus increasingly seeking to network their own products with other systems in order to provide higher value-added solutions for their customers – often a difficult, time-consuming, and costly development process, in particular for SMEs.

The HYDRA project aims to alleviate the problems facing European industries by researching and developing middleware for networked embedded systems that allows developers to develop cost-effective, high-performance AmI applications using heterogeneous physical devices.

Middleware is software that connects different components or applications to enable multiple processes running on one or more machines to interact across a network. The 52-month Hydra project is working on developing middleware that allows developers to create AmI applications, i.e., electronic environments that are sensitive and responsive to the presence of people. HYDRA will showcase its prototype for

M. Eisenhauer (✉)
Fraunhofer FIT, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
e-mail: markus.eisenhauer@fit.fraunhofer.de

P. Rosengren
CNet Svenska AB, Svärdvägen 3B. 182 33 Dandaryd, Sweden

P. Antolin
Telefonica I + D, Zaragoza Area, Spain

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_36,

secure home automation. This will demonstrate how future ambient environments can be designed, realized, and integrated. It will also allow users and developers to provide their input and feedback.

2 Hydra

The HYDRA project [1] develops middleware for networked embedded systems that allows developers to create AmI applications utilizing device and sensor networks.

Device and sensor networking research has seen increasing activity in the last years, with advances in sensor node and radio hardware [2,3]. This work has been instrumental in clarifying the trade-off between computation and communication and the need for in-network processing. Most of this work is based on topographically addressed sensor nodes; other researchers (Heidemann et al. [4]) have based their work on the use of attribute-based naming for structure and data diffusion. The use of attribute-based naming is an interesting concept that will be further investigated in HYDRA. Internet ad hoc routing (Broch et al. survey several protocols [4] such as DSR and AODV) can also be used in sensor networks. SPIN evaluates several variants of flooding for wireless sensor networks [5]. We instead use attributes to name data alone; globally unique identifiers are not used. Through its unique combination of Service-oriented Architecture (SoA) and Model-Driven Architecture, HYDRA will enable the development of generic services based on open standards. In particular, Hydra aims at creating middleware, which operates with limited resources in terms of, e.g., energy and memory. Taking advantage of the HYDRA-middleware interoperability of complex processes as well as heterogeneous infrastructures, services, and devices, Hydra addresses three application domains: home automation, healthcare, and agriculture.

The Hydra middleware as an intelligent software layer, is placed between the operating system and applications. The middleware contains a large number of software components – or managers – who handles the various tasks needed to support cost-effective development of intelligent applications for networked embedded systems. The middleware can be incorporated in both new and existing networks of distributed devices, which operate with limited resources in terms of computing power, energy, and memory usage.

The Hydra middleware allows developers to incorporate heterogeneous physical devices into their applications. It provides easy-to-use web service interfaces for controlling any type of physical device irrespective of its network interface technology. It is based on a semantic Model-Driven Architecture for easy programming and also incorporates means for device and service discovery, peer-to-peer communication and diagnostics. Hydra-enabled devices offer secure and trustworthy communication through distributed security and social trust components of the middleware (Fig. 1).

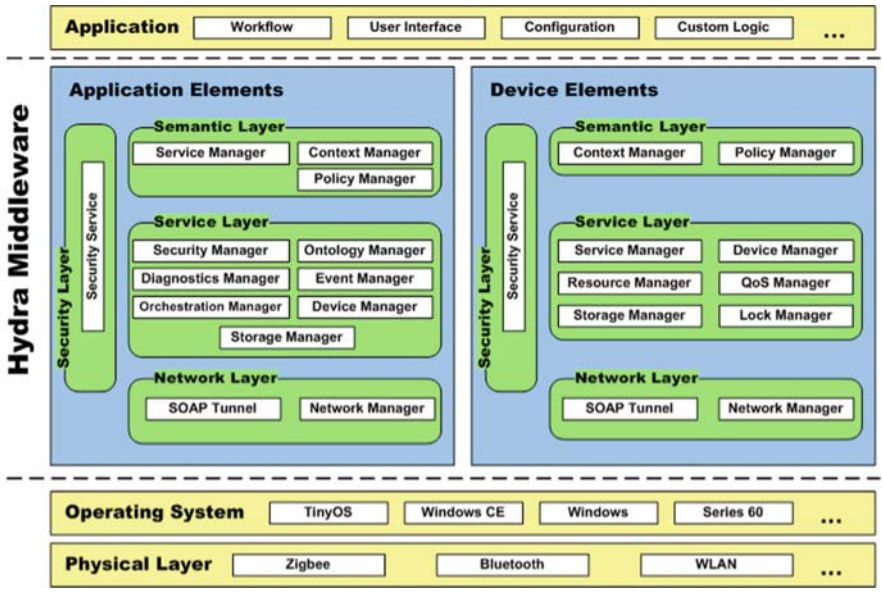


Fig. 1 Software architecture layers

The Hydra Software Development Kit (SDK), Device Development Kit (DDK), and Integrated Development Environment (IDE) will allow developers to create new networked embedded AmI applications and devices quickly and cost effectively. The main research fields in Hydra are briefly summarized below.

2.1 Embedded Ambient Intelligence Architecture

In Hydra AmI applications, any physical device, sensor, actuator, or subsystem can be considered as a unique web service. A major novelty in the Hydra approach is that the middleware provides support for using devices as services both by embedding services in devices and by proxy services for devices. Another novelty is that the middleware supports dynamic reconfiguration and self-configuration, which are indispensable properties in any AmI application [6, 7].

2.2 Wireless Communication and Networks

To assist application developers in addressing a wide variety of mobile and stationary devices and networks, the Hydra middleware hides device-dependent and network-dependent details and provides comprehensive open interfaces to the

display, communication port, input facilities, and memory management of each class of device. The Hydra middleware can also manage communication in the Hydra network, route data, provide session management in the communication, and synchronize the different entities in the network.

A novel implementation in the Hydra middleware is the use of peer-to-peer (P2P) network technologies to identify and utilize the services available in the network, even if they are behind firewalls or NATs [8]. P2P pipes are used as an alternative to WS communication between Hydra-enabled devices.

2.3 Service-Oriented Architecture

The Hydra middleware has features that allow developers to create AmI services and systems through a Service-Oriented (SoA) and Model-Driven Architecture approach. The network part of the Hydra middleware can interconnect devices, people, terminals, buildings, etc. with the SoA providing interoperability at a syntactic level. However, the Hydra middleware also provides interoperability at a semantic level by extending semantic web services to the device level, thus opening up for semantic interoperability of AmI applications.

Hence, the Hydra middleware offers a real novel possibility to discover primitive resource constrained devices, dynamically embed them as Hydra enabled nodes in Ambient Intelligent device networks and provides interoperability between them. In order to achieve this extraordinary discovery capability, the capabilities of the devices must be semantically described in such a way that machine agents can understand and use it [9]. In the Hydra middleware, the semantic description of devices is based on ontologies using OWL, OWL-s and SAWSDL.

2.4 Trust, Privacy and Security

Security goals, such as confidentiality, authenticity, and nonrepudation, can be addressed by a particularly trustworthy design and implementation of wen-service-based mechanisms, enriched by ontologies. The concept behind Hydra security metamodel is semantic resolution of security focusing on moving security from identity-based into a semantic, credential-based framework [10].

2.5 Application Domains

Hydra addresses three application domains: home automation, healthcare, and agriculture.

2.5.1 Building Automation

The field of Intelligent Buildings, Intelligent Homes and Building Management Systems encompasses an enormous variety of technologies, across commercial, industrial, institutional, and domestic buildings, including energy management systems and building controls. The potential of the Hydra middleware in these markets is vast, and peoples' lives are heavily influenced by the effects of Intelligent Buildings technologies. Hydra allows for light-weighted and completely networked "smart homes", both internal and external, controllable, electronically secured, and equipped with different features based on self-learning software.

2.5.2 Healthcare

Hydra improves the productivity of healthcare provisioning.

eHealth services and the development of sophisticated personal wearable and portable medical devices will also allow patients and healthcare professionals to become more mobile and stay longer in the workforce. Hydra will improve the interoperability of intelligent devices and overcome the lack of interconnectivity and interoperability of the various proprietary components and subsystems.

2.5.3 Agriculture

The Hydra middleware will support time-planned data processing, intelligent decision support, and interconnectivity via heterogeneous networks. The Hydra middleware will enable devices and subsystems to communicate and allow developers to develop intelligent, secure applications where devices and subsystems cooperate to perform common tasks in Agriculture.

3 Prototype

Intelligent home automation shows how future ambient environments can be designed, realized, and integrated to provide energy efficiency solutions with a maintained or even increased level of comfort for users.

Taking advantage of the HYDRA-middleware interoperability of complex processes as well as heterogeneous infrastructures, services and devices in future ambient environments can be efficiently modelled and semantically secured; the prototype shows context-awareness and its security and privacy implications in a proof-of-concept implementation.

The HYDRA project has implemented several prototypes in three considered domains within the scope of the project: home automation, e-health, and agriculture. They all make use of the intelligent service layer provided by the middleware that

allows every device, no matter the communication technology it uses (Bluetooth, ZigBee, X-10, etc.), to be presented to the application developer as an UPnP device (proxy) that offers both UPnP and WS services. Thus, the application developer does not need to deal with the particularities of the devices and can access its functionality in a systematic and standardized way.

The prototype we present here implements a set of energy efficiency context-aware ambient intelligent applications in the home automation domain. It is based on the sensors and actuators that are deployed in a living lab in Telefónica I + D premises in Valladolid (Spain), called "Casa Domótica" and where a complete house is available to test the HYDRA applications. A HYDRA middleware instance is deployed also in the house, controlling all the discovered sensors, actuators and devices. Another HYDRA middleware is deployed in a laptop, which may be connected to a different network and still being able to access the devices' services in "Casa Domótica". The context that is taken into account in the prototype is the location of the user (an avatar is used in the application to show the position of the user in the house) and the moment of the day in which the action takes place (whether there is daylight or not). The applications developed in the prototype are:

- *Give me light*: The system provides the user with the needed light regarding his context. For instance, if the user is in the living room and there is daylight, the system will pull up a blind. Otherwise, a lamp will be switched on.
- *Follow me light*: In connection with the "Give me light" application, this application follows the user in the house, switching on or off the lights in the house when the user enters or leaves a particular room respectively.
- *Goodbye, stand-by*: This application takes care of switching off all the lights in the house and of turning down all those devices that are in stand-by mode.
- *Energy-consumption information*: The user gets real-time information about the energy consumption at home as he interacts with the electric appliances. Moreover, each device includes an energy profile, which allows the user to check the consumption of a particular device within the house.

Acknowledgment This work was supported in part by the European Commission under the Integrated Project HYDRA contract FP6-IST-034891.

References

1. Hydra Middleware Project. FP6 European Project. http://www.hydra.eu.com
2. Sohrabi K, Gao J, Ailawadhi V, Pottie G (1999) A self-organizing sensor network. In: Proceedings of the 37th allerton conference on communication, control, and computing, Monticello, IL, USA, Sept 1999.
3. Pottie GJ, Kaiser WJ (2000) Embedding the internet: wireless integrated network sensors. Commun ACM, 43(5):51–58.
4. Heidemann J, Silva F, Intanagonwiwat C, Govindan R, Estrin D, Ganesan D (2001) Building efficient wireless sensor networks with lowlevel naming. In: Proceedings of the 18th ACM symposium on operating systems principles (SOSP), 146–159, October 2001.

5. Broch J, Maltz DA, Johnson DB, Hu Y-C, Jetcheva J (1998) A performance comparision of multi-hop wireless ad hoc network routing protocols. In: Proceedings of the ACM/IEEE international conference on mobile computing and networking, 85–97. Dallas, TX, USA, Oct 1998.
6. Zhang W, Hansen KM (2008) An owl/swrl based diagnosis approach in a web service-based middleware for embedded and networked systems. In: The 20th international conference on software engineering and knowledge engineering, 893–898. Redwood City, San Francisco Bay, CA, USA, Jul 2008.
7. Zhang W, Hansen KM (2008) Towards self-managed pervasive middleware using owl/swrl ontologies. In: Fifth international workshop on modeling and reasoning in context (MRC 2008), 1–12. Delft, The Netherlands, Jun 2008.
8. Milagro F, Antolin P, Kool P, Rosengren P, Ahlsén M (2008) "SOAP tunnel through a P2P network of physical devices." In: Internet of Things Workshop, Sophia Antopolis, French Riviera, France, Sep 2008.
9. Kostelník P, Sarnovský M, Ahlsén M, Rosengren P, Kool P, Axling M (2008) "Semantic devices for ambient environment middleware." In: Internet of Things Workshop, Sophia Antopolis, French Riviera, France, Sep 2008.
10. Hoffmann M, Badii A, Engberg S, Nair R, Thiemert D, Mattheß M, Schütte J (2007) "Towards Semantic Resolution of Security in Ambient Environments." In: Ami.d – 2nd conference for ambient intelligence developments, Sept 2007.

Probabilistic Information Dissemination for MANETs: The IPAC Approach

Odysseas Sekkas, Damien Piguet, Christos Anagnostopoulos, Dimitrios Kotsakos, George Alyfantis, Corinne Kassapoglou-Faist, and Stathes Hadjiethymiades

1 Introduction

One of the main problems in mobile ad-hoc networks (MANETs) is the efficient dissemination of data, typically from the different sources to destination nodes. Solutions that work efficient for a specific setup do not perform well on slightly different applications. Hence, careful examination and selection of dissemination algorithms is needed. Epidemic dissemination is introduced as a method to reliably spread information across a network in which no direct path from source to destination is guaranteed.

The Integrated Platform for Autonomic Computing (IPAC) project aims at delivering a middleware and service creation environment for developing embedded, intelligent, collaborative, and context-aware services in mobile nodes. IPAC relies on short-range communications for the ad hoc realization of dialogs among collaborating nodes. Advanced sensing components leverage the context-awareness attributes of IPAC,[1] thus rendering it capable of delivering highly innovative applications for pervasive computing. IPAC networking capabilities are based on epidemic/rumor spreading techniques, a stateless and resilient approach, and information dissemination among embedded nodes.

Spreading of information is subject to certain rules (e.g., space, time). IPAC nodes may receive, store, assess, and possibly relay the incoming content to other nodes. The first requirement for the IPAC dissemination scheme is to support multiple sinks for the same piece of information. Most of the constraints (energy, limited processing capability) encountered in Wireless Sensor Networks (WSNs) can be

[1]IPAC – Integrated Platform for Autonomic Computing (ICT-224395), is funded by the European Community through FP7 ICT Program. Web site: http://ipac.di.uoa.gr.

O. Sekkas (✉), C. Anagnostopoulos, D. Kotsakos, G. Alyfantis, and S. Hadjiethymiades
Pervasive Computing Research Group Department of Informatics and Telecommunications, University of Athens Panepistimioupolis, Illissia, 15784, Athens, Greece
e-mail: sekkas@di.uoa.gr; bleu@di.uoa.gr; shadj@di.uoa.gr

D. Piguet and C. Kassapoglou-Faist
CSEM, Centre Suisse d'Electronique et de Microtechnique S.A. Jaquet-Droz 1, CH-2002, Neuchâtel, Switzerland

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_37,

found on IPAC nodes as well. On top of that, the mobility of IPAC nodes adds more constraints and problems that must be dealt with when disseminating information. In this sense, the IPAC framework has the operating parameters of a Mobile Sensor Network (MSN), which is a WSN with moving sensors.

The rest of the chapter is organized as follows: previous and related work is presented in Sect. 2. Section 3 describes the problem of information dissemination in the context of IPAC and presents a set of preliminary simulation results in an effort to shed some light on how information dissemination is affected by various model parameters. Based on these results, adaptive probabilistic schemes are proposed and assessed. Section 4 concludes the chapter and presents directions for future work.

2 Related Work

The simplest technique to spread information is flooding. Another one is that the node or process that owns a piece of information, broadcasts it regularly to random subsets of its neighbors with certain probability [1, 2]. The basic parameters of such epidemic dissemination models were defined in [3] and are the number of times a message is forwarded, the buffer capacity for each node or process, the total number of nodes or processes (system size), the number of known nodes or processes (partial view size), and the size of target group of nodes. The main issues associated with information spreading are scalability and reliability of the dissemination.

Such simple models suffer from certain drawbacks: do not scale nicely; impose considerable overhead (traffic) due to deliveries to uninterested nodes, and do not take into account the limited amount of storage available on the nodes [3]. Additionally, they do not take into account the membership issue (i.e., which node knows which other). Work performed so far addresses some of the aforementioned issues. In all cases, the goals of every proposed algorithm or technique are threefold: maximize message delivery rate, minimize message latency, and minimize the total resources consumed in message delivery [4].

The SPIN protocols [5, 6] were introduced to improve the situation. They are based on meta-data (i.e., semantic awareness). They can be bound to the needs of each application, and they impose selective retransmissions of information, thus minimizing retransmission overhead. Nodes transmit after ensuring that the information to be transmitted is useful and after probing their own resource manager. In this manner, the problems of redundant information transmission and resource-blind transmission are addressed. Publish/subscribe models have also been proposed [7] as directed diffusion-based schemes. Probabilistic broadcast and multicast schemes [8,9] have been proposed in order to address the reliability issue raised as one moves away from the flooding concept. Gossip-based broadcast algorithms trade reliability guarantees against "scalability" properties, which is a known pattern in this field. The concept of awareness can apply to network-related properties too, such as the received signal strength indication or RSSI [10].

More limited awareness may be used in random phone call-type algorithms to push rumors to one random partner each time. This does not use awareness beyond immediate neighbor awareness, but can impose considerable overhead traffic in order to reach good reliability [11]. Dissemination schemes that avoid flooding the network without the need to maintain subscription information have also been proposed [12]. A multiepidemic approach that relies on semantic dependencies, which are modeled through a hierarchical representation scheme, is investigated in [13].

3 IPAC Information Dissemination

In IPAC, information dissemination actually boils down to the forwarding of messages by each node, according to specific rules. Messages are broadcast by each node to its neighborhood, i.e., are not explicitly destined to specific neighbors. Nodes receive every message they listen to and process it according to some rules/policies. It is also possible that nodes subscribe for information that they are interested in. In case that relevant information is received through an IPAC message, it is delivered to the appropriate application, and, then, such message is considered again for forwarding. In this chapter, we focus only on the dissemination of messages throughout the network, leaving outside details regarding the application layer.

At each time instant, nodes might be in vicinity and connected to each other thus forming a graph. Note that, due to mobility, such graph is not static and might change quite rapidly. A node, at any time, might generate a message that will have several attributes, typically set by the application that generated it (e.g., time validity – TTL or criticality). A node is a reconfigurable entity that has several parameters tunable to adapt to a specific situation. The concept of IPAC is to let the node itself observe its situation and decide how to optimally tune its operation. A situation can be decomposed into a set of primitive elements. The proposed scheme represents a typical cognitive networking model, where each node is capable of sensing/observing its situation/context, and using this information to intervene with diverse, cross-layer parameters within the node, thus, optimizing its operation.

3.1 Nonadaptive Probabilistic Broadcast Algorithm – L0

In this section, we present the simulation results of a nonadaptive broadcast algorithm. Such simulations assist in the design of an adaptive algorithm. In this setup, the number of broadcast attempts (f_m) and the probability of broadcast by a certain node (β) are the same for every node and do not vary during the simulation. All simulations have been performed using the Omnet++ tool [14].

The basic setup involves $N = 25$ nodes which are placed on a 5×5 squared grid Nodes are fixed (zero mobility). While the number and the relative positions of the nodes remain static, the side size w varies in order to change the network density. For

the lower layers of the network, we employed IEEE 802.15.4 nonbeacon enabled and CSMA MAC layer. Table 1 summarizes the simulation parameters along with the corresponding domains adopted for the simulations.

In this simulation setup, only node with identifier 0 sends packets. The back off parameter is the probabilistic delay upon which every node waits before sending the first copy of a message to mitigate the risk of systematic collisions. From T_m, max of back off and TTL, it is effortless to calculate how many times a message will be sent at most. In our case: $f_m = 10, 5, 4, 3, 2$, and 1 times. The node, which creates a message, sends it definitely ($\beta = 1.0$) at the first attempt, independent of the current value of β. Each message m, which arrives at the network layer, is forwarded only once to the application. Other copies of that message are discarded. The metrics that were monitored through the simulations are:

- *Success rate (good-put)*: (total number of received packets)/(total number of expected packets). Total refers to all the nodes. The total number of expected packets is defined as $(N-1)P_0$.
- *Number of forwardings*: average number of packet transmissions within the network per message.

In Fig. 1a, we provide plots of the success rates (good-put) for $w = 300$ m. We observe percolation phenomena such that full network coverage is obtained after

Table 1 Simulation parameters

N: number of nodes	25
P_0: number of packets sent by node 0 (for other nodes: none)	50
Traffic type	Exponential (10 s)
TTL [s]	20
Backoff [s]	Uniform in [0, 1]
Number of runs per β	5
W: square size [m]	50, 100, 200, 300, 400
β: probability of broadcast	from 0 to 1 with steps of 0.05
T_m: broadcast period [s]	2, 4, 6, 8, 10, 21

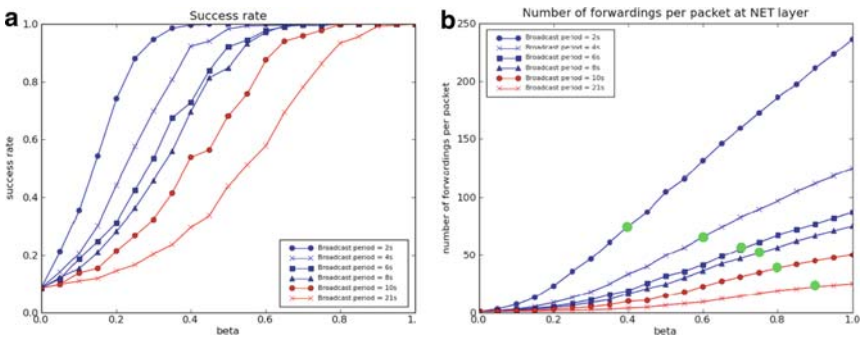


Fig. 1 Side size $w = 300$ m. (**a**) Success rates (good-put). (**b**) Number of packet transmissions per message in the network

a certain value of β. Once this threshold is reached, it is meaningless to further increase β. This percolation value decreases with network density and increases with the broadcast period (i.e., decreases with the number of transmissions), as expected. Once the percolation conditions are known, we can consider the cost factor. In our case, we estimate the cost from the number of transmissions needed within the whole network in order to obtain a given coverage.

In Fig. 1b, we have added enlarged dots to indicate percolation probabilities for $w = 300$ m. We observe that to reach full coverage, it is cheaper, in terms of number of transmissions, to send a message once with high probability than to forward it several times with lower probability. From our observations, we derive hints for the design of the adaptive algorithm:

Hint 1. The broadcasting probability must decrease with the network density.
Hint 2. To reach a desired coverage at the lowest energy cost, one should adopt higher probabilities along with lower number of broadcasts (or longer broadcast periods).

In another scenario, we simulate conditions of network congestion. All nodes send 50 packets and we varied the exponential traffic parameter from a mean of 10 to 1 s. We also introduce an additional metric. which is the number of packets dropped by the MAC layer due to congestion. The derived hint from this simulation experiment is:

Hint 3. The dissemination protocol must observe or be notified of MAC status. When packets are dropped at MAC layer, probability of broadcast must be decreased until the number of dropped packets becomes negligible.

Scenarios involving random mobility have also been simulated. Good-put results showed that sometimes mobility helps achieving full dissemination with lower probabilities while, under certain conditions, mobility undermines good-put. Hence, we could not find a clear rule about how the mobility should influence the probability.

3.2 Adaptive Probabilistic Broadcast Algorithm – L1

In this section, we propose an adaptive probabilistic broadcast algorithm. The difference between the adaptive and the nonadaptive probabilistic broadcast is that the probabilistic parameter β is dynamically adjusted by each node according to Hint 1, Sect. 3.1. Specifically, the network density is inferred from the estimation of the number of neighbors. Each time a packet is received, the address of the sender is registered in a "neighbors" table. At that time, a timer related to the new entry is set to T_{L}. When the timer expires, the entry is removed from the table. When the node receives a packet from a neighbor that is already registered, it restarts the corresponding timer to T_{L}. Each T_{c} seconds, the node counts how many nodes are in its table. After that, it adjusts the value of β according to the number of neighbors estimation with the values presented in Table 2, determined experimentally.

Table 2 Adaptive probabilistic broadcast: values of beta (β) for different neighborhood sizes

Number of neighbors	0, 1, 2, 3	4, 5	6, 7	8, 9	10, 11	12, 13	14, 15	16, 17	18, 19	≥ 20
B	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1

Table 3 Coverage and number of forwardings for adaptive probabilistic broadcast

Grid width [m]	50	100	200	300	400
Good-put	1.000	0.999	0.996	0.980	0.055
Number of forwardings[a]	14.9	15.7	20.7	22.0	2.2

Grid layout, 25 nodes. Only node 0 (top, left) sends 50 packets. Mean inter-packet interval: 10 s. TTL = 20 s. Broadcast period = 21. Initial value of β when the node is turned-on: 0.9

[a]Average number of times a particular message is forwarded in the network, including the first emission when it is created. If $\beta = 1.0$ and broadcast period >TTL (one broadcast at each node), average number of forwarding per message = number of nodes

Table 4 Good-put and average number of transmissions per message when all nodes send 50 packets

Grid width [m]	50	100	200	300	400
Good-put	0.996	0.962	0.952	0.964	0.111
Number of forwardings	3.9	5.0	15.9	20.0	3.6

We adopt the same simulation setup regarding the grid layout, the number of nodes and other settings (TTL, etc.) in Sect. 3.1. The timer parameters of the adaptive scheme are $T_{\mathrm{L}} = 10\,\mathrm{s}$ and $T_{\mathrm{c}} = 5\,\mathrm{s}$. The broadcast period is set to TTL + 1 to ensure only one broadcast. Therefore, there is only one value per grid size. We examine the behavior of β derived from the network nodes. As we can conclude from Table 3, the adaptive algorithm allows full or, almost full, coverage even when the number of broadcasts is only one.

Once the network density is high, we notice that the average number of transmissions per message is higher than necessary. If traffic is too low, the values of T_{L} and T_{c} may be too low to allow nodes to find out all their neighbors. To verify this assumption, the simulation is reexecuted with all 25 nodes set to send 50 packets. The results are provided in Table 4 and plotted in Fig. 2.

As expected, the algorithm is more efficient in the presence of enough traffic. We observed also a border effect: nodes which are near the border of the playground (nodes 10 and 14) detect fewer neighbors and infer a lower density.

We have also run a mobility scenario with the adaptive algorithm. Evidently, we obtain that increasing the number of broadcast attempts (that is decreasing the broadcast period) definitely achieves better network coverage. This further supports our strategy of tuning the probability according to the density and not on the mobility. Using mobility to control the number of broadcast attempts helps obtaining full, or almost full coverage while keeping the algorithm simple.

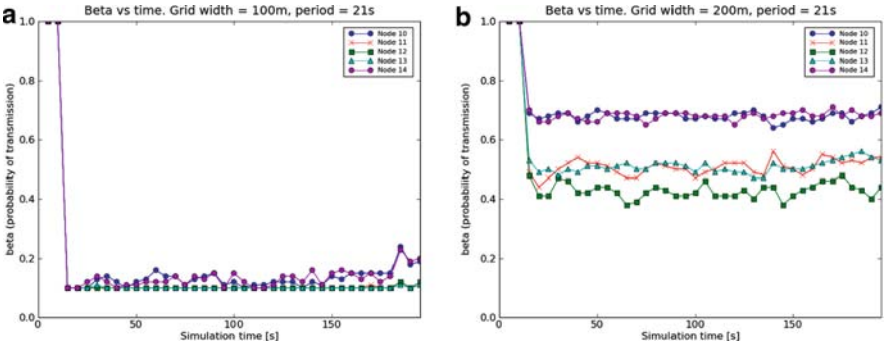


Fig. 2 Probability β vs. time for playground sizes of (**a**) 100 m and (**b**) 200 m. All 25 nodes generate and transmit 50 messages. Initial value of $\beta = 1.0$

3.3 Adaptive Probabilistic Broadcast Algorithm – L2

In this section, we describe an alternative approach to achieve information dissemination in IPAC. We describe a second adaptive probabilistic broadcast algorithm (L2). We provide a preliminary discussion, introducing useful notations and ideas, as well as a detailed description of the proposed algorithm.

Let m be a message received by a node. We define t_m (reception time for message m), TTL_m (the TTL of message m). Upon receiving message m, the node should consider transmitting the message within time frame $[t_m, t_m + \mathrm{TTL}_m]$. The message may be transmitted multiple times within this time frame, i.e., f_m times. The number of transmissions f_m must increase with TTL_m. The transmission period is $T_m = \mathrm{TTL}_m / f_m$. The node, at the beginning of each transmission period, decides stochastically whether to transmit or not with probability β. The probability is calculated according to the number of neighbors, as well as changes in the node's neighbor list. Specifically, β starts from an initial value of β_0 and is adjusted dynamically according to network density. Another factor that influences β is node mobility and the number of messages retransmitted by neighbors. A minimum value of $\beta_{\min}$ is assumed to avoid having all nodes seizing transmitting.

Moreover, a relative mobility factor M_n is defined. Each node keeps a table of the nodes seen in its neighborhood, and the table is updated when a node receives or overhears a packet. When a timer expires, the corresponding entry is removed from the table. When an entry is removed or inserted in the table, a change counter is incremented. The node periodically observes the change counter with period T_c. The counter is set to zero at the beginning of each period. At the end of each period, the node decides to alter the mobility factor M_n depending on the observed number of changes N_c in the neighborhood:

$$M_n = (N_c - N_{\mathrm{mean}})/N_{\mathrm{mean}}.$$

where N_{mean} is the mean of the observed values of N_c at each round. In other words, M_n represents the deviation from the expected number of changes.

A network density factor D_n is also defined. The network density factor D_n can be altered periodically by observing the total number of nodes K_n present in the neighborhood table used to compute the mobility factor. Specifically:

$$D_n = (K_n - K_{\text{mean}})/K_{\text{mean}}$$

where K_{mean} is the mean of the observed values of K_n at each round. In other words, D_n represents the deviation from the expected number of neighbors.

3.3.1 Simulation Layout

The basic setup involves $N = 25$ moving nodes, which are initially randomly placed on a varying width squared playground. Nodes are moving randomly according to the following mass mobility scenario:

- At the beginning, nodes are randomly placed on the playground field.
- Each node changes its speed and direction a number of times. The interval of time between changes is normally distributed ($\sim N(10, 0.5)$, figures in seconds).
- At each change, every node turns by a certain angle ($\sim N(0, 30)$, figures in degrees).
- At each change, every node selects its new speed ($\sim N(0.1, 0.1)$, figures in m/s).
- When a node reaches an edge of the playground, it bounces on it.

The goal of mass mobility is to take into account the inertia of moving objects. Table 5 describes the simulation parameters.

Only node 0 sends packets. The setup regarding the messages m that are generated and arriving in networking level, as well as the metrics value that are monitored (*Success rate* and *Number of forwardings*) are similar to those of the Sect. 3.1. We notice that for low density, the average number of transmissions per message as well as good-put gets lower. We would like to examine the values of β that the

Table 5 Simulation parameters

N: number of nodes	25
P_0: number of packets sent by node 0 (for other nodes: none)	50
Traffic type	Exponential (10 s)
TTL [s]	20
Backoff [s]	Uniform in [0, 1]
Transmission periods	10, 21
w: square size [m]	200, 300, 400
β_0	0.5, 0.7, 0.9
Node (average) velocity (m/s)	5, 10, 15, 20

Table 6 Good-put and number of forwardings vs. mean speed and playground size

	Field side (m)	200	300	400
Average speed: 5 m/s	Good-put	0.993583, 0.901000	0.771667, 0.554000	0.417667, 0.294167
	Number of forwardings	23.312, 11.326	15.932, 7.422	8.572, 4.408
Average speed: 10 m/s	Good-put	0.991500, 0.897250	0.823500, 0.561167	0.487250, 0.287500
	Number of forwardings	22.566, 11.264	16.81, 7.51	9.708, 4.306
Average speed: 15 m/s	Good-put	0.996583, 0.921167	0.891333, 0.595583	0.505583, 0.292250
	Number of forwardings	22.86, 11.64	18.276, 8.034	9.504, 4.566
Average speed: 20 m/s	Good-put	0.993000, 0.937417	0.893083, 0.599833	0.550333, 0.305917
	Number of forwardings	22.498, 11.72	18.146, 8.324	18.146, 8.324

$\beta_0 = 0.5$. In each cell, the first value corresponds to a broadcast period of 10 s and the second one to a broadcast period of 21 s

nodes derive. In Table 6, we summarize our findings. Results are shown for broadcast periods 10 and 21 s (one and two broadcasts at most) and $\beta_0 = 0.5$ and indicate increased network coverage as node mobility (node speed) increases. This observation becomes more evident when the network is sparse in the case of $w = 400$ m. This means that mobility helps the dissemination process.

The main difference of algorithm L2 with the algorithm L1 is that it tries to infer mobility and computes the broadcast probability β accordingly. Moreover, while L1 relies on the detected number of neighbors to infer density and tunes the probability accordingly, L2 starts from a reference value β_0 and increases or decreases it according to the variation of density. The simulations show that the derived values of β heavily depend on β_0.

4 Conclusions and Future Work

Disseminating information within a wireless ad-hoc network calls for the use of a flooding technique. In IPAC, by proposing a probabilistic broadcast technique, we aim to reduce the amount of message transmissions. Simulation results show that for a given network density, there is a minimal probability of broadcast $\beta_{\min}$, (percolation probability), which achieves full coverage. Trying different broadcast periods, we found out that the percolation probability decreases with the number of broadcast attempts. However, in terms of efficiency, using a low number of broadcast attempts and higher probabilities involves fewer transmissions than the opposite.

The simulations also showed the importance of implementing a congestion control mechanism. If the network creates too many messages, the mechanism will not allow full coverage, but will help reach the maximal possible coverage given the actual generated traffic. Mobility scenarios were simulated as well. Due to their significant randomness, we cannot conclude whether probability should be increased or decreased depending on the mobility.

We also propose an adaptive probabilistic broadcast algorithm (L1) based on network density inferred by the estimation of the number of neighbors. Simulation results prove that a good coverage can be obtained in very different conditions of network density, size, and mobility. Moreover, the algorithm is able to detect high density situations and divide the amount of transmitted packets by more than 5 for a network of 25 nodes. Another advantage of the scheme is that it does not require control signaling. Hence, the protocol overhead is minimal.

A variant of the adaptive algorithm (L2) is also introduced. The main difference with the first algorithm is that it tries to infer mobility and takes it into account when computing the probability of broadcast. Also, it starts from a reference value β_0 and increases or decreases it according to the variation of density. The simulations show that the derived values of β heavily depend on β_0. Our findings show that the first scheme is more likely to provide a good coverage irrespective of network density.

As future work, we are planning to run simulations in order to optimize certain parts of the presented algorithms, such as congestion control (which could depend on the criticality of messages) and number of broadcasts according to mobility.

References

1. Demers AJ, Greene DH, Hauser C, Irish CW, Larson J (1987) Epidemic algorithms for replicated database maintenance. In: PODC, Vancouver, British Columbia, Canada, pp 1–12
2. Franklin MJ, Zdonik SB (1996) Dissemination-based information systems. Data Eng Bull 19(3):20–30
3. Eugster PTh, Guerraoui R, Kermarrec AM, Massoulie L (2004) From epidemics to distributed computing. IEEE Comput 37(5):60–67
4. Vahdat A, Becker D (2000) Epidemic routing for partially-connected ad hoc networks. Technical Report CS-200006, Duke University
5. Heinzelman W, Kulik J, Balakrishnan H (1999) Adaptive protocols for information dissemination in wireless sensor networks. In: Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking, pp 174–185
6. Kulik J, Rabiner W, Balakrishnan H (2002) Negotiation-based protocols for disseminating information in wireless sensor networks. Wirel Netw 8(2–3):169–185
7. Estrin D, Govindan R, Heidemann J, Kumar S (1999) Next century challenges: scalable coordination in sensor networks. In: Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking, pp 263–270
8. Eugster PTh, Guerraoui R, Handurukande SB, Kouznetsov P, Kermarrec AM (2003) Lightweight probabilistic broadcast. ACM TOCS 21(4):341–374
9. Luo J, Eugster P, Hubaux JP (2003) Route driven gossip: probabilistic reliable multicast in ad hoc networks. In: Proceedings of the 22nd IEEE conference on computer communications, San Francisco, California, USA, pp 2229–2239
10. Miranda H, Leggio S, Rodrigues L, Raatikainen K (2006) A power-aware broadcasting algorithm. In: Proceedings of the 17th annual IEEE international symposium on personal, indoor and mobile radio communications, Helsinki, Finland, pp 1–5
11. Karp R, Schindelhauer C, Shenker S, Vöcking B (2000) Randomized rumor spreading. In: Proceedings of the 41st FOCS, Redondo Beach, CA, pp 565–574
12. Datta A, Quarteroni S, Aberer K (2004) Autonomous gossiping: a self-organizing epidemic algorithm for selective information dissemination in wireless mobile ad-hoc networks. In: International Conference on Semantics of a Networked, pp 126–143

13. Anagnostopoulos C, Hadjiefthymiades S (2008) On the application of epidemical spreading in collaborative context-aware computing. ACM SIGMOBILE MC2R 12(4):43–55
14. Varga A (2001) The OMNeT++ discrete event simulation system. In: Proceedings of the European Simulation Multiconference. SCS – European Publishing House, Prague, Czech Republic, pp 319–324

Part V
Security and Privacy Issues

An Overview of Privacy and Security Issues in the Internet of Things

Carlo Maria Medaglia and Alexandru Serbanati

1 Introduction

The trend of having ever more objects included in the IT data flows and ever more connected devices, moving toward mobile and decentralized computing is evident. The Internet and ancillary technologies are the base that provide the needed connectivity. In the last few years, the idea of connecting existing computing devices gave place to the concept of "connecting things." The Internet of Things, as it is called, has drawn a lot of attention from many academics and public research institutions. While there is no global consensus on the meaning of the term [1–3], it is clear that the main idea behind the IoT concept is the ability to connect loosely defined smart objects and enable them to interact with other objects, the environment, or more complex and legacy computing devices. The communication infrastructure will be based on an extension of the Internet, which will enable transparent use of the object resources across the globe. Smart objects will densely populate human life and human environment [4], interacting with both by providing, processing, and delivering any sort of information or command. Objects and environment will be able to tell us about them, their state, or their surroundings and can be used remotely. Sensors will be integrated in buildings, vehicles, and common environment, carried by people and attached to animals and will communicate among them locally and remotely in order to provide integrated services.

For example, mobile devices could adopt silent mode when entering a meeting room if this is the request of the meeting moderator, alert the user and turn off the radio before entering sensitive medical areas or detect when the user enters the car and connect to its sound system. Wireless sensors could let people check where their pet is in real-time as well as control the temperature of each room of their house while they are out. Emergency services could be remotely and automatically alerted if fire is detected in a building or if a patient's medical parameters drop beyond a critical threshold.

C.M. Medaglia (✉) and A. Serbanati
Centro per le Applicazioni della Televisione e delle Tecniche di Istruzione a Distanza (CATTID), University "Sapienza", Rome, Italy
e-mail: carlomaria.medaglia@uniroma1.it

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_38,

With such a deep penetration of technology, which will introduce a new kind of automation and remote interaction, it is likely that new security and privacy issues will rise.

2 Short Term

The first steps of this process are already ongoing, with autoID and sensor and actuator networks as peripheral enabling technologies. In this context, only presence and sensor data can be provided by the peripheral part to the central structure. Radio Frequency Identification (RFID) tags are passive (powered by the RF field emitted by the RFID reader) while more complex objects are battery powered. Their battery lifetime is in inverse proportion to that of their processing and communication potential. Self- and context-awareness, provision of web services[1] as well as mobility inside a global, morphologically dynamic network are still missing at this stage. Also, a fundamental lack is the absence of a unified communication standard across the different parts of the Internet of Things network.

Even if some of the auspicated enabling technologies for the IoT are still missing, it is evident that the resulting overall number of connected devices will be very large and the Internet, as we nowadays conceive it, would soon be overwhelmed. One of the shared ideas about the possible solution is the adoption of the IPv6 standard, which will provide a sufficient number of available unique addresses for many years to come, thanks to a 128 bit addressing instead of current 32 bit of IPv4. This scenario, dominated by current technological limits, is what the authors call the short-term scenario.

The current scenario follows two main ways of collecting information in the environment: RFID and Wireless Sensor Networks (WSN). Note that while authors have chosen a technology from each of the aforementioned peripheral enabling technologies, both are wireless because: (a) this is the current IoT forecast trend in order to provide mobility and service portability, and (b) wireless devices are more challenging from a security point of view as they share the physical medium with other, potentially malicious devices.

2.1 RFID and Identification

The most common RFID implementations use passive tags, which uniquely and wirelessly identify the items to which they are attached, enabling they presence

[1] Authors use the term "services" when referring to high level services provided by business systems to their users, while the term "web service" is properly used for a software capable of providing a standard interface with other computing devices across the network. The term "infrastructure service" will be used referring to a software run by the governance of the IoT providing smart objects critical information for the operation of the IoT itself.

monitoring. Recent tags, especially UHF ones because of their higher data throughput and widespread, also have a small amount of memory in which business data can be stored, but usually this data is either unprotected or read-only locked. EPC Global Gen-2 tags also provide a wider range of primitive functions onboard though labels provide no security feature by default. *Kill* and *Lock* commands are available. The former feature is not useful in the IoT context and, as seen in [5], it may lead the way to other threats as eavesdropping the communication session could give an attacker the 16bit PIN needed to kill the tag which, usually, is the same for all the tags in a given system.

In RFID systems, the reader never authenticates and tags are by default set to respond to the interrogation of any compliant reader, which poses a concrete threat to privacy. Not only authorized readers can read the tag, but also rogue ones. Also, even readers that should be authorized in a context could read the tag on unsolicited occasions. Approaches to this issue in logistics envisage the killing of tags, use of active jamming or even Faraday cages [6], but these are not compliant with such a pervasive architecture as that of IoT.

Also, the authentication of the RFID tags itself is subject to issues: while the primary goal of this technology is to provide a means of identification, it is not a secure way of. ID-writable tags are available in case of simpler RFID technologies. More recent standards (such as the aforementioned Gen-2) provide the primitives for developing more complex authentication features and a good amount of academic research is drawn by this subject [7]. Yet, the shared communication, physical medium, and the reduced computational capabilities make it difficult to develop an absolutely secure system [8] based on passive RFID.

Currently, RFID solutions should be used in noncritical contexts: in the IoT, RFID can provide information about object presence and, eventually sensor-collected data but system designers should address the risk malicious alteration of such information. Also, for RFID solutions to be integrated in the IoT, a special middleware must be used to provide a suitable, possibly web service based, interface to remote interaction. In this case, services will of course be provided by a central architecture or by the middleware. Security issues for this part of the network are out of the scope of this work.

13.56 MHz RFID-based Near Field Communication (NFC) is sometimes seen as an answer to some of the security issues of RFID. Mifare and NFC compliant devices provide authentication and symmetric cipher, though it has been already demonstrated that reverse engineering is practicable and can compromise the entire system [9].

The fact that the technology is usually embedded in mobile devices also provides interesting options for enhancing overall system security. As seen in [10], the availability of network communication and consequent access to a Public Key Infrastructure (PKI), together with the peer-to-peer NFC operation mode and a programmable environment may give place to secure applications.

2.2 WSN and Networking

The first and most important architectural differences between WSNs and RFID are the networking and processing capability. Nowadays, the greater part of WSNs are based on different implementations of the IEEE 802.15.4, a standard for low-rate WPAN, which provides different network topologies, among which mesh networking.

WSNs were born for field survey or control in military and ecological contexts. Such battery-powered computing devices had long autonomy and small form-factor requirements and thus usually had constrained hardware (low processing power, limited connection and storage capabilities) while compared to other devices.

These requirements very well suit the idea of IoT and mesh networking is also very interesting because, in this way, devices are not bound to operate in a specific area. Albeit the processing power is minimal, it is sufficient to provide some automation. WSN-based systems are still centralized, being controlled or used by more complex and powerful devices. Sometimes [11] such devices can be attached to the system to act as a gateway or provide services to users across the Internet. Home automation is one of the most widespread applications of such technology.

WSNs are a key enabling technology in the evolution of IoT as the presence of a network architecture facilitates the integration in a larger framework (i.e., the IoT infrastructure) and the provision of services. Services that could be technically provided through the means of WSNs are very appealing and pervasive of human life. Being potentially very pervasive and directly impacting users' lives, securing such systems must be taken into account. A general overview of security in WSNs is given in [12].

Authorization prior to inclusion of nodes in a WSN is very important as, even if other (higher level) security mechanisms could prevent rogue nodes from deciphering or injecting packets, they could easily provoke Denial of Service (DoS), for example, by overwhelming the (reduced) network bandwidth.

Authentication should also be taken into account as, failing that, there is a concrete risk of running business processes or providing services on top of malicious data. Authentication should be done against secure PKIs, probably run by Certification Authorities. This will become particularly important for mobile devices as these could provide mobile gating capabilities for less complex smart objects, which cannot have a dedicated connection to the Core Infrastructure. Creating the tools for enabling trust in such a scenario will likely speed the adoption of the IoT paradigm.

Data confidentiality and integrity are also issues for the possible consequences on user privacy and safety. Failing this, private sensor data could be available to malicious users, actuators could be commanded to perform unauthorized actions or the correct functioning of the system could be altered by corrupting packet loads.

Talking about the IEEE 802.15.4 standard, it provides some security features. This standard defines Physical and Medium Access Control layers. First of all, an Access Control List (ACL) can be defined and only frames from the nodes listed in it are admitted to be received. Basic access control, message integrity, message confidentiality and replay protection are provided.

There are different implementation of the 802.15.4, the most interesting of which are ZigBee and SunSPOT. ZigBee provides an application layer security (APS) while SunSPOT users (still at version 0.4) provide SSL. SunSPOT is an exception to the reduced resources characteristic of WSN devices as it works on a 200 MHz ARM7, its OS is open source, it can be programmed in Java, and thus could implement any potential security feature. The only limit is that security features usually draw upon the reduced processing power and bandwidth, producing a significant overhead over the business logic and the messages' payload respectively.

Evolutions in IoT will likely see the presence of PKIs to establish trust between different component devices. According to the current architecture, which is still centralized, object to object connectivity is not contemplated. The 6LoWPAN project [13] aims to bridge this gap providing IPv6 address compression and communication gating low-rate WPANs (i.e., 802.15.4 compliant). PKIs assume a critical role in this context as security features implemented on WSNs must be forwarded to the entire IoT or, where possible, protocols for scaling Internet (i.e., computer) level security features to low rate WPANs (i.e., smart objects) should be provided. Though current devices are not yet sufficiently powerful, this would also open the way to policy regulated service providing as authentication could be provided in both directions.

3 Long-Term Vision

Future devices will likely be as powerful and resourceful as any current mobile or even fixed device and they will have all the privacy and security issues that such current devices have. Miniaturization and increase in spectrum efficiency will enable a denser use of devices which, in turn, will be more sophisticated.

In this new scenario, standardization, semantics, and the availability of smart-object-based services will very likely be the key to the success of this paradigm. In this context, many new private, public, and business services can be conceived. An interesting set of visionary scenarios explored by a high-level visionary Panel of experts of the European Commission can be found in [4]. These efforts toward understanding the future scenario should serve as a base for better extending the current legislation to the new issues brought by the IoT paradigm. Such architecture also poses a great challenge for what concerns its governance. In [14], the European Commission places

> the definition of a set of principles underlying the governance of the Internet of Things and the design of an architecture endowed with a sufficient level of decentralized management

as the first action to be undertaken to promote the evolution of the IoT.

Privacy is also one of the main concerns: issues will arise as data collection, storage, mining, and provision will be completely different from what we now know and legislation shall change accordingly. The number of entities providing services as well as the occasions in which personal data could be collected by such entities

will be greater than what the human user could manage by himself. The solution will be a personal policy-based privacy management system that will automatically negotiate and handle privacy issues for the user according to the rules set by the governance. It will be very important though to provide the user with the right tools for letting him ultimately in control of his own privacy. Such a system might be somehow compared to the identity management system, which is under development by the PRIME project [15].

For example, in order to enforce privacy, devices should have at least two context-based operating modes: public and personal. In the former state, the object should advertise its presence and provide its services to all nearby devices. In the later, it should listen only for other object's presence advertisement and inform about the services it provides only those objects with which a close relationship subsists and public key is already possessed (for example, those belonging to the same user).

In a wider view, smart objects should be able to transparently manage interaction with the environment by using user-defined policies. As previously mentioned, it will be important to have the user in control and that he feels so.

Also, a new set of issues will rise from the high mobility of smart objects and the services they provide. Devices will travel across the world, always and transparently providing the user with their functionalities. To this end, they will locally connect to other objects or gateways. They will have to manage both situations in which they – either directly or not – can access the IoT infrastructure and relative services, and contexts in which they will only be able to communicate with nearby devices. Ad hoc security solutions and policies should be developed for managing mutual authentication, policy enforcement, and basic communication security in both situations.

Security management of nomadic devices will also be an issue. Mobile or worn smart objects, for example, will be able to connect to the infrastructure through different connections in time (some of which may be public while other may be provided by TLC for example). These devices though will need to safely interact with other devices spread across the word, which are not aware of their movement, connection status, etc.

The services too will have to be redesigned as they will become probably portable from one device to a possibly completely different one. It is too soon though to understand how services will actually evolve.

4 Conclusions

It is certain that many other new solutions are needed in order to enable the long-term vision. As for all the newborn visions, there is little agreement in the scientific community about the architectural and technical solutions to adopt. There is though a common feeling that standardization will be one of the key enabling factors. Thus, IoT security design should also follow this trend in order to enable an open, pervasive and interoperable yet secure infrastructure. For the sake of privacy and

flexibility, smart objects should be capable of implementing individual, user set policies. Infrastructural security services should also be accessible transparently and regardless of the connection used by nomadic smart objects.

References

1. Gershenfeld N, Krikorian R, Cohen D (2004) The Internet of things. Sci Am 291(4):76–81
2. Furness A (2008) A Framework Model for The Internet of Things. In: GRIFS/CASAGRAS Workshop, Hong Kong, December 2008
3. Presser M et al. (2008) Real World Internet (Position Paper). Future Internet Assembly, Madrid, Spain, December 2008
4. Hourcade JC, Nuevo Y, Wahlster W, Saracco R, Reinhard P (2009) Future Internet 2020: visions of an industry expert group. Future Internet Final Report, Belgium, May 2009
5. Duc DN, Park J, Lee H, Kim K (2006) Enhancing security of EPCglobal GEN-2 RFID tag against traceability and cloning. The 2006 symposium on cryptography and information security, Hiroshima, Japan
6. Korkmaz E, Ustundag A (2007) Standards, security & privacy issues about radio frequency identification (RFID). RFID Eurasia, 2007 1st Annual, Istanbul, Turkey
7. Chien HY, Chen CH (2007) Mutual authentication protocol for RFID conforming to EPC Class 1 Generation 2 standards. Comput Stand Interfaces 29(2):254–259
8. Peris-Lopez P, Hernandez-Castro JC, Estevez JM, Ribagorda A (2009) Cryptanalysis of a novel authentication protocol conforming to EPC-C1G2 standard. Comput Stand Interfaces 31(2):372–380
9. Garcia FD, de Koning Gans G, Muijrers R, van Rossum P, Verdult R, Schreur RW, Jacobs B (2008) Dismantling MIFARE classic. Proceedings of ESORICS 2008, Malaga, Spain, pp 97–114
10. Aigner M, Dominikus S, Feldhofer M (2007) A system of secure virtual coupons using NFC technology. Pervasive Computing and Communications Workshops, 2007. PerCom Workshops '07. Fifth annual IEEE international conference on, 19–23 Mar 2007, pp 362–366
11. Arch Rock. Arch Rock PhyNetTM. http://www.archrock.com/product/
12. Boyle D, Newe T (2008) Securing wireless sensor networks: security architectures. J Netw (JNW) 3(1):65–77
13. Mullingan G (2007) The 6LoWPAN architecture. Proceedings of the 4th workshop on embedded networked sensors, Cork, Ireland, pp 78–82
14. European Commission (2009) When your yogurt pots start talking to you: Europe prepares for the internet revolution. European Commission's Press Release, June 2009
15. Hansen M, Krasemann H (2008) PRIME whitepaper. Available at https://www.prime-project.eu/prime_products/whitepaper/PRIME-Whitepaper-V3.pdf, May 2008

Privacy Challenges in RFID Systems

Yong Ki Lee, Lejla Batina, and Ingrid Verbauwhede

1 Introduction

Radio frequency identification (RFID) is the technology to identify or track an object using wireless communication. Most RFID systems are composed of three parts: RFID tags, readers, and servers. A tag is attached to an object and it communicates with a reader to transfer its identity and possibly to exchange some data. A server collects and utilizes data of tags via readers. In general, there are two types of tags: active tags and passive tags. Active tags have a battery and can initiate communication, while passive tags derive energy from a reader's carrier signal.

RFID applications are radically expanding from supply chains management and access control and inventory to health care, new-born's safety, road pricing, transport control, etc. [1]. In short, RFID systems are expected to replace the bar code systems completely in near future. However, this expansion induces various security and privacy issues. If no proper security solution is applied, tags can allow for an unauthorized identification and/or tracking. Moreover, most of the international and/or industrial standards use either a simple password-based access control or they have no security at all to keep implementations of a tag very cheap [2, 3]. This can cause serious security and privacy risks since an eavesdropped password can be easily used to compromise or track a tag.

Recently, many security solutions were designed using cryptographic hash functions or private-key encryption algorithms that require less hardware and power resources than public-key algorithms. However, due to the limitations of hash functions and private-key algorithms, they cannot satisfy all the desired properties for general RFID systems, which are system scalability and security against cloning attacks, replay attacks, tracking attacks, and Denial of Service (DoS) attacks [4].

Y.K. Lee (✉) and I. Verbauwhede
University of California, Los Angeles, 56-125B Engineering IV Building 420 Westwood Plaza, Los Angeles, CA 90095-1594, USA
e-mail: jfirst@ee.ucla.edu

L. Batina and I. Verbauwhede
ESAT/SCD-COSIC, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001, Belgium

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_39,

In this chapter, we give an overview of security and privacy solutions for RFID not only for the standards but also in the research community. Depending on the required security and/or operational properties, different cryptographic primitives are needed. In addition, we present our novel authentication protocols, which satisfy all the required properties for RFID systems such as scalability, anticloning, and protection against tracking and impersonation attacks [5].

2 Overview

2.1 Desired Properties in RFID Systems

There are some generally required operational and cryptographic properties for RFID systems as follows:

1. *System scalability*: Some randomized authentication protocols, e.g. [6,7], are not scalable since the computational workload on the server increases linearly with the number of tags. Considering that in general RFID systems include a large number of tags, this is a required property.
2. *Anticloning*: If a group of tags shares the same secret key and uses it for the authentication, the tags are vulnerable to cloning attacks. If an attacker succeeds to crack one of the tags, he can use the revealed secret to clone some other tags. Therefore, a secret key should be pertinent only to a single tag so that a revealed secret key cannot be used for any other tag.
3. *Replay attack (impersonation attack)*: An attacker should not be able to generate a valid set of messages for a new challenge if he does not know the secret keys of a tag. An attacker may actively query a tag and/or perform some polynomial time computation utilizing known information such as the system parameters and the history of exchanged messages.
4. *Anonymity (security against tracking attacks)*: If an attacker can differentiate between different tags from the exchanged messages, he is possibly able to track a tag, and hence its owner, and collect data for malicious purposes. Therefore, the messages should be properly randomized so that it is infeasible to extract any information about a specific tag.
5. *Backward/forward anonymity*: Even if all the information of a tag (including the secret keys) is revealed to an attacker at a certain moment, an attacker should not be able to track a tag in the past or future communications. We put this strong property as an option in the proposed protocols.
6. *Denial of service (DoS) attacks*: In some of the proposed RFID protocols, tags update their secret information to randomize the responses to a reader. In general, the secret updates must be synchronized between a tag and a server. Otherwise, the later authentications will fail since a server cannot recognize the updated secret of a tag. However, a perfect synchronization cannot be guaranteed in RFID systems since it can be easily disturbed by an attacker. The solution to overcome

the DoS attack is that tags block the secret updates after certain number of unsuccessful authentications. However, this causes tracking attacks since the responses of tags become fixed.

2.2 Security of RFID Standards

New standards such as ISO/IEC and EPCglobal emerged owing to growing applications for various industries. The existing standards are established depending on applications and radio frequencies. Some of the most prominent examples are ISO/IEC 14443/15693 for contactless smart cards, ISO 11784∼5/14223 for animal tracking, and ISO/IEC 18000 for item managements. EPCglobal has announced four standards, which are all for item management with passive RFID tags: Class-0 UHF (Ultra High Frequency), Class-1 Generation-1 HF (High Frequency), Class-1 Generation-1 UHF and Class-1 Generation-2 UHF [8]. Although standards from EPCglobal are industrial standards, they draw great attention from the RFID community. Especially, EPCglobal class-1 Gen-2 has been standardized as the ISO/IEC 18000-6C in 2006. This cooperation of ISO/IEC and EPCglobal results in more confidence of EPCglobal for RFID vendors and wider variety and lower prices for end-users.

The security features of the major standards are summarized in Table 1 [2, 8]. In most of the standards, authentication features are based on a simple password system and many others do not have any protection. Therefore, the security can be easily compromised since a password can be simply eavesdropped. For the privacy protection, a tag is killed when a product is purchased by an end-user. However, simply killing a tag is not desirable since there are many situations and environments where a tag can be utilized after the purchase [9, 10]. The cover-coding in EPCglobal Class-1 Gen-2 is used to mask reader-to-tag communications, where a reader performs an XOR operation for data encryption with a random number from a tag. Then, a tag can recover the received messages by doing another XOR operation. Assuming that the signal from a tag to a reader is too weak to be eavesdropped by an

Table 1 Security features in RFID standards

Standards	Security features
EPCglobal Class-0 Gen-0 UHF	Self destruct feature (24-bit password)
EPCglobal Class-1 Gen-1 HF	Self destruct feature (24-bit password)
EPCglobal Class-1 Gen-1 UHF	Self destruct feature (8-bit password) Cover-coding for reader-to-tag communication
EPCglobal Class-1 Gen-2 UHF	Self destruct feature (32-bit password) Access control (32-bit password)
ISO/IEC 18000-3	48-bit password for reading
ISO/IEC 18000-2/14443/15693, ISO 11784∼5	None in standard

attacker, this cover-coding scheme is an effective encryption scheme. However, an enhanced receiver or an implanted receiver near to a tag can make the cover-coding scheme useless.

2.3 Security for Non-standard RFID

In order to protect the privacy of tags, many solutions are proposed in the literature using different cryptographic primitives. Cryptographic primitives can be classified into four categories: Random number generators (RNG), cryptographic hash functions, private-key algorithms, and public-key algorithms. Most of the RFID standards use only RNG, which is not sufficient to provide the requirements. Therefore, nonstandard solutions rely on stronger cryptographic primitives to enable stronger security and privacy protection.

First of all, in order to protect from replay attacks, authentications should be a challenge-response type. A generic challenge-response RFID protocol can be defined as in Fig. 1 [4] where k is private information such as a tag's secret key and ID. In order to facilitate the privacy requirements, the function f must be a one-way function whose output is undistinguishable from random. This function can be constructed with one of the cryptographic primitives.

2.3.1 Protocols Using Hash Functions

The function f in Fig. 1 can be replaced with a hash function as shown in Fig. 2, where multiple hash inputs need to be combined with some operations such as the string concatenation or the bitwise XOR operation.

In this case, however, the system is not scalable since a reader (or server) needs to compute $H(\mathrm{ID}_i, c, r)$ for each tag index i until a match is found. If there is a match, a tag is recognized as a valid one, otherwise rejected. Some relevant works can be found in [6, 7, 11–16]. Some of the published work does not satisfy either

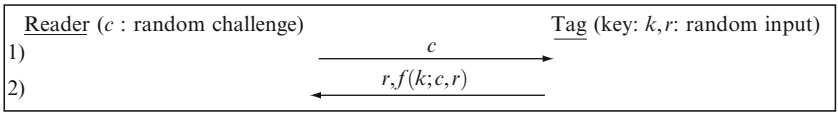


Fig. 1 A generic challenge-response RFID protocol

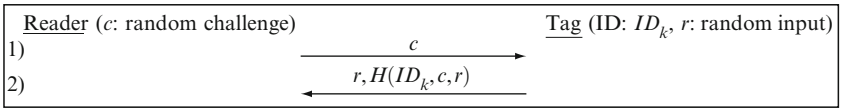


Fig. 2 Hash based randomized access control (H-RAC)

system scalability or the privacy requirement, and the others are vulnerable to the Denial of Service attacks. If the privacy is required, the system eventually becomes unscalable as shown in Fig. 2. This is due to the use of hash functions.

2.3.2 Protocols Using Private-Key Algorithms

In order to obtain the system scalability, the function f should be invertible. By applying a secret-key algorithm, a tag can transfer its ID encrypted and a reader can decrypt the messages. However, if a tag uses its own secret key, which is different from the others, a reader needs to apply every possible key of all tags since a reader does not know a tag's ID at the moment. If a reader finds a proper key, he can verify a tag's ID by decrypting the messages. However, this procedure makes the system unscalable. Therefore, in order to overcome this problem, the secret key must be shared among tags so that a reader can apply the same secret key for any tag, which makes the system vulnerable to cloning attacks. The protocol as described in Fig. 3, where SE_K is a private-key encryption with the private key K, can be used. Note that the transmission of r in plain text is not necessary since the message can be decrypted without it. Some relevant work can be found in [17–19]. However, none of the previous work was able to overcome the limitation of private-key algorithms.

2.3.3 Protocols Using Public-Key Algorithms

In order to satisfy the desired cryptographic and operational properties mentioned in this paper, a public-key algorithm is indispensable as shown in [4]. A public-key algorithm can be directly applied to a generic challenge-response RFID protocol as in Fig. 4, where R_{SK} and R_{PK} are a reader's secret and public key pair. Since a reader can use the same decryption key for any of tags, the system is scalable while maintaining the security against cloning attacks.

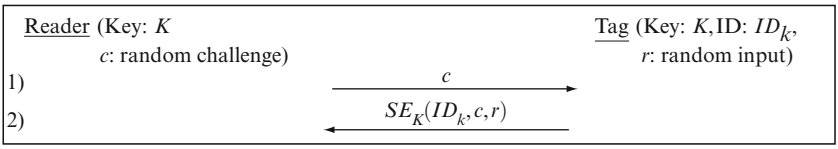


Fig. 3 Secret-key based randomized access control (SK-RAC)

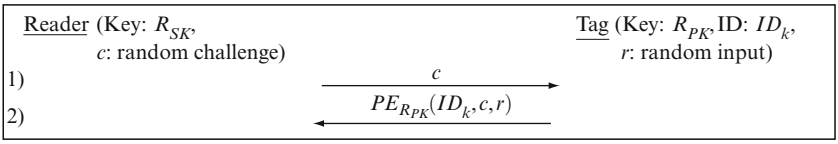


Fig. 4 Public-key based randomized access control (PK-RAC)

In many cases, however, authentication protocols are separately designed instead of directly applying a public-key encryption algorithm (as in Fig. 4) due to its high computational complexity. Some well-known public-key based authentication protocols are the Schnorr protocol [20] and the Okamoto protocol [21]. These protocols have security proofs in a classical model, but they are not proper for RFID systems since the vulnerability to tracking attacks remains, as shown in [22].

A compact public-key processor that is suitable for Fig. 4 can be found in [23], where a variant of Rabin cryptosystem is used. The EC-RAC protocol for efficient RFID authentication is proposed in [22]. However, it was broken in [24], and the randomized Schnorr protocol is proposed for the replacement. The revision of EC-RAC and its security analysis are presented in [5].

2.3.4 Comparison of Cryptographic Features

A comparison of cryptographic features for RFID systems is summarized in Table 2. The level 0 and level 1 are covered by some international standards, and the others can not be covered by standards since they require more complex cryptographic primitives. Depending on the allowed cryptographic primitives, achievable properties differ, and in order to satisfy all the properties, a public-key algorithm is required. However, some of the properties may not be needed depending on the application. For example, in systems with remote car key immobilizers, the number keys (tags) is not so big, so H-RAC could be enough.

In the remainder of this paper, we discuss Elliptic Curve based authentication protocols, which are presented in [5].

Table 2 Security and privacy features for RFID

Security level	Level 0	Level 1	Level 2	Level 3	Level 4
Primitives	Nothing	RNG	RNG, hash function	RNG, private-key	RNG, public-key
Authentication	Password	Cover coding	C/R	C/R	C/R
Replay attacks	Vulnerable	Vulnerable	Secure	Secure	Secure
Cloning attacks	Vulnerable	Vulnerable	Secure	Vulnerable	Secure
Tracking attacks	Vulnerable	Vulnerable	Secure	Secure	Secure
System scalability	Scalable	Scalable	Un-scalable	Scalable	Scalable
Examples	EPCglobal, ISO/IEC 18000-3	ISO/IEC 18000-6C	H-RAC	SK-RAC	PK-RAC, [5, 24]

Cover coding: password can be transmitted with cover coding
C/R challenge/response
[24]: Randomized Schnorr protocol, [5]: Revised EC-RAC

3 Authentication Protocol Design

3.1 System Parameters

RFID systems are in a somewhat different situation from conventional protocols. Unlike conventional authentication protocols, a tag's ID or public-key is not public information since revealing the ID (or public-key) on the fly can cause tracking attacks. Therefore, we also call a public key of a tag a *verifier*. Moreover, RFID protocols are many to one protocols, i.e. many RFID tags communicate with one server. Because of this, tags' verifiers can be kept securely in the server in order to be used for authentications.

First, we assign each tag two secret keys, x_1 (ID) and x_2 (password), similarly to conventional password protocols. Note that the ID is also secret information just like the password. The corresponding ID-verifier and password-verifier, $x_1 P$ and $x_2 P$, are securely stored in the server unlike general public-keys. For the attacking model, we suppose that an attacker knows the system parameters which can be revealed by cracking any of the tags. The system parameters and the storage of each entity are summarized in Table 3. Note that the base point P must be chosen to have a prime order as required by ECC standards [25, 26].

We consider authentication protocols as the combination of the secure ID transfer scheme and the secure password transfer scheme. These parts can be independently designed and analyzed, and can be composed differently depending on the application.

3.2 Component Design

In the ID transfer scheme (Fig. 5), a tag generates a random number r_{t1} and T_1, and transfers T_1 to the server. Then, the server responds with a random challenge r_{s1}, and a tag produces and transfers T_2 to the server. Finally, the server calculates a tag's ID-verifier $x_1 P (= X_1)$.

It is possible to use only the ID transfer scheme for a tag's authentication. The server may authenticate a tag by checking the existence of the decrypted ID-verifier

Table 3 System parameters

System parameters	y : Server's secret-key	$Y(= yP)$: Server's public-key
	x_1 : Tag's ID	$X_1(= x_1 P)$: Tag's ID-verifier
	x_2 : Tag's password	$X_2(= x_2 P)$: Tag's password-verifier
	P : Base point in the EC group	n : Prime order of P
Server's storage	y, X_1, x_1, X_2, P, n	
Tag's storage	x_1, x_2, Y, P, n	
Attacker's storage	Y, P, n : Publicly known information	

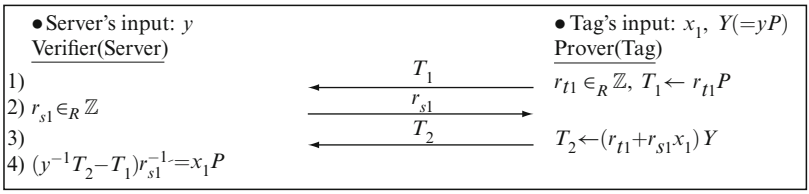


Fig. 5 Secure ID transfer (EC-RAC 1)

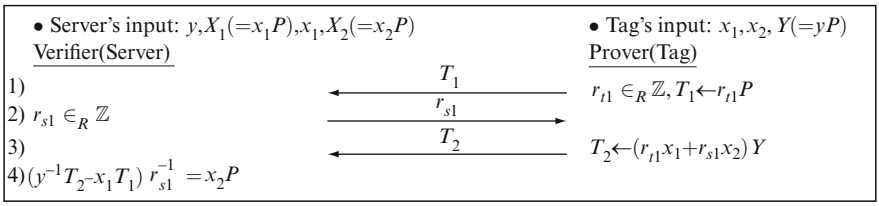


Fig. 6 Secure password transfer

in the list. However, a large number of tags may weaken the security level of the system since the probability that a randomly selected ID is identified as a valid one increases with the number of tags. Therefore, if the number of used tags is large, the password transfer scheme should be added.

Since the password transfer scheme (Fig. 6) is performed after the ID transfer scheme, the server already knows the tag's ID-verifier (X_1). Therefore, the server can look for x_1 and X_2, which are paired with X_1 in the local database.

3.3 Authentication Protocol Construction

We propose three schemes that can be combined differently as summarized in Table 4, where [24] is included for comparison. In EC-RAC 1, we are just using the ID transfer scheme for a tag's authentication. The server authenticates a tag by checking the existence of a tag's ID-verifier in the list. This would be effective to minimize the computation workload on a tag if the number of tags is relatively small. Although EC-RAC 1 is comparable to the randomized Schnorr protocol [24] with similar cryptographic properties, EC-RAC 1 has better performance than the randomized Schnorr protocol in a server (the number of EC point multiplication is smaller).

The combination with the secure password transfer scheme can be done in two different ways resulting in EC-RAC 2 and 3 with different amounts of computation and security properties. All the protocols are scalable and secure against cloning attacks, replay attacks, tracking attacks, and DOA attacks.

Table 4 Authentication protocol constructions and their security properties

Protocols		EC-RAC 1	EC-RAC 2	EC-RAC 3	[24]
EC point multiplication	Server	2	4	4	3
	Tag	2	3	4	2
Properties	Number of tags	Small	Large	Large	Small
	Backward/forward un-traceability	Secure	Vulnerable	Secure	Unknown

Common properties: scalability, protection against cloning, replay, tracking, and DoS attacks
Unknown: there is no proof for the backward/forward un-traceability

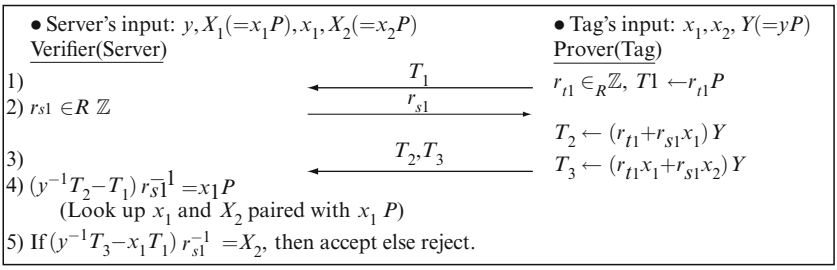


Fig. 7 EC-RAC 2 flow

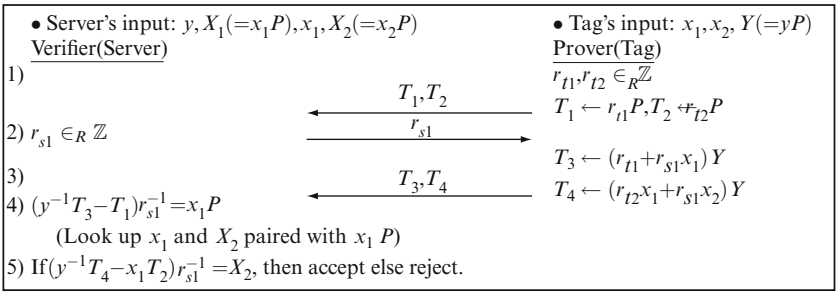


Fig. 8 EC-RAC 3 flow

The first combination is shown in Fig. 7, where the random point $T_1(= r_{t1}P)$ is used not only for the ID transfer scheme but also for the password transfer scheme. This minimizes the amount of computation on a tag, but it results in a weakness against forward/backward tracking attacks.

EC-RAC 2 can be revised to EC-RAC 3 (Fig. 8) to obtain security against forward/backward tracking attacks. In this case, a tag generates two random numbers and each one is used only once to encrypt its ID-verifier or password verifier.

4 Conclusion

We presented an overview of RFID authentication protocols and their achievable properties, which differ depending on the cryptographic primitives used. In addition, several Elliptic Curve based authentication protocols are presented, showing a much stronger properties in RFID systems.

Acknowledgments This work is supported in part by the IAP Programme P6/26 BCRYPT of the Belgian State, by FWO project G.0300.07, by the European Commission under contract number ICT-2007-216676 ECRYPT NoE phase II, by K.U. Leuven-BOF (OT/06/40), NSF CCF-0541472 and SRC.

References

1. Seaner J (2006) EPC/RFID update. http://www.chemicalstrategies.org/pdf/workshop_events/JSeaner_RFID.pdf,EPCglobal
2. Phillips T, Karygiannis T, Kuhn R (2005) Security standards for the RFID market. IEEE Secur Priv 3(6):85–89
3. Razaq A, Luk WT, Shum KM, Cheng LM, Yung KN (2008) Second-generation RFID. IEEE Secur Priv 6(4):21–27
4. Burmester M, Medeiros B, Motta R (2008) Anonymous RFID authentication supporting constant-cost key-lookup against active adversaries. Int J Appl Cryptogr 1(2):79–90
5. Lee YK, Batina L, Verbauwhede I (2009) Untraceable RFID authentication protocols: revision of EC-RAC. In: IEEE international conference on RFID, pp 178–185
6. Ohkubo M, Suzuki K, Kinoshita S (2003) Cryptographic approach to "privacy-friendly" tags. In: RFID Privacy Workshop, MIT, Cambridge, MA
7. Weis SA, Sarma SE, Rivest RL, Engels DW (2003) Security and privacy aspects of low-cost radio frequency identification systems. In: The first international conference on security in pervasive computing – SPC'03
8. Karygiannis T, Eydt B, Barber G, Bunn L, Phillips T (2007) Guidelines for securing radio frequency identification (RFID) systems: Appendix A – RFID standards and security mechanisms. In: NIST Special Publication 800-98, pp A1–A5
9. Garfinkel SL, Juels A, Pappu R (2005) RFID privacy: an overview of problems and proposed solutions. IEEE Secur Priv 3(3):34–43
10. Kumar R (2003) Interaction of RFID technology and public policy. RFID Privacy Workshop, MIT, Boston, MA
11. Avoine G, Oechslin P (2005) A scalable and provably secure hash-based RFID protocol. In: IEEE international workshop on pervasive computing and communication security – Persec'05
12. Burmester M, van Le T, de Medeiros B (2006) Provably secure ubiquitous systems: universally composable RFID authentication protocols. In: IEEE/CreateNet international conference on security and privacy in communication networks – SECURECOMM'06
13. Gao X, Xiang Z, Wang H, Shen J, Huang J, Song S (2004) An approach to security and privacy of RFID system for supply chain. In: IEEE international conference on E-commerce technology for dynamic E-business – CEC-East'04
14. Lee YK, Verbauwhede I (2005) Secure and low-cost RFID authentication protocols. In: IEEE international workshop on adaptive wireless networks – AWiN05, pp 1–5
15. Tan CC, Sheng B, Li Q (2008) Secure and serverless RFID authentication and search protocols. IEEE Trans Wirel Commun, 7(4):1400–1407
16. Tsudik G (2006) YA-TRAP: yet another trivial RFID authentication protocol. In: IEEE international conference on pervasive computing and communications – PerCom'06

17. Feldhofer M (2004) An authentication protocol in a security layer for RFID smart tags. In: IEEE Mediterranean electrotechnical conference – IEEE MELECON'04
18. Feldhofer M, Dominikus S, Wolkerstorfer J (2004) Strong authentication for RFID systems using the AES algorithm. In: Cryptographic hardware and embedded systems – CHES'04, LNCS, vol 3156. Springer, Berlin
19. Toiruul B, Lee K (2006) An advanced mutual-authentication algorithm using AES for RFID systems. Int J Comput Sci Network Secur 6(9B):156–162
20. Schnorr C-P (1989) Efficient identification and signatures for smart cards. In: Advances in cryptology – CRYPTO'89, LNCS, vol 435, Springer, Berlin
21. Okamoto T (1992) Provably secure and practical identification schemes and corresponding signature schemes. In: Advances in cryptology – CRYPTO'92, LNCS, vol 740, Springer, Berlin
22. Lee YK, Batina L, Verbauwhede I (2008) EC-RAC (ECDLP based randomized access control): provably secure RFID authentication protocol. In: IEEE international conference on RFID, pp 97–104
23. Oren Y, Feldhofer M (2008) WIPR – a public key implementation on two grains of sand. In: Conference on RFID security, Budapest, Hungary. http://iss.oy.ne.ro/WIPR, July 2008
24. Bringer J, Chabanne H, Icart T (2008) Cryptanalysis of EC-RAC, a RFID identification protocol. In: International conference on cryptology and network security – CANS'08, LNCS, vol 5339, Springer, Berlin
25. NIST (1999) Recommended elliptic curves for federal government use. http://csrc.nist.gov/groups/ST/toolkit/documents/dss/NISTReCur.pdf
26. SECG (2000) SEC 2: recommended elliptic curve domain parameters. http://www.secg.org/download/aid-386/sec2_final.pdf

Security and Privacy Protection of Contactless Devices

Olivier Savry and François Vacherand

1 Introduction

In the last decade, there was a tremendous growth of contactless applications. Teleticketing and more generally access control were the first large-scale and public initiators, and then credit cards and electronic purse moved to this technology. Finally, e-Identity such as National Identity card and e-Passport started national or worldwide applications. This emerging technology brought new and fair daily usage, but offered new ways of fraudulent or malicious actions directly related to the economical or security aspects of most of the applications.

This chapter focuses on the security of the contactless channel, mainly at the physical level, and on physical attacks and associated countermeasures. The contactless technology has particular characteristics, which were not yet addressed in wireless radio technology. These two technologies share the data transmission through the RF channel, basically bidirectional transmissions. So far, the threats in term of security, which jeopardize the transmission of information are a priori the same ones: remote listening or eavesdropping.

However, there is an important difference: the contactless system has an active element, the reader, and a passive element, the card or the tag, which are remotely powered thanks to the electromagnetic field of the reader. This system, contrary to the traditional radiofrequency emitter/receiver couple, is thus strongly dissymmetrical. This dissymmetry is pointed out by the evidence that there is no ON/OFF switch on these passive devices and that they can be activated automatically without the help of their owner. Transactions can thus start without the awareness of the users, which introduces a more or less strong potential threat depending on the type of application.

Compared to contact smart cards, it is worth pointing out that the RF channel opens a new potential weakness by enabling and performing information communication over the air.

O. Savry and F. Vacherand
Commissariat à l'énergie atomique, LETI, 38054 Grenoble, France
e-mail: olivier.savry@cea.fr; francois.vacherand@cea.Fr

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_40,

First, the article attempts to introduce the economical and societal context of the RFID technology. Then a risk analysis will introduce a state-of-the-art of the attacks relating to this technology. Finally, countermeasures such as the Noisy Reader, which enables preventing eavesdropping on the communication and the Contactless Privacy Manager (CPM), which enables protecting from unauthorized readings will be presented.

1.1 Threats

1.1.1 Economical and Societal Context and Trusted Computing

The main debate on RFID at the societal level is security vs. privacy. Many companies that deploy contactless systems are profit companies and fraud is a severe competitor. Security is basically promoted by operators for evident reason of preventing economical losses or unauthorized accesses to subscription-based services. In some commercial or economical aspects, it is also mandatory to protect the users. The targeted asset of attacks is the business. Fraudulent attacks on the contactless technology have to be nullified.

Privacy is only a matter of concerns for users that will live more and more in a digital world with one or several digital doubles. Here the asset is individual freedoms and more specifically in this case, the protection of digital personal data [1].

With the trumpeted arrival of Ambient Intelligence, digital data privacy is a key point. Several contactless low resources smart devices have shown some connections with privacy concerns:

- RFID electronic tags for items identification in supply chain at retail stores
- e-Citizen contactless cards (e-passport, e-identity, etc.) for person identification

For the daily usage of this emerging and rapidly spreading technology, some questions arise from the basic user, such as:

- Do I communicate with the right reader?
- Is somebody spying my transaction?
- Is somebody trying to communicate with my passive device?
- What are the data exchanged over the air?

Here, it should be pointed out that not only the users but also operators that deploy the system ask quite the same questions. The former want to protect their privacy and the latter to protect their business. In both cases, they need to trust the trusted computing supported by these contactless portable objects and must understand the underlying security in order to trust it.

RFID electronic tags: The concept of ambient intelligence (or AmI) is a vision where humans are surrounded by computing and networking technology unobtrusively embedded in their surroundings. The privacy concerns have been raised very early with the introduction of RFID smart devices. In particular, the introduction and potential generalization of electronic tags on mass markets products and at the

retails level at every shop or supermarket have enlarged the potentiality of traceability of persons. The interoperability of automatic data capture systems and the combination of tagged items carried by a person may be used to trace and to locate this person when moving in a commercial center or urban areas.

E-Citizen contactless cards: Concerns are also very high with the use of contactless e-passport [2, 3]. Some hostile distant reader may check your nationality and trigger attacks. A same scheme may be ruled out with an identity card. High sensitive privacy personal data may be potentially downloaded from contactless health cards without the agreement of the owner.

Among the basic fears that users, consumers and citizens raise when addressing the contactless technology, one can mention:

1. A human person cannot sense the electromagnetic field
2. The identification code is unique. Comparing with the bar code, the RFID code includes the serial number of the items that enables total traceability. Groups of electronic identifiers (RFID) turn out to be a signature of a person and so to be profiled and traced easily
3. RFID readers are small enough to be embedded in a cell phone
4. The localization of person is possible when he is close to a reader
5. Citizens and consumers do not have access to the history of the readings

In that "privacy situation," the main drawback of these contactless systems is that there is no switch on/switch off mechanism to assess the will of the owner to use them. They can be remotely and covertly activated and triggered without the awareness of the owner. So trust starts to shift down.

Privacy is the ability of an individual or group to keep their lives and personal affairs out of public view, or to control the flow of information about them. It is clear that AmI may endanger the privacy of citizens. On the other hand, AmI aims to offer a lot of on-line services that anybody can get only if they identify themselves or the items they are dealing with. So that is the difficulty.

1.1.2 Risk Analysis

Risk analysis must be performed upon contactless systems on two main topics: prevention of economical fraud and protection of privacy. The targeted assets and motivation of attackers are not quite the same. Possibly some countermeasures could help both.

Anyway, in both cases threats and vulnerability must be addressed, focusing on the contactless RF interface. Risk analysis on contact chip is nowadays a stable and very accurate process.

Among vulnerabilities of the contactless interface that are used to perform relevant attacks that jeopardize security and privacy, it is important to list:

- Bidirectional data communications over the air
- Unidirectional power transfer over the air
- Clock transfer over the air
- Passive devices and no on/off switch

- Load based retro modulation
- Singulation protocol
- Misuse of critical commands (i.e., kill command)

1.2 Basic Attacks

1.2.1 Passive Listening

Eavesdropping

In this case, a hacker discreetly listens to a running transaction in order to attempt to retrieve information that he could use in a fraudulent manner at a later date. In that situation, the owner of the contactless device is normally aware that a transaction is being carried out. Basically, data are the contents of the exchanged messages and can be easily recovered if they are not ciphered. It is the case of electronic RFID tags for example.

Data emitted by the card as well as data emitted by the reader can be eavesdropped. Because only the latter one emits an electromagnetic field, it is easier to intercept its broadcasted data messages [2]. Basically, the messages emitted by the reader can be listened at a larger distance (around 20 m) than those emitted by the contactless device (around 4 m). The eavesdropping operational distance depends upon the sensitivity of the receiver and of the receiving antenna.

RFA

In the world of contact cards, a well known class of attacks named single power analysis (SPA), differential power analysis (DPA), or electromagnetic analysis (EMA) enables the hackers to recover the secret key used by the cryptographic processor of the card by recording passively the power consumption variations of the chip, directly or through electromagnetic probes, before running some computation to guess the right key.

This classical side channel attack can be transposed to the contactless world. Moreover, the technique used to modulate the return link can be helpful. The retro modulator is based on the variation of the load of the chip, so equivalently on its power consumption. Because power consumption is basically representative of the current computation work of the chip, the chip activity directly modulates the embedded emitter. Thus, power consumption activity of the chip is directly transmitted through the RF channel. Finally a hacker can register the modulated signal and recover the key, especially if the SPA attack is sufficient to succeed in, such as for RSA exponentiation.

1.2.2 Remote Activation

Skimming

Skimming is an active attack. Because a contactless card has no ON/OFF switch, it is not necessary to ask for authorization from the card holder via a voluntary gesture on his/her part to activate this card. In this case, an unauthorized reader may try to take control of the card. Simple authentication of the reader by the card may solve the problem, except if the attacker has previously got the password or the key [4].

E-passport could be an ideal target of that kind of attack. If not prevented, it is possible to remotely get sensitive information on a person. Of course, some access control functions have been embedded to manage the trouble. Because a human being is not able to sense an electromagnetic field, this kind of attack is very worrying for privacy protection.

Relay

Relay attack is a more serious one because of its basic simplicity [5,6]. The objective is to trigger a normal transaction between a true reader and a true card with a purely transparent relay of communication. The relay is made of a fake card and a fake reader that are connected together with any communication link, fast enough to respect the timeouts of the protocol. Of course, it works because the two linked true contactless devices were designed to work together. With such an attack, it is possible to substitute a fake contactless device with a true one which can be very far from the true reader.

Relay attacks involve two different devices and, as a consequence, two attackers that should coordinate each other except if the relay is really short. The relay attack is based on a specific weakness of the contactless smartcards or RFID tags that is the possibility to activate the device without the consent of the user. The reader will assume that the card, and by implication the user, is in close vicinity [6]. Using this attack on cryptographic authentication schemes, the attacker would be able to convince both reader and card that they share a common secret key.

Man in the Middle

“Man in the middle” attack is an over set of relay attack. It is indeed quite similar but with the distinctive feature that in this attack the bit streams that flow into the relay can be modified. When the communication is ciphered, the attacker has to know the secret key and the algorithm to change the data.

Fault Injection

The basic idea here is to use the RF channel as an entrance door into the chip. It is possible to modify the power, the frequency, or the phase of the carrier in order

to inject an operating fault during the microcontroller code execution in order to disturb the execution of the program and prevent a control to work and try to get confidential information.

1.2.3 Denial of Service

Jamming

With very basic equipment, it is possible to prevent a reader or a tag to work with jammers. The idea is to emit a noisy signal in the same bandwidth as the reader or the tag RF channel in order to blur the communication. Only few watts are required to do that.

Destruction

This attack consists in destroying some part of the contactless system in order to prevent any future operation in an irreversible way. Well known destructive attacks through the RF channel are:

1. High power RF field. According to the current RFID standards, contactless cards must resist to a high electromagnetic up to a given threshold. Beyond this threshold, the manufacturer doesn't and cannot guarantee the reliability of the product.
2. Misuse of the kill command. Some electronic tags such as ePC products support the kill command, the purpose of which is to inhibit definitively the tag for any further operation in order to preserve privacy. Malicious use of that possibility may endanger tags which is no so difficult because the command is only protected with a 32 bits password.

Blockers

With blockers [7], a new class of attacks is presented: attacks that aim at preventing the right operation of the system. Blockers use the basic protocol in such a way that the transaction does not progress anymore and loops endlessly in a preliminary identification phase. Anti-collision phase of the protocol is a nice place to do that. At that stage of the protocol, the reader tries to identify all the tags that are present in the electromagnetic field. To complete its inventory the reader checks if there is still a remaining tag. It is possible to abuse the reader with a device that does not go silent when the reader said it is identified, or that emulates all the possible identification codes or that triggers collisions continuously.

2 Countermeasures

2.1 The RFID Noisy Reader: How to Prevent from Eavesdropping on the Communication?

Communication eavesdropping is one of the main threats in RFID systems. Encryption of the data is in fact expensive in term of computing resources and is time consuming whereas contactless cards and especially RFID tags are limited in energy supply [8]. Some papers assert that the communication between a card and a reader could be eavesdropped from up to 4 m. The goal of the noisy reader is to create a secure channel on the physical level without modifying the RFID standards and so by using the existing RFID tags or contactless cards [9, 10].

2.2 Principle of the Noisy Reader

The proposed mechanism stems from the observation that, while responding, a tag is powered by a constant amplitude signal emitted by the reader. The idea (illustrated in Fig. 1) is to engineer a reader which, during the tag reading, emits a suitably shaped "noisy" signal instead of a constant amplitude one. This noisy signal is in turn modulated in the tag response. Since the reader is the only one that knows the specific noise pattern, it is also the only one that can subtract the (modulated) noise

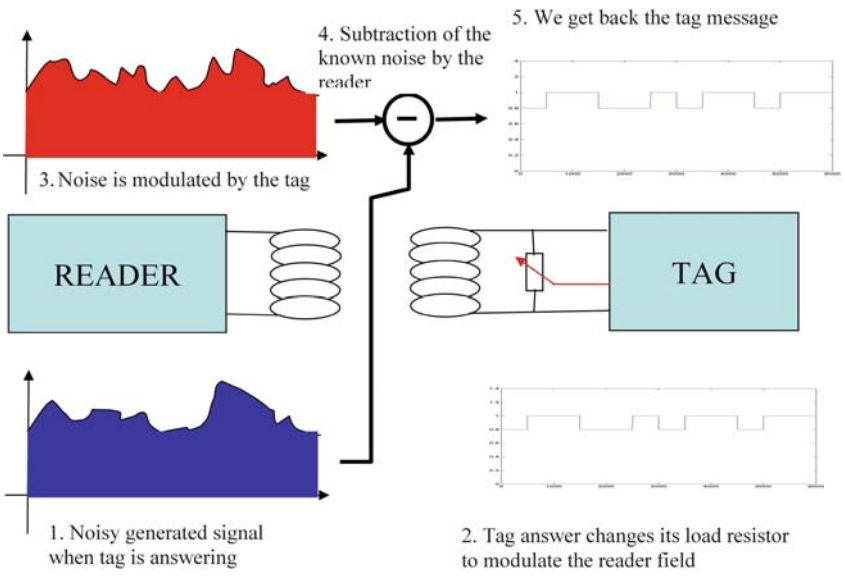


Fig. 1 Principle of the noisy reader

from the actual tag response and "decrypt" the information transmitted by the tag. Any other spying reader in the field will not be able to filter out the noise and hence properly read the tag response.

2.3 Technical Design and Implementation

The solution was developed for the ISO 14443 standard (type B). One of first problem to solve was to define a proper noise that should be in the same spectral bandwidth as the tag to prevent from an easy filtering from a hacker and to keep energy efficiency. Thus as a tag, the noise should use a subcarrier at 847 kHz. Moreover, since some tags not only modulate their amplitude with a resistive load but also their phase with a capacitive load, the noise should be modulated in phase and amplitude. To perform those requirements, the noise is generated by using two random numbers that are picked from a Tausworthe RNG. This choice has been governed by the quick implementation of this solution but for more security especially against statistical attacks, a DES already implemented in the reader will be used. This RNG is fed by three seeds randomly calculated with two unsynchronized counters and by using an IQ configuration to modulate the phase and amplitude.

The noise is then emitted via its own antenna that is in null coupling with the reader antenna. These two antennas have a special design assuring the same radiation diagram that will prevent from an easy spatial noise separation. Thus, the reader will not see the emitted noise while a spying probe in the field will be jammed. Experiments show that this coupling was not sufficient to subtract the noise in the reader. Indeed, the card or even the spying probe will reflect an image of the noise to the reader. Then, we implemented a subtraction chain with a correlation with the subcarrier just after a period of calibration to recover the emitted message.

2.4 Performances

Performances were first assessed in the case of a spy probe positioned directly on top of the RFID tag card. Our measurements show (see Fig. 2) that the decoding performance depends on the noise level.

A too small noise power results in both the reader and the spy to be able to perfectly decode the information. Rather, a too large noise power makes impossible, for both the noisy reader and the spy, to decode information. Figure 2 shows that there exists a noise range (i.e., an utilization "window") where the reader remains able to decipher the tag message with almost 100% of success while the spying probe displays almost 0% of decoded frames. Clearly, the "optimal" amount of noise depends on the distance between the reader and the tag. To this purpose, dynamic adjustment of the proper noise power level was accomplished by leveraging the information provided by the AGC (automatic gain control), already implemented in the reader.

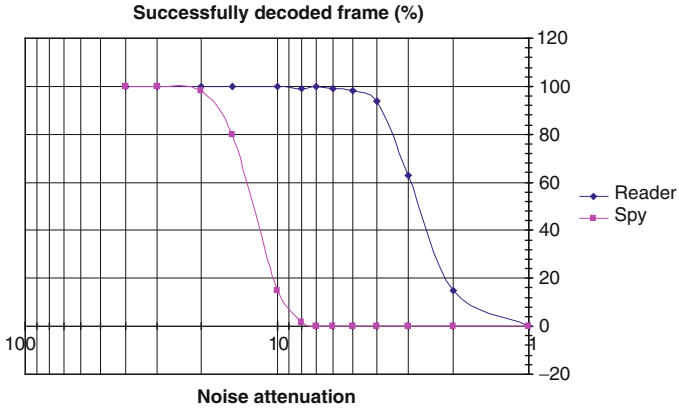


Fig. 2 Noisy reader performance

2.5 *The Contactless Privacy Manager: How to Protect from Unauthorized RFID Readings?*

A fundamental threat emerging in RFID systems is the fact that information stored in RFID tags can be easily read by any attacker equipped with an ordinary reader. While cryptographic-based solutions can be envisioned to restrict the reading only to authorized readers/entities, these would go along with a significant extra cost, with a supplementary crypto-management burden, and with the need to employ nonlegacy RFID tags [11].

2.6 *Contactless Privacy Manager*

To face this issue, we engineered a device called "Contactless Privacy Manager" (CPM) [4, 12]. The CPM can be either directly carried by the user on the body (similarly to any other tag), or conveniently integrated in a portable user device such as a mobile phone. Goal of this device is to intercept unauthorized reading attempts, and jam the relevant tag responses by artificially creating a collision over the air interface and preventing the tag reading. As a result, the CPM creates a virtual protecting sphere against reading attempts from inquisitive RFID readers.

The CPM operation differs from previous solutions, such as the well known "blocker tag" approach, in two fundamental aspects. A first, functional, difference is that the CPM is flexibly designed to permit the user to explicitly and dynamically (run-time) decide which tags should be authorized for reading and which other tags should instead be protected. A second, technical, difference is that the CPM is devised for RFID standards employing a time-slotted singulation (anti-collision) protocol.

2.7 Designed Solution

A proof-of-concept implementation of the Contact-less Privacy Manager has been developed for the ISO 15693 RFID standard at the 13.56 MHz frequency band. To prevent that multiple tags in the field of the reader respond at a same time, this standard implements a time-slotted singulation (anti-collision) protocol. Specifically, the response time following an "inventory" command issued by the reader is divided in 16 time-slots. Tags will answer in a specific time slot depending on the bit mask issued by the reader in the inventory command and their identifiers (UID).

The idea of the CPM is to keep in its memory a list of "authorized" tags. Based on these stored authorized tag UIDs, the CPM is able to subdivide the time slots in two categories: the "authorized" time slots, in which an authorized tag may transmit, and the "forbidden" slots in which no authorized tag transmission is deemed to occur.

The CPM is composed of two parts: a "reader" part whose goal is to detect the occurrence of another tag's transmission inside a time slot (thanks to the detection of a SOF – Start of Frame – symbol), and achieve perfect synchronization with the considered transmission, and a "tag" part. If a transmission is detected in a forbidden slot (meaning that a nonauthorized tag is responding), the CPM tag part will jam such response by transmitting a "dummy" (meaningless and artificially generated) UID, thus causing a collision.

Conversely, if a transmission is detected in an authorized slot, the CPM will transmit the UID of the corresponding authorized tag. Thanks to the achieved synchronization, this will allow an external reader to properly decode the signal in the case the responding tag is in fact the authorized one. Note that if another tag (i.e., a nonauthorized one) transmits in the authorized slot, the CPM transmission of the authorized UID (hence differing from that of the nonauthorized tag) will again cause a collision. Furthermore, we remark that the CPM limits its transmission only to the UID of the tag and not the different preambles (such as flags or CRC) of the standard frame. Thus, if the authorized tag is in fact not in the field, the reader reads only a part of the frame from the CPM and then it cannot understand the message (hence it properly concludes that the transmission was an error).

3 Conclusion

In this chapter, we introduced the security and privacy threats, which could jeopardize the development and deployment of the RFID technology. A survey of the basics attacks on an RFID system enables us to determine its weaknesses. Then, we focused on two mains issues. First, we developed the Noisy Reader to prevent from the eavesdropping on the communication by emitting noise during the contactless device response. Secondly, we focused on the skimming threat by defining a third object we named the CPM, which is able to protect from unauthorized readings of the tags or contactless cards we may have on us.

References

1. Federal Office for Information Security (2004) Security aspects and prospective applications of RFID systems. Germany
2. Ko G, Karger P (2004) Preventing security and privacy attacks on machine readable travel documents. In: Security and Privacy for Emerging Areas in Communications Networks, SecureComm 2005, pp. 47–58, University of Columbia and IBM Research Division
3. Schneier B (2005) Fatal flaw weakens RFID passports. In: Wired News, n°69453
4. Savry O, Vacherand F, Crochon E (2004) Contactless privacy protection device. Patent WO2006/035177
5. Kfir Z, Wool A (2004) Picking virtual pockets using relay attacks on contactless smartcard systems. In: Security and Privacy for Emerging Areas in Communications Networks, SecureComm 2005, pp. 47–58
6. Hancke G (2004) A practical relay Attack on ISO 14443 Proximity Cards. In: IEEE Symposium on Security and Privacy (S&P'06)
7. Juels A et al (2003) The blocker tag: selective blocking of RFID tags for consumer privacy. In: 8th ACM Conference on Computer and Communications Security, pp. 103–111, ACM Press
8. Garfinkel SL, Juels A, Pappu R (2005) RFID privacy: An overview of problems and proposed solutions. In: IEEE security and privacy, vol. 3, no. 3, pp. 34–43, IEEE Computer Society
9. Castelluccia C, Avoine G (2006) Noisy tags: a pretty good key exchange protocol for RFID. In: Domingo-Ferrer J, Posegga J, Schreckling D (eds.) CARDIS No7, vol. 3928, pp. 289–299, Springer-Verlag, Tarragona, ESPAGNE (2006)
10. Savry O, Pebay-Peyroula F, Reverdy J, Robert G (2007) The RFID noisy reader: how to prevent from the eavesdropping on the communication. In: Paillier P, Verbauwhede I (eds.) Cryptographic Hardware and Embedded Systems - CHES 2007, vol. 4727, pp. 334–345, Springer
11. Kirschenbaum I, Wool A (2006) How to build a low-cost, extended-range RFID skimmer. In: Proceedings of the 15th conference on USENIX Security Symposium, vol. 15, USENIX Association
12. Rieback M et al (2006) A platform for RFID security and privacy administration. In: Proceedings of the 20th conference on Large Installation System Administration, pp. 8–16, Usenix Association

Private Location-Based Information Retrieval via k-Anonymous Clustering

David Rebollo-Monedero, Jordi Forné, and Miguel Soriano

1 Introduction

The right to privacy was recognized as early as 1948 by the United Nations in the Universal Declaration of Human Rights, Article 12. With the advent of the Internet of things, according to which the Internet connectivity paradigm shifts toward almost every object of everyday life, privacy will undeniably become as crucial as ever. In this spirit, we consider a particular application of location-based Internet access, which will serve as motivation for an architecture of private information retrieval, where anonymity is attained by means of clustering of user coordinates.

Specifically, consider Internet-enabled devices equipped with any sort of location tracking-technology, frequently operative near a fixed reference location, for example a home computer or a cell phone that is most commonly used from the same workplace. Suppose that such devices access the Internet to contact information providers, occasionally to inquire about location-based information that does not require perfectly accurate coordinates, say weather reports, traffic congestion, or local news and events. Even if authentication to the information providers is carried out with pseudonyms or authorization credentials, accurate location information could be exploited by the providers to infer user identities, for example with the help of an address directory such as the yellow pages. Analyzing both location-based and location-independent queries coming from these devices, information providers could profile users according to their queries, in terms of both activity and content, thereby compromising their privacy.

At this point, we would like to describe a possible mechanism to counter this, at a functional level, solely to motivate our work. A *trusted third party* (TTP) collects accurate location information corresponding to the home location of these devices,

D. Rebollo-Monedero (✉), J. Forné, and M. Soriano
Information Security Group, Department of Telematics Engineering, Universitat Politècnica de Catalunya (UPC), E-08034 Barcelona, Spain
e-mail: david.rebollo@entel.upc.edu; jforne@entel.upc.edu; soriano@entel.upc.edu

M. Soriano
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Parc Mediterrani de la Tecnologia (PMT), 08860 Barcelona, Spain

D. Giusto et al. (eds.), *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*, DOI 10.1007/978-1-4419-1674-7_41,

possibly already publicly available in address directories. This party performs k-anonymity clustering of locations, that is, group locations minimizing the distortion with respect to centroid locations common to k nearby devices. Intuitively, while the same measure of privacy may be applied to all devices, devices with a home location in more densely populated areas should belong to smaller clusters and enjoy a smaller location distortion. The devices trust this intermediary party to send them back the appropriate centroid, which they simply use in lieu of their exact home location, and together with their pseudonym, to access *location-based service* (LBS) providers. Ideally, the TTP would carry out all the computational work required to cluster locations while minimizing the distortion, in a reasonably dynamic way that should enable devices to sign up for and cancel this anonymization service based on the perturbation of their home locations.

In this chapter, we develop a multidisciplinary solution to the application of private retrieval of location-based information motivated above. Our solution relies on a location anonymizer, is based on the same privacy criterion used in microdata k-anonymization, and provides anonymity through a substantial modification of the Lloyd algorithm, a celebrated quantization design algorithm, endowed with a numerical method to solve nonlinear systems of equations inspired by the Levenberg–Marquardt algorithm.

In summary, we consider location-aware devices, commonly operative near a fixed reference location. Accurate location information is collected by a trusted third party and our modification of the Lloyd algorithm is used to create distortion-optimized, size-constrained clusters, where k nearby devices share a common centroid location. This centroid location is sent back to the devices, which use it whenever they need to contact location-based information providers, in lieu of the exact home location, to enforce k-anonymity.

This chapter is organized as follows: Section 2 reviews the state of the art on privacy in LBSs. An architecture for k-anonymous retrieval of location-based information is proposed in Sect. 3. Section 4 develops a modification of the Lloyd algorithm for distortion-optimized, size-constrained clustering to implement the key functionality of the architecture described. Conclusions are drawn in Sect. 5.

2 State of the Art on Privacy in LBSs

The simplest form of interaction between a user and an LBS provider involves a direct message from the former to the latter including a query and the location to which the query refers. An example would be the query "Where is the nearest bank from my home address?," accompanied by the geographic coordinates or simply the address of the user's residence. Under the assumption that the communication system used allows the LBS provider to recognize the user ID, there exists a patent privacy risk, namely, the provider could profile users according to their locations, the contents of their queries, and their activity.

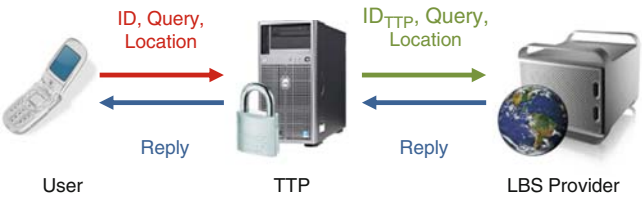


Fig. 1 Anonymous access to an LBS provider through a TTP

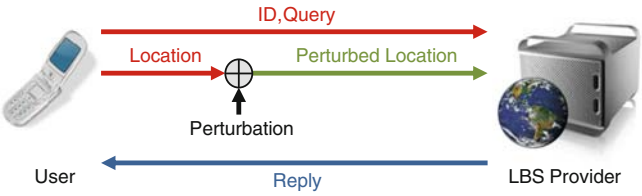


Fig. 2 Users may contact an untrusted LBS provider directly, perturbing their location information to help protect their privacy

An intuitive solution that would preserve user privacy in terms of both queries and locations is the mediation of a TTP in the location-based information transaction, as depicted in Fig. 1. The TTP may simply act as an *anonymizer*, in the sense that the provider cannot know the user ID, but merely the identity ID_{TTP} of the TTP itself inherent in the communication. Alternatively, the TTP may act as a *pseudonymizer* by supplying a pseudonym ID' to the provider, but only the TTP knows the correspondence between the pseudonym ID' and the actual user ID. A convenient twist to this approach is the use of *digital credentials* [1–3] granted by a trusted authority, namely digital content proving that a user has sufficient privileges to carry out a particular transaction without completely revealing their identity. The main advantage is that the TTP need not be online at the time of service access to allow users to access a service with a certain degree of anonymity.

Unfortunately, this approach does not prevent the LBS from attempting to infer the real identity of a user by linking their location to, say, a public address directory, for instance by using *restricted space identification* (RSI) or *observation identification* (OI) attacks [4]. In addition, TTP-based solutions require that users shift their trust from the LBS provider to another party, possibly capable of collecting queries for diverse services, which unfortunately might facilitate user profiling through cross-referencing inferences. Finally, traffic bottlenecks are a potential issue with TTP solutions, and so is any sort of infrastructure requirement in certain ad hoc networks.

We shall see that the main TTP-free alternatives rely on perturbation of the location information, user collaboration, and user–provider collaboration. The principle behind TTP-free perturbative methods for privacy in LBSs is represented in Fig. 2. Essentially, users may contact an untrusted LBS provider directly, perturbing their location information in order to hinder providers in their efforts to compromise user privacy in terms of location, although clearly not in terms of query contents and

activity. This approach, sometimes referred to as obfuscation, presents the inherent trade-off between data utility and privacy common to any perturbative privacy method.

A wide variety of perturbation methods for LBSs has been proposed [5]. We cannot but briefly touch upon a few recent ones. In [6], locations and adjacency between them are modeled by means of the vertices and edges of a graph, assumed to be known by users and providers, rather than coordinates in a Cartesian plane or on a spherical surface. Users provide imprecise locations by sending sets of vertices containing the vertex representing the actual user location. Alternatively, [7] proposes sending circular areas of variable center and radius in lieu of actual coordinates.

Regarding TTP-free methods relying on the collaboration between multiple users, [8] considers groups of users that know each other's locations but trust each other, who essentially achieve anonymity by sending to the LBS provider a spatial cloaking region covering the entire group. Recall that a specific piece of data on a particular group of individuals is said to satisfy the *k-anonymity* requirement (for some positive integer k) if the origin of any of its components cannot be ascertained beyond a subgroup of at least k individuals. The concept of k-anonymity, originally proposed by the *statistical disclosure control* (SDC) community [9, 10], is a widely popular privacy criterion, partly due to its mathematical tractability. However, this tractability comes at the cost of important limitations, which have motivated a number of refinements [11–14]. As many collaborative methods, the one just described guarantees k-anonymity regarding both query contents and location.

Another effort toward k-anonymous privacy in LBSs, this time without the assumption that collaborating users necessarily trust each other, is that of [15]. Fundamentally, k users add zero-mean random noise to their locations and share the result to compute the average, which constitutes a shared perturbed location sent to the LBS provider. Unfortunately, some of these users may apply noise cancellation to attempt to disclose a slow-changing user's location. To counter this, privacy homomorphisms may prove more convenient in the computation of this shared perturbed location [16].

A third class of TTP-free methods such as [17] builds upon cryptographic methods for *private information retrieval* (PIR) [18], which may be regarded as a form of untrusted collaboration between users and providers. Recall that PIR enables a user to privately retrieve the contents of a database, indexed by a memory address sent by the user, in the sense that it is not feasible for the database provider to ascertain which of the entries was retrieved [18]. Unfortunately, PIR methods require the provider's cooperation in the privacy protocol, are limited to a certain extent to query-response functions in the form of a finite lookup table of precomputed answers, and are burdened with a significant computational overhead.

Not surprisingly, a number of proposals for privacy in LBSs combine several of the elements appearing in all of the solutions above. Hybrid solutions more relevant to this work build upon the idea of *location anonymizers*, that is, TTPs implementing location perturbative methods [19], with the aim of hindering RSI and OI attacks, in addition to hiding the identity of the user. Many of them are based on the k- anonymity and cloaking privacy criteria [4, 15, 20–23].

3 A Functional Architecture for k-Anonymous LBSs

Throughout the chapter, the measurable space in which a random variable (r.v.) takes on values will be called *alphabet*. We shall follow the convention of using uppercase letters for r.v.s and lowercase letters for particular values they take on. Probability density functions (PDF) and probability mass functions (PMF) are denoted by p and subindexed by the corresponding r.v.

We formalize the architecture already motivated in Sect. 1. Specifically, and according to the terminology of Sect. 2, we describe a protocol, sketched in Fig. 3, for k-anonymous location-based information retrieval with a location anonymizer. Summarizing Sect. 1, users are assumed to frequently operate near a fixed reference location, which we call home location, represented by values of an r.v. X in an arbitrary alphabet, possibly discrete or continuous, for example, Cartesian or spherical coordinates, or vertices of a graph modeling geographic adjacencies. A TTP playing the role of location anonymizer collects accurate home location information either from the users or from publicly available address directories. This party performs k-anonymous clustering of locations, that is, group locations around centroid locations common to k nearby devices. Users trust this intermediary party to send them back the appropriate centroid, which they simply use in lieu of their exact home location whenever they access LBS providers. The centroid is represented by the r.v. $\hat{X}$, which may be regarded as an approximation to the original data, defined in an arbitrary alphabet, commonly but not necessarily equal to the original data alphabet. For higher privacy protection, users may in addition utilize anonymizers, pseudonymizers, or digital credentials as explained in Sect. 2.

The clustering function implemented by the location anonymizer is depicted in Fig. 3. Precisely, this function is defined to be a quantizer satisfying cell probability constraints, introducing a distortion as small as possible. Formally, the quantizer $q(x)$ is a map assigning X to a quantization index Q in a finite alphabet $\mathscr{Q} = \{1, \ldots, |\mathscr{Q}|\}$ of a given size. The reconstruction function $\hat{x}(q)$ maps Q into the aggregated key attribute value $\hat{X}$.

For any nonnegative (measurable) function $d(x, \hat{x})$, called distortion measure, the associated expected distortion $\mathscr{D} = \mathrm{E}d(X, \hat{X})$ is a measure of the discrepancy between the key attribute values and their aggregation values, which reflects the loss in data utility. $p_Q(q)$ denotes the PMF corresponding to the cell probabilities. A widely popular, mathematically tractable type of distortion measure is $d(x, \hat{x}) = \|x - \hat{x}\|^2$, for which $\mathscr{D}$ becomes the mean-squared error (MSE).

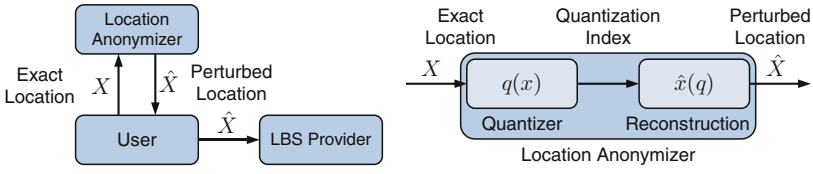


Fig. 3 Architecture with location anonymizer

The k- anonymity requirement in the clustering problem is formalized, from a more general perspective, by means of cell probability constraints $p_Q(q) = p_0(q)$, for any given PMF $p_0(q)$.

We would like to stress that our formulation of the probability-constrained quantization problem may also find applications in microdata anonymization and a variety of resource allocation problems. Nevertheless, we focus on the motivating application of this chapter, namely location k-anonymization. In this important case, let n be the number of home locations to be clustered. The k-anonymity constraint could be translated into probability constraints by setting $|\mathcal{Q}| = \lfloor n/k \rfloor$ and $p_0(q) = 1/|\mathcal{Q}|$, which ensures that $n\ p_0(q) \geqslant k$. More generally, for a given probability p_0, we could naturally speak of p_0-anonymization, a term more suited to continuous probability models of user locations.

Given a distortion measure $d(x, \hat{x})$ and probability constraints $p_Q(q) = p_0(q)$ (along with the specification of the number of quantization cells $|\mathcal{Q}|$), we wish to design an optimal quantizer $q^*(x)$ and an optimal reconstruction function $\hat{x}^*(q)$, in the sense that they *jointly* minimize the distortion $\mathcal{D}$ while satisfying the probability constraints. This problem is addressed in the next section.

4 Modified Lloyd Algorithm for k-Anonymous Clustering

This section investigates the problem of distortion-optimized, probability-constrained quantization, formulated in Sect. 3, and motivated as the functionality implemented by a location k-anonymizer. The quantizer design method proposed is a substantial modification of the *Lloyd algorithm* [24, 25], a celebrated quantization design algorithm, endowed with a numerical method to solve nonlinear systems of equations inspired by the Levenberg–Marquardt [26] algorithm.

Recall that in the context of source coding, quantizers are required due to the need to represent the data in a countable alphabet, such as the set of finite bit strings, suitable for storage and transmission in computer systems. Clearly, quantization comes at the price of introducing a certain amount of distortion between the original data and its reconstructed version. *Optimal quantizers* are those of minimum distortion for a given number of possible indices. The formulation of the problem of minimum-distortion quantization in conventional data compression is identical to the formulation of our probability-constrained version in Sect. 3, without the constraints.

In this section, we propose heuristic optimization steps for probability -constrained quantizers and reconstruction functions, analogous to the nearest-neighbor and centroid conditions found in conventional quantization [27, 28]. Next, we modify the conventional Lloyd algorithm by applying its underlying alternating optimization principle to these steps.

Finding the optimal reconstruction function $\hat{x}^*(q)$ for a given quantizer $q(x)$ is a problem identical to that in conventional quantization [27, 28]:

$$\hat{x}^*(q) = \arg\min_{\hat{x}} \mathrm{E}[d(X, \hat{x})|q]. \tag{1}$$

In the special case when MSE is used as distortion measure, this is the *centroid step* $\hat{x}^*(q) = \mathrm{E}[X|q]$. On the other hand, we may not apply the nearest-neighbor condition in conventional quantization directly, if we wish to guarantee the probability constraints $p_Q(q) = p_0(q)$. We introduce a cell *cost function* $c : \mathcal{Q} \to \mathbb{R}$, a real-valued function of the quantization indices, which assigns an additive cost $c(q)$ to a cell indexed by q. The intuitive purpose of this function is to shift the cell boundaries appropriately to satisfy the probability constraints. Specifically, given a reconstruction function $\hat{x}(q)$ and a cost function $c(q)$, we propose the following *cost-sensitive nearest-neighbor step*:

$$q^*(x) = \arg\min_q d(x, \hat{x}(q)) + c(q). \tag{2}$$

This is a heuristic step inspired by the nearest-neighbor condition of conventional quantization, which states that an optimal quantizer must satisfy $q^*(x) = \arg\min_q d(x, \hat{x}(q))$ [27, 28].

The step just proposed leads to the question of how to find a cost function $c(q)$ such that the probability constraints $p_Q(q) = p_0(q)$ are satisfied, given a reconstruction function $\hat{x}(q)$. For discrete probability distributions of X, it is easy to see that such $c(q)$ may not exist. In the continuous case, we propose an application of the Levenberg–Marquardt algorithm [26], an algorithm to solve systems of nonlinear equations numerically, or similarly but slightly more simply, a Tychonov regularization of the Gauss–Newton algorithm [29], for example with backtracking line search [30] along the descent direction. To estimate the Jacobian more efficiently, slightly increase each of the coordinates of $c(q)$ at a time, exploiting the fact that only the coordinates of $p_Q(q)$ corresponding to neighboring cells may be changed.

Ideally, we wish to find a pair of quantizers and reconstruction functions that *jointly* minimize the distortion. The conventional Lloyd algorithm [24, 25] is essentially an alternating optimization algorithm that iterates between the nearest-neighbor and the centroid optimality conditions, necessary but not sufficient conditions, hoping to approximate a jointly optimal pair $q^*(x)$, $\hat{x}^*(q)$, but only guaranteeing that the sequence of distortions is nonincreasing. Experimentally, the Lloyd algorithm very often shows excellent performance.

Recall that our modification of the nearest-neighbor condition (2) is heuristic, in the sense that this work does not prove it to be a necessary optimality condition. We still use the same alternating optimization principle, albeit with a more sophisticated nearest-neighbor condition, and define the following *modified Lloyd algorithm for probability-constrained quantization*:

1. Choose an initial reconstruction function $\hat{x}(q)$ and initial cost function $c(q)$.
2. Update $c(q)$ to satisfy the probability constraints $p_Q(q) = p_0(q)$, given the current $\hat{x}(q)$. To this end, use the method described at the end of Sect. 4, setting the initial cost function as the cost function at the beginning of this step.
3. Find the next quantizer $q(x)$ corresponding to the current $\hat{x}(q)$ and the current $c(q)$, according to (2).
4. Find the optimal $\hat{x}(q)$ corresponding to the current $q(x)$, according to (1).
5. Go back to step 2, until an appropriate convergence condition is satisfied.

The initial reconstruction values may simply be chosen as $|\mathscr{Q}|$ random samples distributed according to the probability distribution $p_X(x)$ of X. An effective cost function initialization is $c(q) = 0$, because it ensures that the corresponding quantizer cells cannot have zero volume. Note that the numerical computation of $c(q)$ in Step 2 should benefit from better and better initializations as the reconstruction values become stable. If the probability of a cell happens to vanish at any point of the algorithm, this can be tackled by choosing a new, random reconstruction value, with cost equal to the minimum of the rest of costs. The stopping convergence condition might for instance consider slowdowns in the sequence of distortions obtained.

5 Conclusion

According to the vision of the Internet of things, the paradigm of Internet connectivity is expected to shift to almost every object of everyday life. Concordantly, we shall expect privacy, particularly in LBSs, to rapidly gain even greater importance.

In this spirit, here we propose a multidisciplinary solution to an application of private retrieval of location-based information with location-aware devices, commonly operative near a fixed reference location. Our solution relies on a location anonymizer, is based on the same privacy criterion used in microdata k-anonymization, and provides anonymity through a substantial modification of the Lloyd algorithm, a celebrated quantization design algorithm, endowed with a numerical method to solve nonlinear systems of equations inspired by the Levenberg–Marquardt algorithm.

The k-anonymous location clustering mechanism implemented by the location anonymizer is regarded more generally as a problem of minimum-distortion, probability-constrained quantization, which also addresses applications of similarity-based, workload-constrained resource allocation. We extend the Lloyd–Max algorithm from conventional quantization design to probability-constrained quantization. The centroid condition remains the same, but the nearest-neighbor condition is expressed in terms of an additive cost function that shifts cell boundaries to satisfy the probability constraint.

Our framework enables us to represent a quantizer unambiguously and compactly, simply as a list of reconstruction values and costs, one per cell, rather than an arbitrary clustering of a large cloud of points. This is particularly useful when a model of the data is given by means of a PDF, for which a probability-constrained quantizer is to be designed only once, but later on applied repeatedly to dynamic sets of samples distributed according to the original model.

Acknowledgment This work was partly supported by the Spanish Research Council (CICYT) through projects CONSOLIDER INGENIO 2010 CSD2007-00004 “ARES,” TSI2007-65393-C02-02 “ITACA,” and TEC-2008-06663-C03-01 “P2PSec.”

References

1. Chaum D (1985) Security without identification: transaction systems to make big brother obsolete. Commun ACM 28(10):1030–1044
2. Benjumea V, López J, Linero JMT (2006) Specification of a framework for the anonymous use of privileges. Telemat Informat 23(3):179–195
3. Bianchi G, Bonola M, Falletta V, Proto FS, Teofili S (2008) The SPARTA pseudonym and authorization system. Sci Comput Program 74(1–2):23–33
4. Gruteser M, Grunwald D (2003) Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the ACM international conference on mobile systems, applications, and services (MobiSys). ACM, San Francisco, CA, May 2003, pp 31–42
5. Duckham M, Mason K, Stell J, Worboys M (2001) A formal approach to imperfection in geographic information. Comput Environ Urban Syst 25(1):89–103
6. Duckham M, Kulit L (2005) A formal model of obfuscation and negotiation for location privacy. In: Proceedings of the international conference on pervasive computing. Lecture Notes in Computer Science (LNCS), vol 3468. Springer, Munich, Germany, May 2005, pp 152–170
7. Ardagna CA, Cremonini M, Damiani E, De Capitani di Vimercati S, Samarati P (2007) Location privacy protection through obfuscation-based techniques. In: Proceedings of annual IFIP working conference on data and applications security. Lecture Notes in Computer Science (LNCS), vol 4602. Springer, Redondo Beach, CA, Jul 2007, pp 47–60
8. Chow C, Mokbel MF, Liu X (2006) A peer-to-peer spatial cloaking algorithm for anonymous location-based services. In: Proceedings of the ACM international symposium on advances in geographic information systems (GIS), Arlington, VA, Nov 2006, pp 171–178
9. Samarati P, Sweeney L (1998) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. SRI Int Tech Rep, pp 1–19
10. Samarati P (2001) Protecting respondents' identities in microdata release. IEEE Trans Knowl Data Eng 13(6):1010–1027
11. Truta TM, Vinay B (2006) Privacy protection: p-sensitive k-anonymity property. In: Proceedings of the international workshop on privacy data management (PDM), Atlanta, GA, 2006, p 94
12. Sun X, Wang H, Li J, Truta TM (2008) Enhanced p-sensitive k-anonymity models for privacy preserving data publishing. Trans Data Privacy 1(2):53–66
13. Machanavajjhala A, Gehrke J, Kiefer D, Venkitasubramanian M (2006) l-Diversity: privacy beyond k-anonymity. In: Proceedings of the IEEE international conference on data engineering (ICDE), Atlanta, GA, Apr 2006, p 24
14. Rebollo-Monedero D, Forné J, Domingo-Ferrer J (2008) From t-closeness to PRAM and noise addition via information theory. In: Privacy Stat. Databases (PSD). Lecture Notes in Computer Science (LNCS). Springer, Istambul, Turkey
15. Domingo-Ferrer J (2006) Microaggregation for database and location privacy. In: Proceedings of the international workshop on next generation information technologies and systems (NGITS). Lecture Notes in Computer Science (LNCS), vol 4032. Springer, Kibbutz Shefayim, Israel, Jul 2006, pp 106–116
16. Solanas A, Martínez-Ballesté A (2008) A TTP-free protocol for location privacy in location-based services. Comput Commun 31(6):1181–1191
17. Ghinita G, Kalnis P, Khoshgozaran A, Shahabi C, Tan K-L (2008) Private queries in location based services: anonymizers are not necessary. In: Proceedings of the ACM SIGMOD international conference on management of data, Vancouver, Canada, Jun 2008, pp 121–132
18. Ostrovsky R, Skeith III WE (2007) A survey of single-database PIR: techniques and applications. In: Proceedings of the international conference on practice and theory in public-Key cryptography (PKC). Lecture Notes in Computer Science (LNCS), vol 4450. Springer, Beijing, China, Sep 2007, pp 393–411
19. Mokbel MF (2006) Towards privacy-aware location-based database servers. In: Proceedings of the IEEE international conference on data engineering workshops (PDM), Atlanta, GA, p 93

20. Gedik B, Liu L (2005) A customizable k-anonymity model for protecting location privacy. In: Proceedings of the IEEE international conference on distributed computing systems (ICDS), Columbus, OH, Jun 2005, pp 620–629
21. Cheng R, Zhang Y, Bertino E, Prabhakar S (2006) Preserving user location privacy in mobile data management infrastructures. In: Proceedings of workshop on privacy enhancing technologies (PET). Lecture Notes in Computer Science (LNCS), vol 4258. Springer, Cambridge, UK, 2006, pp 393–412
22. Gedik B, Liu L (2008) Protecting location privacy with personalized k-anonymity: architecture and algorithms. IEEE Trans Mob Comput 7(1):1–18
23. Bamba B, Liu L, Pesti P, Wang T (2008) Supporting anonymous location queries in mobile environments with PrivacyGrid. In: Proceedings of the international world wide web (WWW) conference, Beijing, China, Apr 2008, pp 237–246
24. Lloyd SP (1982) Least squares quantization in PCM. IEEE Trans Inform Theory IT-28: 129–137
25. Max J (1960) Quantizing for minimum distortion. IEEE Trans Inform Theory 6(1):7–12
26. Marquardt D (1963) An algorithm for least-squares estimation of nonlinear parameters. SIAM J Appl Math (SIAP) 11:431–441
27. Gersho A, Gray RM (1992) Vector quantization and signal compression. Kluwer, Boston, MA
28. Gray RM, Neuhoff DL (1998) Quantization. IEEE Trans Inform Theory 44:2325–2383
29. Björck A (1996) Numerical methods for least squares problems. SIAM, Philadelphia, PA
30. Luenberger DG, Ye Y (2008) Linear and nonlinear programming, 3rd edn. Springer, New York

Index

D