ST Journal of Research

Wireless Sensor Networks

May 2007

Volume 4, Number 1
Wireless Sensor Networks

LEGAL INFORMATION
DIRETTORE RESPONSABILE
Maria Teresa Gatti

REDAZIONE
STMicroelectronics Srl
Via C. Olivetti 2
20041 Agrate Brianza (Milano) - Italy

LAYOUT AND PRODUCTION
La Carta Stampata
Milano, Italy

PHOTOS
STMicroelectronics - Zerodue
Milano, Italy

STAMPA
C&M Print Sas
Pioltello (Milano), Italy, via Roma 4

Iscrizione presso il Tribunale di Monza
Nr. 1650/2003

Wireless Sensor Networks: a journey towards Ambient Intelligence

Ambient Intelligence is a vision where environment becomes smart, friendly, context-aware and responsive to any type of human needs. In this world, computing and networking technology coexist with people in a ubiquitous and pervasive way. In this world, numerous miniature and interconnected smart devices create a new intelligence and interact with each other seamlessly.

Wireless Sensor Networks (WSNs) are one of the first real world examples enabling that vision of Ambient Intelligence. Within a WSN, sensory data originate from multiple sensors of different types spread over potentially wide areas. Virtually any kind of physical quantities such as temperature, voltage, pressure, acceleration, ultrasound, gas flow, etc. can be acquired. Likewise, actuators can be used as a means to influence or control the environment. For coordination and information exchanges, the various nodes communicate over a wireless mesh network in a self-organized manner. Originally motivated by military applications such as battle-field surveillance, the applications of WSNs are many and varied. Industrial and home automation benefit from the capability to wirelessly monitor data that would be difficult or too expensive to monitor using wired sensors. Likewise, health care or environmental control applications take advantage of the miniaturization and self-organized nature of WSNs. The many potential applications recently motivated major players of the semiconductor industry to invest in this promising field. STMicroelectronics has been among the first companies to enter the WSN market with the introduction of sensors, low-power microcontrollers, wireless radio devices, etc., in its portfolio. STMicroelectronics is a promoter of the ZigbeeTM alliance, an industry consortium working on the definition of network, security and application layers towards truly interoperable WSNs.

One major technical challenge of WSNs is to produce low-cost, low-power and tiny sensor and actuator nodes. Research work is thus needed and involves numerous fields such as Radio Frequency communication, ad-hoc networks, low-power microcontrollers, transduction principles, energy management, middleware, etc. Another key aspect is the coexistence and interoperability of WSNs with existing technologies such as the Internet, broadband wireless networks or database management systems. Given this wide list of topics, this issue of the ST Journal of Research includes several overview papers that help the reader capture the most important challenges as well as a selection of contributions covering specific design topics. Eleven papers have been selected from worldwide recognized experts and ST researchers in the field.

The first section of this special issue includes three tutorial papers. A general introduction to the problems and challenges of developing Wireless Sensor Network devices is given, in the first paper by Chalard et al. from STMicroelectronics. A particular emphasis on ad-hoc networking technologies is proposed in the second paper by Blasi et al. from STMicroelectronics and Bononi from Universita di Bologna. Sheets et al. from the Univeristy of California, Berkeley and Cervini from STMicroelectronics describe in the third paper the main technical challenges faced in the PicoRadio Project carried out at the Berkeley Wireless Research Center.

Four contributions describing the application-specific nature of WSN optimization are included in the second part of the issue. An indoor climate monitoring prototype developed by our STMicroelectronics colleagues illustrate the challenges of integrating WSNs with the Internet via gateway and database client-server tools. Thonet and Bruel from Schneider Electric discuss the mass market adoption of ZigbeeTM solutions in the segments of Building Automation, Home control, Automated Meter Reading and Industrial Automation. The third paper, authored by Yao from UCLA and Lorenzelli from STMicroelectronics, introduces different techniques for source and node localization. In the last paper of this section, Farella et al. from University di Bologna and Urbino expose the reader to the application of WSN and MEMS sensors in the context of body area networks, more specifically posture and activity recognition.

As mentioned above, a WSN device is a complex integrated system implementing a combination of functions such as communication, security and signal or application processing. The last part of the issue includes four papers dedicated to specific design experiences carried out in those areas. Two papers explore the promising use of low data rate UltraWideBand radio communications in the context of WSN. Specifically, Flury et al. from EPFL analyze the impact of impulsive interference on network throughput and energy consumption. Di Benedetto et al. from University La Sapienza focus their work on medium access control schemes proposed for the IEEE 802.15.4a standard. In the third paper Necchi and Lavagno from Politecnico di Torino together with Pandini and Vanzago from STMicroelectronics present their work in trading off power consumption and die area for the design of an asynchronous 8-bit processor. Finally, in the last paper, performances of two Elliptic Curve Cryptography coprocessors are explored by Bertoni from STMicroelectronics together with Breveglieri and Venturi from Politecnico di Milano. Wireless Sensor Networks have the potential to open a new era. However, as illustrated by the broad range of topics covered by this special issue, technical issues still exist in the production of low-cost small devices operating for years in a self-organized way and seamlessly integrated with Internet access and appropriate human interfaces. We believe that only companies having the capability to master this broad spectrum of activities and to succeed in both "more Moore" and "more-than-Moore" directions will provide winning solutions for Ambient Intelligence.

The Guest Editors would like to thank all the Authors as well as all those involved in the review and editorial process, with a special mention to Antoine Hue for his huge contribution.

Julien Zory & Laura Vanzago
Guest Editors

WIRELESS SENSOR NETWORKS DEVICES: OVERVIEW, ISSUES, STATE-OF-THE-ART AND PROMISING TECHNOLOGIES

Laurent Chalard, Didier Helal,
Lucille Verbaere, Armin Wellig, Julien Zory

STMicroelectronics

Revolutionary networking concepts and unprecedented mix of technical challenges have made Wireless Sensor Networks (WSN) one of the major research trends of the 21th century. However, only very recently did such systems start to appear as products on the market. In this new context, this paper aims at providing a general status on WSNs, with a special focus on the device aspects, independent of application peculiarities. We first investigate why, despite years of research and development and technical maturity, WSN products and solutions are yet neither fully adopted nor widely deployed. We then review the state-of-the-art technologies that build most of today's available products and finally briefly discuss some advanced techniques for future generations of WSN systems.

1. INTRODUCTION

Wireless Sensor Network is a generic term for a system made of several autonomous sensors and/or actuators, called nodes, capable of exchanging information with others by means of a dedicated wireless network. Beyond this basic definition, additional specific characteristics are generally assumed de facto. Among others, nodes are expected to be simple and low cost, which implies, for instance, favoring their redundancy rather than their performance; furthermore, they shall be low power, which implies limited communication range and hence use of multi-hop connections; additionally, they are assumed to be autonomous with underlying features like self-configuration and self-healing of the network.

This common understanding of WSN probably originates from its initial research context, supported among others by numerous funded projects in the US (e.g., SmartDust, NEST). The vision was then a kind of self-organized, homogeneous, tiny, multi-hop, and resource-constrained set of devices. Applications were initially mostly military and slowly started to encompass other monitoring fields, like agriculture or species tracking. More recently, several technological advances made such systems a promising alternative for numerous commercial applications as well: home automation, healthcare, industrial plant monitoring,

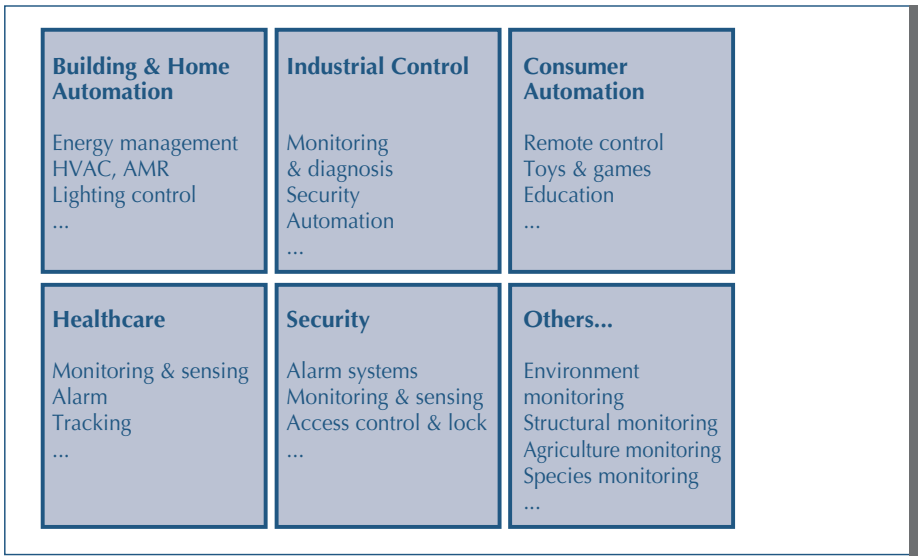

Figure 1: Possible application fields of Wireless Sensor Networks.

or asset tracking, to name a few (see Fig. 1). Such potential mass production markets put great expectations on WSN.

One of the most famous initiatives consolidating the possible deployment of WSN systems was the IEEE802.15.4 Task Group (TG), which specified in 2003 a physical (PHY) and a Medium Access Control (MAC) layer dedicated to low data rate communication devices [1]. The main motivation to develop a dedicated standard, and not to rely on existing technologies like Bluetooth or WLAN, was to ensure low complexity energy-efficient implementations so as to enable multi-year battery lifetime and, hence, to open the door to new application fields (e.g., hard-to-access or non-accessible nodes) and to lower the maintenance costs. Most of today's commercially available radio solutions use the 2.4GHz unlicensed worldwide ISM band, whose regulation allows for 1mW of transmitted power, thus spanning a range between 30 to 100m as a function of the environment. Its packet-based processing allows multiple users via Carrier Sense Multiple Access and Collision Avoidance (CSMA/CA), and its flexibility accommodates various application requirements (e.g., latency, number of devices, network topology, encryption, and duty cycle). Future releases are already under development, mainly via the IEEE802.15.4a TG, which focuses on an Ultra Wide Band (UWB) alternative with location capabilities and the IEEE802.15.4b TG, which enhances the current Narrow Band (NB) standard.

In order to consolidate deployment of low data rate systems based on IEEE802.15.4, as well as to speed up their development, the Zigbee Alliance specified in 2004 a generic protocol stack that provides all networking features up to the application interface [2]. It supports various pre-defined application profiles, e.g., home automation, in order to allow designers to focus on the application while lower layers are configured accordingly. Note that even if the Zigbee Alliance is an initiative of several industry leaders, the specification itself is open and unlicensed. Still, it has not initially been developed for WSN specifically, and numerous proprietary solutions claim to better fit specific needs; consequently, there is a hard push to get shares of the potentially huge WSN market.

Today, first generation WSN systems are commercially available, based on either standard or proprietary technologies. End products range from home automation (Control4, Eaton's

HomeHeartBeat, Kalirel's Cyclope), to security (RayMarine's LifeTag), industrial control (Siemens' Apogee), automatic meter reading (Wellspring) and others. The list increases every day. Application designers can choose among various technologies, hardware devices, software, tools, and all-in-one solutions (ST's SN250, TI's CC2430). In 2005, the Zigbee chipset market had already registered 2.5 million unit shipments and generated revenues of $11.3 million [3]. Moreover, an exponential growth is expected for the coming years, as will be further discussed.

The remainder of this article is structured as follows. Section 2 addresses the ambiguous situation of WSN today by investigating what key issues prevent or simply delay its true market ramp-up. Section 3 then provides an insight of the WSN state-of-the-art technology for some of the core hardware and software elements. Section 4 focuses on some of the research fields expected to drive tomorrow's WSN technology. Finally, Section 5 summarizes the key highlights of the paper and provides concluding remarks.

2. GREAT EXPECTATIONS, YET MANY QUESTION MARKS

Today, the WSN market as a whole is rather confused. Despite the already wide offerings in low data rate products, there are many concerns that have prevented a massive adoption by the market up to now. Although only time will tell what is going to happen, it is worth highlighting the most important issues of today. These are addressed below.

2.1 Market uncertainties

Pioneer research work on modern WSNs started back in the 80's under the DARPA umbrella; the Zigbee alliance was formed in October 2002 and the IEEE802.15.4 standard published in October 2003. However, the time between technology concept and real market adoption is always longer than expected. Recent examples include Bluetooth or UMTS, and this will apply as well to WSN in general and to Zigbee in particular. After a few years of strong design activity, mostly by small companies, first generation hardware and software solutions are being shipped, and big silicon manufacturers recently have become very active as well. In this ramp-up phase, home automation has the biggest potential; however, consumer demands for low cost and plug-and-play solutions might delay the true mass-market volume. In the meantime, industrial applications lead the path to market adoption. Once high volumes and technology maturity are achieved, many other application segments will benefit from WSN. Several market analysts report that the overall WSN market could hit the 1 billion dollar target as soon as 2009-2010.

The setup of standard solutions will definitely help in reducing device costs and creating true interoperability. However, up to now, most WSN solutions on the market were built upon proprietary solutions, and it is still unclear whether Zigbee will succeed in all market segments or not. Some of Zigbee's competitors are trying to promote simpler solutions tailored to specific market segments (e.g., Insteon in Home Automation) or building upon other technological bricks (e.g., Zwave from Zensys).

2.2 Spectrum and regulations

Today, most WSN solutions use NB modulations and operate in license-free bands. The 433MHz band is available worldwide (with minor modifications in Japan) with the major restriction that applications such as audio or continuous data transmission are typically not allowed in this band. The 868MHz (in Europe) and 915MHz (in USA) bands are not application restrictive and allow for low cost implementation but suffer from "not being worldwide." It is thus no surprise that most Zigbee vendors today select the 2.4GHz band for worldwide operation. However, the use of several air interfaces operating in different frequency bands potentially limited to specific countries only adds up to the WSN market confusion. Regulation for UWB is under discussion.

Moreover, with a maximum of 80MHz of allocated bandwidth (in the 2.4GHz NB band) to be used simultaneously by multiple

WSNs (e.g., HVAC and alarm system) and shared with other types of devices operating in this band (Bluetooth, Wifi, microwave ovens, etc.), the issue of coexistence is foreseen as one major threat to WSN mass-market adoption.

2.3 Reliability

The Zigbee Alliance slogan is "Wireless Control that Simply Works." Reliability is probably the number one requirement for adoption by the mass-market. Before being low cost, low power, self-organized, or super-smart, wireless sensor networks should simply work. Major progress has already been made during the past years on this aspect, but quite a few challenges might still strongly affect the image of (standard or proprietary) WSN solutions. Transmit power variations upon manufacturing and temperature conditions, sensitivity to Electro Magnetic Interference, or limited battery lifetimes are typical factors which may affect the node itself; however, those can be easily anticipated. More troublesome are interference, shadowing, or multi-path fading effects which affect the wireless link reliability; those are fairly hard to systematically cope with as they highly dependent upon channel conditions. Even more application specific, hence difficult to anticipate, are all the reliability problems related to the wireless network configuration where orphanage procedures, route discoveries, or security checks are needed.

2.4 Ease of use

Some equipment manufacturers may choose to purchase System-on-chip (SoC) solutions, select the right antenna, and dig into board design and complex programming tasks. However, for those application developers who still are not familiar with wireless connectivity, complete packaged modules start to be available on the market. The work effort is then limited to supplying energy (battery), connecting a few sensors, and programming the application software, potentially on an external microcontroller.

Unfortunately, the end-user experience might be affected as well by the wireless connectivity. It is nice to get rid of the cables as long as all this remains transparent to the user. Nodes should be easy to deploy, wireless links self-configured, and networks self-healing. Likewise, upgrades and remote maintenance should be enabled by the appropriate interfaces and tools.

2.5 Application requirements

How could pressure monitoring of pipelines in the desert and tracking of precious medical equipment in a hospital share the same requirements? The answer is probably that they cannot. The WSN design space is very large and linked with constraints from the device (size, cost, power, etc.), the air interface (data rate, range, robustness, etc.), the network configuration (topology, scalability, ad-hoc vs infrastructure, etc.), and even the application features (localization, QoS, mobility, latency, etc.). In the absence of a single killer application, next generation WSN solutions will have to offer the appropriate tradeoff between flexibility and efficiency. It is expected that current solutions will evolve incrementally to optimize existing applications (sensitivity, security, etc.) while disruptive changes will be needed to enable new ones (energy scavenging, localization, etc.).

3. KEY COMPONENTS SURVEY & STATE-OF-THE-ART

As illustrated on the left side of Fig. 2, any WSN mote is a complex device that embeds several dedicated hardware and software components, irrespective of the application. Several motes are then combined to build either simple (e.g., star, tree) or very complex ad-hoc mesh networks, which can potentially be bridged to other types of information networks such as internet (right side of Fig. 2).

The purpose of the following section is to briefly describe the most relevant components and to comment on the status of their key features and performance so as to get a better picture of where WSN technology stands today.

3.1 Sensors & actuators

A sensor is a physical device that probes physical, biological, or chemical properties of its environment and converts these

properties into an electrical signal. Sensors for temperature, light, oxygen, distance, blood pressure, moisture, and torque are some of the many examples. An actuator typically accepts an electrical signal and converts it into a physical action to act upon the environment. Sensors and actuators belong to the broader family of transducers.

Classical transducers such as temperature or pressure sensors are available as off-the-shelf components and can be easily integrated at the board or package level. More complex ones like CMOS image sensors, inertial sensors, or micro-fluidic actuators have recently emerged, thanks to technological advances. Those "smart" sensors typically require dedicated logic for calibration, signal processing, or analog-to-digital conversion and sometimes include a micro-controller. The LM135 precision thermal sensor, TDA0161 proximity detector, and LIS3LV02DQ three-axis digital accelerometer from STMicroelectronics are typical examples of sensor technology [4]. As recently demonstrated with the In-Check platform (a biochip for amplifying tiny DNA samples), it is expected that ongoing efforts on CMOS and Micro Electro-Mechanical System (MEMS) technologies will further enlarge transducer potential in the near future.

Transducer manufacturers serve a vast variety of applications using diverse physical interfaces and control protocols. The IEEE1451 is pursuing an effort to develop a family of open, network-independent communication interfaces for connecting transducers to microprocessors, instrumentation systems, and control/field networks [5]. It particularly defines the concept of Transducer Electronic Data Sheets (TEDS), a memory device attached to the transducer, which stores transducer identification, calibration, correction data, measurement range, and manufacturer-related information, among others. This allows connecting virtually any Smart Transducer Interface Module to any Network Capable Application Processor in a "Plug-and-Play" manner.

3.2 Transceivers

Each node of a WSN system comprises a transceiver unit, which is in charge of the wireless communication with peers. This dedicated hardware usually supports the complete PHY layer, as well as some time-critical operations of the MAC layer which would not be efficiently performed in software. Regarding the MAC, mechanisms such as CSMA fit WSN constraints well because neither central coordination nor synchronization between nodes is required; nevertheless, time synchronization is a hot topic, for it allows the effective implementation of power duty cycle solutions. Conversely, applications with very low traffic may actually rely on a simpler Aloha medium access mechanism.

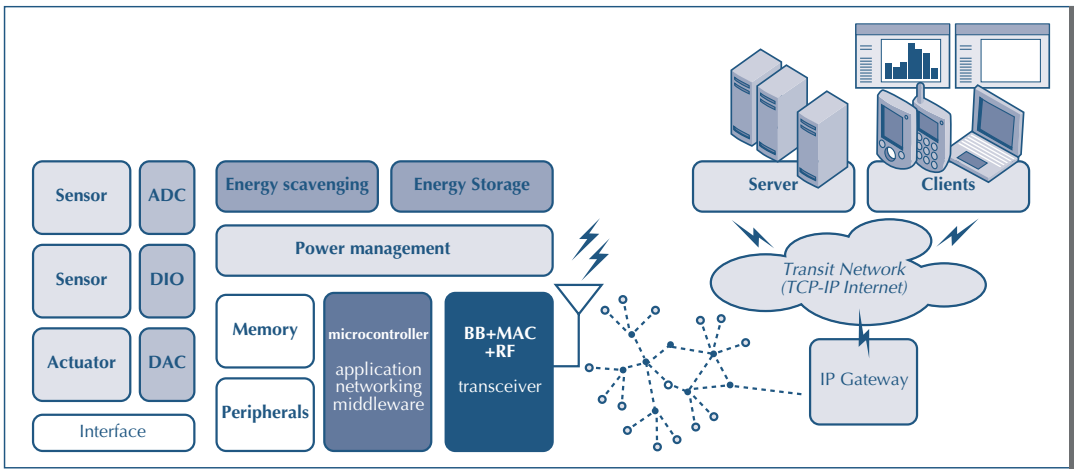

FIGURE 2: WSN MOTE (LEFT) AND WSN COMPLETE INFORMATION PROCESSING SOLUTION (RIGHT).

PHY layers for WSN systems are typically based on Radio Frequency (RF) technologies. Although the preference for unlicensed bands has already been discussed for NB systems, an additional advantage of such frequencies is that the corresponding technology is mature, hence lower in cost. Moreover, transmitted power and typical receiver sensitivity, around 1mW and -100dBm respectively, fit well for indoor operation. Typical NB modulations are Phase Shift Keying enhanced with Direct Sequence Spread Spectrum to increase link reliability and ease packet detection and synchronization. Data rates over-the-air range up to tens of kbit/s for NB systems.

The communication device usually deserves careful design because its power consumption is often far from negligible. Currently, IEEE802.15.4-compliant transceivers at 2.4GHz typically require 20-25mA in both transmit (at 0dBm) and receive mode, at 1.8V. Future generations shall get down to figures of 10-15mA [6]. Initial UWB devices are expected to be more demanding, but this could be compensated for by reduced duty cycles due to the higher peak rates over the air. In the longer run, advanced research targets even lower figures. For NB alternatives, a 2.4GHz transceiver shall require less than 1mA in both modes [7]. Research on UWB foresees down to 0.4mW for the transmitter and 0.6mW for the receiver at 1Mbit/s [8]. Hence, probability is high for future WSN transceivers to benefit from a significant decrease in power consumption as well as size and cost.

3.3 Networking

Networking is a huge research topic in WSN; thus, it would be quite ambitious to provide an exhaustive state-of-the-art discussion in a few lines. Nevertheless, because it encompasses key features of any WSN solution, it at least deserves a quick overview.

In its simplest form, a WSN may rely on a central "base-station" collecting sensory data from all other devices in a synchronized and (infra)structured way. Infrastructure-based systems typically enable very efficient usage of the available bandwidth, though with high complexity or limited scalability (number of nodes and covered area). A more adequate approach to low cost and low power wireless networks consists of implementing an infrastructure-less network, also known as ad-hoc network, where heterogeneous devices communicate in a self-organized manner.

Mesh topologies, self-organization, and multi-hopping are commonly assumed in WSN. Compared to wired networks, wireless networks benefit from the broadcast nature of wireless communications but suffer from potentially severe and irregular channel conditions. The Zigbee consortium has defined what could prove to be the first large scale implementation of WSN, although alternatives like Z-Wave or Insteon products exist. For all those technologies, the networking layer is responsible for many functions such as network creation procedures, address assignment, packet routing, security, application management, etc. Advanced features such as self-healing capability, energy aware routing, or mobility support are being explored by the research community and will be progressively introduced in commercial products.

The networking functionality is commonly implemented all in software and shall run on low-end microcontrollers. The footprint of a Zigbee stack is in the range of 64kbytes of program memory and requires a few kbytes of RAM. Much simpler network protocols may be implemented for those applications which require only basic features (fixed addressing, star topology, etc.), albeit at the cost of reduced flexibility and interoperability.

3.4 Microcontroller

As discussed in the previous sections, increased networking capabilities, data logging features, or self-healing capabilities of WSN motivate flexible solutions. A microcontroller (MCU) is a self-sufficient and cost-effective "computer-on-a-chip" used to control the functionality and data flow of (embedded) electronic devices and to store/process data. On top of the processing core - a Central Processing Unit (CPU) sometimes enhanced with Digital Signal Processing (DSP) hardware - today's MCU cells include customized memory sub-systems, non-volatile memory

(embedded Flash), and interfaces such as ADC, UART, SPI, counters, and timers. In this way, it can interact with embedded sensors and communication devices to compose a fully-integrated sensor node. There are many types of microcontrollers with respect to the micro-architecture (CISC versus RISC), instruction set bit width (i.e., Atmel's ATmega165P 8-bit [9], TI's MSP430 16-bit [10], or ARM's Cortex-M3 32-bit [11]), different low-power operating modes, and peripheral diversity or energy-efficiency (Watt/MIPS) characteristics. Commercially available state-of-the-art MCUs report power consumption figures on the order of 250µA/MIPS in active mode, 1µA in real-time clock (or idle) mode, and 0.1µA during RAM retention (standby mode). It should be noted that the exact figures depend on the actual application, number, and type of active peripherals and operating temperature. The ideal choice of the MCU is the one that matches its performance level with application needs.

3.5 Energy management

In WSN, sophisticated energy management techniques are essential since WSN devices (running on a single battery) are expected to last for several years. Energy-aware *computing* includes ultra-low power microcontrollers (see Section 3.4) and Dynamic Power Management (DPM, [12]) hardware to provide requested performance and service levels while minimizing power consumption (by turning off inactive sensors/peripherals and parts of the IC). Among the key characteristics of energy-aware *software* (e.g., TinyOS) are power-aware scheduling, exploitation of concurrency, and the support for (high-level) programming routines to efficiently control the low power operating modes of MCUs. Finally, power management techniques of the *radio* transceiver include wake-up mechanisms (e.g., based on carrier sensing) in conjunction with advanced MAC protocols (e.g., WiseMAC [13]).

Voltage regulators and DC/DC converters are key building blocks in the implementation of dynamic power management techniques. While the voltage regulator provides a constant DC voltage regardless of changes in the load current, the DC/DC converters match the voltage level of the variable power supply to the individual voltage levels required by the different sub-circuits (so-called "voltage islands"). Whereas today's commercial WSN products rely solely on battery power, energy-harvesting techniques coupled with secondary (rechargeable) batteries may soon replace conventional solutions in the WSN market [14].

As a case study, a WSN mote - composed of a MCU, embedded memory, sensor(s), and a 2.4GHz IEEE802.15.4 radio transceiver - is reported in [15] and is compatible with energy harvesting techniques such as piezoelectric materials or solar cells. For very low-duty cycle sensing applications (≤ 10Hz), the power consumption was reduced from several tens of mWatts below one mWatt by implementing DPM, where both the radio and sensors sleep between the "sensor's sampling periods." Note that, in this particular hardware configuration, the radio transceiver is the biggest power drainer with ~45mWatts in a continuous radio link setup. The power consumption was greatly reduced (to ~5mWatts in continuous mode) by processing and/or logging data, rather than streaming it over the air. Thus, there is an important "local computation/storage" versus "radio transmission" architectural trade-off to be optimized as a function of the application profile.

3.6 Others

On top of the aforementioned components, a WSN mote includes many others which serve the application purpose; interfaces of various kinds (Analog-Digital Converter, Universal Synchronous-Asynchronous Receiver/Transmitter, Serial Peripheral Interface, etc.), general purpose Inputs/Outputs, support for external crystal oscillator or integrated RC oscillator, peripherals (timers, encryption hardware acceleration engine, etc.), middleware (operating system, drivers, etc.), and memory storage for program and data are typical.

Then, beyond the single mote, a WSN complete solution is a very complex integrated system where sensory data collected throughout a network of motes can be made accessible to the

"end-user" by different means. In its simplest form, data might be gathered in a so-called sink mote device. Bridges to other kinds of information networks are, however, very common. Gateways can, for instance, be developed to create legacy industrial networks, Ethernet TCP-IP, or powerline backbones. On the other side of this transit network, sensory data management is handled by traditional database client-server infrastructures.

4. FUTURE WSN TECHNOLOGIES

Addressing the state-of-the-art of current technologies allows picturing where WSN systems stand today, but a closer look at what may build next generations deserves some discussion as well. Many trends are being investigated by STMicroelectronics, some of which are related to communication aspects (transceivers and networking) and are presented in the following section. Specifically, improvements of current narrowband systems, UWB technology including a dedicated prototype, and position-enabled sensor networks are addressed, followed by a glimpse at even longer term disruptive technologies.

4.1 Improved Narrowband systems

As previously discussed, due to gaps in technology maturity, first generation WSN systems shall remain NB-based for some time. While disruptive approaches like UWB could take the lead some day, it is still worth investigating NB evolutions as well because these are expected to be implemented and become profitable in a shorter time.

Even if there are many different NB alternatives, some of their weaknesses are common independently from their specific characteristics. Among such key limitations, co-existence in unlicensed ISM bands has already been pointed out. Another well-known weakness of NB systems is their sensitivity to multipath fading, which is particularly critical in indoor environments. Given their typical carrier frequencies and bandwidths, NB receivers suffer flat fading, which can result in more than 30dB of signal attenuation [16]. The statistics of such fades mostly depend on mobility, and static scenarios might appear as the worst cases, since links suffering deep fades would then remain weak, if not broken, forever. To overcome such effects, a solution well adapted to ad-hoc networks consists of using multi-hop capabilities by clever and dynamic routing, depending on the reliability of each link of the network; obviously, such an alternative is expensive but nevertheless appealing since many applications may require similar features for other purposes as well (e.g., routing).

Another way to deal with multipath fading consists of exploiting space diversity by means of multiple antennas at the transmitter and/or the receiver side. Let's for instance assume 2 antennas per receiver. Provided they are at a distance of at least a fraction of the wavelength apart, the correlation between both received signals is usually low enough to justify some kind of selection or combination [17]. Depending on the application and the environment, various alternatives allow then a tradeoff of performance for complexity. The simplest one consists of a straightforward combination of the antenna signals, which is rarely optimal from a performance perspective, but does not require more processing or even energy than used by a single antenna receiver. A more complex scheme consists of selection diversity, which means choosing the "best" antenna at any time. In a general WSN context, such a selection is not trivial because the typical information exchange duration is too short to allow reliable real-time evaluation of each alternative. However, such a scheme may increase link reliability in static scenarios where a dedicated setup phase is used for measurements, while selection diversity is used afterwards with the assumption that the channel conditions do not change in-between. A more efficient use of space diversity consists of combining both signals but in a weighted way so as to maximize the signal-to-noise radio (SNR). This usually requires demodulating each signal independently to get the weighting coefficients and only then to combine them accordingly. The price to pay in complexity and power consumption is high, given that two receivers are necessary down to the combination process. Nevertheless, sub-optimal alternatives that lower complexity at the price of some

performance decrease do exist; for instance, [18] suggests using a single receiver to demodulate both signals while allowing their sub-optimal separation by artificially delaying one of them before combination.

4.2 UWB technology

A plurality of UltraWideBand (UWB) systems is possible, each with relative advantages and disadvantages. While an OFDM-based modulation is adopted for high data rate applications [19], an impulse-based radio signaling scheme, as illustrated in Fig. 3 (right side), is adopted for Low Data Rate (LDR) WSN systems. More specifically, the MAC and PHY layers are defined within the (on-going) IEEE802.15.4a standardization framework [20], which proposes 3 device options operating respectively in the DC-966MHz, 2.4GHz ISM, and 3.1-10.6GHz frequency ranges. The UWB preamble is composed of single (ternary-) coded pulses, whereas the payload data is modulated using a combination of Burst Position Modulation (BPM) and Binary Phase Shift Keying (BPSK), with each symbol being composed of an active burst of UWB pulses.

in very good interference rejection and thus is the key enabler for network co-existence. Another key feature is the fine time resolution (< 1 nanosecond) proper to UWB signaling, resulting in accurate Time-of-Flight (ToF) estimations which allow for accurate ranging (or distance measurement) among peers.

Despite these promising features, the design of energy-efficient and low-cost UWB transceivers is still a challenging task and is addressed in the following.

4.2.1 UWB analog front-end

A standard-compliant transmitter should generate pulses of 500MHz bandwidth minimum and up to 1GHz (optional). The possible center frequencies of the pulse are clustered in a low band group (3494.4MHz, 3993.6MHz, or 4492.8MHz) and a high band group (ranging from 6489.6MHz to 9484.8MHz). The payload burst structure requires that the pulse generator reaches a pulse repetition frequency of 499.2MHz together with the capability of flipping the pulses' polarity at the same rate. An accurate control of the pulse envelope will also help

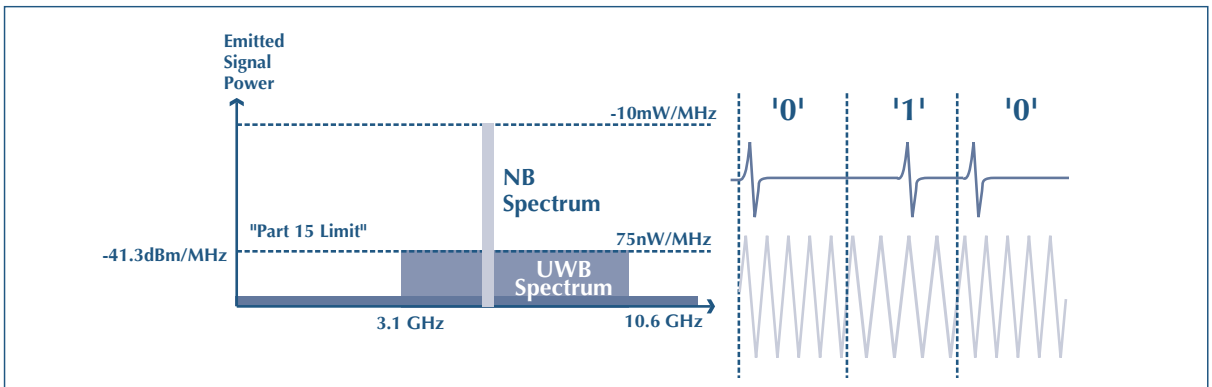

FIGURE 3: SPECTRUM CHARACTERISTICS (LEFT) AND TEMPORAL SHAPE (RIGHT) OF UWB PMM (GREEN) AND NB FSK (BLUE) MODULATED SIGNALS.

Compared to NB systems, the UWB does not suffer from multi-path fading effects since the individual time-delayed multi-path replicas of a short pulse are resolvable at the receiver. Moreover, the inherent low duty cycle of pulse-based signaling schemes coupled with pseudo-random time-hopping techniques results relax necessary filtering to fit inside the power emission masks imposed by the regulation.

At the receiver, the architecture choice is not only driven by the targeted data rate but also by the requirement for time-of-flight

estimation. A coherent receiver offers accurate synchronization, which helps in both directions but imposes tighter requirements on the frequency synthesis and digital baseband operations. The use of a non-coherent receiver relaxes phase noise specifications and overall power consumption at the expense of a performance loss and has been the selected architecture for a first UWB LDR prototype by STMicroelectronics.

The need for spatial resolution will lead to relatively high sampling rates that the low data rate does not require. A quick look to the literature shows that ADC technology has progressed a lot during the last decade, and state-of-the-art efficiencies on the order of tens of picoJoules per sample have been demonstrated over a few bits [21]. Another means to reducing the power consumption linked to the ADC is to take advantage of the wide band characteristics of the UWB signals, which can be strongly clipped with very limited impact on their correlation properties [22]. It has also been shown that in low SNR and low signal-to-interferer ratios, the resolution of the ADC has low impact on performances [23]. This method offers the additional advantage of not having to use a variable gain amplifier in the receive chain.

The objective of the non-coherent receiver is to measure the power of an incoming signal, usually averaging the measure over a sufficient duration so as to detect variations of the mean signal level compared to its variance. So-called correlation detectors can be implemented through an analog correlator using a multiplier to combine the incoming signal and its replica (or signals coming from two antennas).
An alternative introduced by Faran and Hills [24] based on the work of Van Vleck is to use a polarity coincidence correlator which performs the same correlation after strongly clipping the signals. Cheng [25] showed that the loss due to clipping with respect to the analog multiplier is less than 4dB under any condition of signal power, noise power, and shape of auto-covariance. This tradeoff brings a much lower complexity through the use of only 1 bit ADC.

4.2.2 UWB digital processing

The digital baseband (BB) receiver functions within an IEEE802.15.4a-compliant (non-coherent) UWB-LDR system consist of synchronization, channel and drift estimation, Synchronization Frame Delimiter (SFD) detection and ranging, as well as data demodulation and decoding. The task dependencies are shown in Fig. 4 and infer sequential processing of most of the tasks to allow energy-efficient (clock) gating of inactive processing modules. Synchronization is the most processing-intensive task due to the low-duty cycle of pulse-based preamble waveforms - defined by the pulse repetition period (PRP) - and the high Nyquist rate of UWB signals. The “two-dimensional” search space is defined by the number of clock cycles per PRP (time-alignment) and the code length of one synchronization symbol (phase-alignment). To reduce this search space, the analog front end implements a form of dynamic sub-sampling motivated by the observation that UWB communications in realistic propagation environments demonstrate dense diffuse multipath effects with rms delay spreads of tens of nanoseconds. Our sub-sampling scheme exploits this “spreading” of the pulse energy by accumulating consecutive (Nyquist rate) samples to form one combined detection statistic, which is then transferred to the synchronization hardware composed of a matched filter and a threshold-based periodicity detection module to guarantee an optimal detection performance. It is up to the BB to control the sub-sampling period, which is either set statically (based on analytical channel models) or updated dynamically, based on ongoing channel estimations and a-priori information. On top of reducing the effective cycles per PRP, sub-sampling reduces the acquisition time for a fixed number of hardware resources and associated clock frequency requirements (design exploration parameter).

After initial time and phase alignment, the channel delay profile is estimated by comparing the time-aligned matched filter statistics with an energy threshold to decide whether or not a given sub-sampling phase contains sufficient signal energy to be considered for further processing. To improve the SNR, the

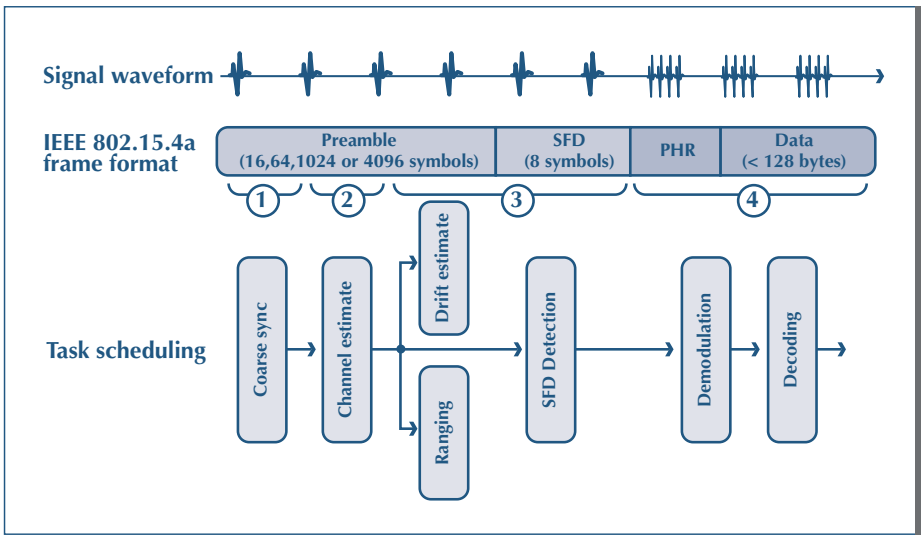

FIGURE 4: FUNCTIONALITY OF AN IEEE 802.15.4A-COMPLIANT DIGITAL BASEBAND RECEIVER; IN THE MANDATORY MODE, EACH SYNCHRONIZATION SYMBOL IS COMPOSED OF A TERNARY CODE OF LENGTH 31.

resulting binary channel mask is fed to the Trellis-based drift estimation [26], SFD detection composed of thresholding, and matched filtering, ranging based on threshold-based first-path detection and demodulation implementing a rake-like energy detection scheme adapted to a 2-BPM modulation.

4.3 Position-enabled sensor networks

Pervasive, ubiquitous, invisible, adaptive, or embedded are among the many adjectives commonly used to describe the Ambient Intelligence concept linked with WSN. Indeed, the ability to create seamless communications between hundreds of intelligent tiny devices anywhere and anytime is what enables a whole class of new applications. Networking is crucial in WSN to perform many actions such as remotely controlling the devices, querying one particular sensor, or adding some level of coordination and self-organization among the nodes. Unfortunately, networking in general and more particularly wireless networking brings up a series of issues. An ideal network should be self-configured and should guarantee packet delivery and quality of service, resist malicious eavesdropping, offer self-healing capability, etc. Refer to [27] for an outlook of those network related topics.

In addition to the transducer data itself, it can be very useful to transmit the context (location, time, etc.) and the destination (user profile, history, etc.) as well. Knowing the location or position of a particular node in a widespread network is particularly interesting from several points of view. At the radio level, this information could be useful to dynamically adapt the emission power so as to save energy or to reduce the risk of packet collisions. At the networking level, this information is useful to, for example, optimize broadcast or routing procedures. At the security level, the position of devices can be used to detect malicious device intrusion or to limit the risk of packet interception. However, the biggest benefit of position information goes to applications. It can greatly simplify the maintenance tasks of existing WSN applications; it can enhance already existing applications with location-based query mechanisms, as an example; it can finally contribute to the creation of brand new applications such as structural health monitoring or indoor navigation systems, where accurate position information is mandatory.

Military radars and GPS are two typical examples of outdoor positioning techniques that have existed for many years. The challenge of positioning in WSN is the need for a low cost, low power, and scalable indoor positioning system with a reasonable accuracy. One important part of a positioning

system is its capability to measure the distance between two devices, called ranging. Several techniques exist that rely on a given signal (infrared, electromagnetic, ultrasound, etc.) and a measurement mean (angle of arrival, signal propagation loss, time of flight, etc.). Since RF devices are already in use for data communications in WSN, radiolocation-based techniques are commonly used and offer the advantage of having good obstacle penetration. Received Signal Strength techniques are preferred for narrowband radio systems (such as the IEEE802.15.4) but suffer from imperfect radio propagation loss models, manufacturing variation, battery levels affecting the emission level, and wireless channel effects varying in space and time. Time of Arrival techniques offer a better accuracy for UWB radios, thanks to the excellent signal resolution of nanosecond long impulses. The effect of multipath propagation and potential clock drift, however, usually degrades the achievable accuracy, and a tradeoff must be found between the cost/power of the solution and the ranging accuracy.

The 2-D position of a device can be easily inferred from the known position of three reference (or anchor) devices by performing one-to-one ranging measurements and using simple triangulation methods. However, in widespread WSN, it is usually not possible to have all nodes in the vicinity of at least three (or four in 3-D scenarios) anchor nodes, due to limited radio range and the relatively higher cost of anchor nodes. Distributed techniques to compute an estimate of each (so called) blind node position from a small subset (typically 5 to 20%) of anchors have been the subject of a lot of research in the past years [28]. The simplest techniques simply combine approximations of the distance to anchor nodes over a multihop link with triangulation-like methods. More advanced algorithms exist that try to iteratively minimize a global error function that models the discrepancies between latest distance and position estimates.

4.4 Disruptive technologies

Beyond the progressive enhancements of current technologies, some more revolutionary ones are worth mentioning in the context of WSN because they may open the door to even wider applications and contexts of use. Even if the list is not exhaustive, this section highlights some research fields that might sooner or later impact the design of WSN devices, especially those regarding power consumption and integration aspects.

4.4.1 Semiconductor processes

The market penetration of sensor networks is closely linked to robustness of the wireless link while form-factor and cost are strong but secondary constraints. Out-of-band noise and interference rejection are key to robustness and cannot always be realized in classical CMOS technology, due to limited quality coefficient of passives. At the expense of higher costs, the Silicon-on-Insulator technology can improve filter performances due to excellent isolation to the substrate through higher quality of buried oxide. If not sufficient, alternatives exist that requires module assembly such as on-wafer mounted bulk acoustic wave filters or multi-layer thin film technology and system-in-package that both provide the capability of integrating high quality passives and antennas. For some types of sensors such as accelerometers or gyroscopes of digital micro mirrors, surface-micro-machining is an alternative that is compatible with classical CMOS process. Use of this technology can also be extended to energy scavenging.

4.4.2 Energy scavenging

Thanks to technology and design advances, the power needed by a sensor to operate is decreasing to levels as low as 100uW on average. The use of a 1cm3 Li-ion battery would enable a sensor lifetime of 3 years maximum. Alternative technologies make it possible within the same volume to scavenge or harvest the same amount of energy from sources in the environment of the device, that is, without limiting the lifetime of the sensor. The main sources that have been envisioned up to now are kinetic, thermal, and electromagnetic. The first one is mainly the transformation of vibrations to electricity through piezoelectric, electrostatic, or magnetic induction effects and can provide up to 300uW/cm3. The second one is the transformation of a heat

flux into electricity and can provide up to 40uW/cm3. The third one retrieves energy from light - at levels up to 15000uW/cm3 outdoor and down to 10uW/cm3 indoor - or from RF radiation up to tens of milliWatts in the case of a strong source such as an RFID tag reader. The best candidates for integration into CMOS technology are the electrostatic through MEMs cantilevers and heat generators.

4.4.3 Wake-up radio

Given the low duty cycles of WSN applications, the ability to partly switch off the node during inactive periods can provide important power consumption savings. Existing MAC strategies combine communication scheduling mechanism and carrier sense. In the case of UWB, the signal is subject to low signal-to-interferer ratios that prevent the use of a simple peak detector. Examples of passive detection of UWB pulse trains using specific IDT techniques have been demonstrated in the literature but suffer from the very high insertion losses of the Surface Acoustic Wave devices (~60dB). An active detection has also been proposed, taking advantage of the very high gain of super-regenerative architecture [29]. Given the power consumption, the budget of the wake-up system should be limited to a fraction of that of the transceiver, a very attractive concept requiring strong and innovative efforts.

4.4.4 Low power design

Besides the more "classical techniques" to reduce active and standby power - including low-power memory subsystems, individually enabled peripherals, pipelines, clock and power gating, dynamic voltage scaling, or multi-threshold standard CMOS cell libraries to trade-off speed versus leakage - the authors would like to point out two promising research activities that will further improve the SoC's energy-efficiency referred to as asynchronous and subthreshold design. While the first technique reduces the power consumption by omitting clocks for sequencing, the latter one operates the circuit in the transistors' subthreshold region. In [30], an 80% improvement in energy-efficiency was reported for an asynchronous 8-bit processor implementing the Atmel AVR instruction set architecture in STMicroelectronics' 0.13-μm CMOS process. In [31], an FFT processor design in a 0.18-μm CMOS process which operates down to 0.18 Volts at the expense of computing bandwidth has been realized.

5. CONCLUSIONS

This paper has presented an extended survey and status of Wireless Sensor Networks. The purpose was both to describe the state-of-the-art of current technologies and to address future trends and expectations as well. Application issues have been left aside, allowing us to focus on generic subsystems and typical features of WSN systems like wireless communication, networking aspects, or energy management.

The evolution of the WSN context from research concept to mass market products has first been addressed to better understand which are today's open issues other than technological limitations. Specifically, market uncertainties, spectrum and regulations, reliability, setup/configuration complexity, and a lack of a killer application have been discussed. This pointed out that technology is only one among many factors impacting the possible future of WSN systems.

As a second step, the key features of such systems have been addressed and analyzed one by one to get a picture of where WSN technology stands today. On the hardware level, the state-of-the-art subsystems such as transceivers and microcontrollers have been described, and relevant metrics such as power consumption have been assessed. On a more global level, energy management architectures and techniques have been introduced. On an even higher level, networking aspects and topologies have been addressed as well. This overview has highlighted the fact that a clever design of WSN nodes requires care of the global system as much as that of each independent function.

Finally, research fields and promising technologies that could shape the future of WSN have been investigated. Possible

enhancements of current narrowband systems, for instance, by means of space diversity, were addressed. UWB technology has been introduced, and a dedicated prototype has been described. Innovative networking features such as positioning have also been investigated. Possibilities of disruptive technologies such as advanced processes, energy scavenging, or wake-up radio have also been addressed.

In conclusion, it has been shown that, even if the WSN market suffers many uncertainties, today all building blocks and several integrated solutions are already available to fulfill requirements of the first generations of applications. Research continues to be very active in investigating revolutionary approaches to answer to higher needs while providing new or better features. In other words, the potential of improvement of WSN technologies is still huge, and concepts such as energy scavenging may even open the way to application fields that have not yet been thought of, fields which would expand WSN deployment even further.

REFERENCES

[1] IEEE802.15.4 specifications, October 1st 2003, http://www.ieee802.org/15/pub/TG4.html

[2] Zigbee specifications, version 1.0, December 14th 2004, http://www.zigbee.org

[3] Frost & Sullivan research, "WORLD UWB & ZIGBEE CHIPSET MARKET", May 16th 2006, http://www.frost.com

[4] http://www.st.com/stonline/products/families/sensors/sensors.htm

[5] IEEE1451 specifications, http://ieee1451.nist.gov

[6] W. Kluge *et al*., "A FULLY INTEGRATED 2.4GHZ IEEE 802.15.4 COMPLIANT TRANSCEIVER FOR ZIGBEE APPLICATIONS", IEEE Int. Solid-State Circuit Conference (ISSCC), pp. 372-373, 2006.

[7] B. W. Cook *et al*., "AN ULTRA-LOW POWER 2.4GHZ RF TRANSCEIVER FOR WIRELESS SENSOR NETWORKS IN 0.13UM CMOS WITH 400MV SUPPLY AND AN INTEGRATED PASSIVE RX FRONT-END", IEEE ISSCC, pp. 370-371, 2006.

[8] I. D. O'Donnell, "A BASEBAND, IMPULSE ULTRA-WIDEBAND TRANSCEIVER FRONT-END FOR LOW POWER APPLICATIONS", PhD Thesis, U.C. Berkeley, May 8th 2006.

[9] A. M. Holberg and A. Saetre, "INNOVATIVE TECHNIQUES FOR EXTREMELY LOW POWER CONSUMPTION WITH 8-BIT MICROCONTROLLERS", ATMEL White Paper on the AVR-ATmega165P, February 2006.

[10] F. Forster, "ULTRA-LOW POWER MICROCONTROLLERS FOR EMBEDDED SECURITY APPLICATIONS", TI White Paper on the MSP430F2xx, February 2006.

[11] P. Clarke, "STMICROELECTRONICS LICENSES ARM'S SANDCAT CORE", EE Times, July 2005.

[12] L. Benini, A. Bogliolo and G. De Micheli, "A SURVEY OF DESIGN TECHNIQUES FOR SYSTEM-LEVEL DYNAMIC POWER MANAGEMENT", IEEE Trans. on Very Large Scale Integration (VLSI) Systems, vol. 8, no. 3, pp. 299-316, 2000.

[13] A. El-Hoiydi and J.-D. Decotignie, "WISEMAC: AN ULTRA LOW POWER MAC PROTOCOL FOR THE DOWNLINK OF INFRASTRUCTURE WIRELESS SENSOR NETWORKS", 9th Int. Symposium on Computers and Communications (ISCC), vol. 1, pp. 244-251, July 2004.

[14] B. H. Calhoun *et al*., "DESIGN CONSIDERATIONS FOR ULTRA-LOW ENERGY WIRELESS MICROSENSOR NODES", IEEE Trans. on Computers, vol. 54, no.6, pp. 727-739, June 2005.

[15] S. W. Arms et al., "Power Management for Energy Harvesting Wireless Sensors", SPIE Int. Symposium on Smart Structures & Smart Materials, pp. 1-9, March 2005.

[16] T. S. Rappaport, "Wireless Communications, Principles & Practice", Prentice Hall PTR, 1996.

[17] G. E. Corazza et al., "A Characterization of Indoor Space and Frequency Diversity by Ray-Tracing Modeling", IEEE Journal on Sel. Areas in Communications, vol. 14, no. 3, pp. 411-419, April 1996.

[18] T. Weiguo and Y. Chenyang, "A Delay Diversity System Based on IEEE 802.15.4 With RAKE-like Receiving", 7th Int. Conference on Signal Processing (ISCP), vol. 2, pp. 1686-1690, 2004.

[19] MB-OFDM PHY 1.1 Specification, http://wimedia.org

[20] IEEE Std 802.15.4a, "Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs)", Draft P802.15.41, April 2006.

[21] G. Van der Plas, S. Decoutere and S. Donnay, "A 0.16pJ/conversion-step 2.5mW 1.25Gs/s 4b ADC in a 90nm Digital CMOS process", ISSC conference, February 2006.

[22] J.H. Van Vleck and D. Middleton, "The spectrum of clipped noise", IEEE proceedings, vol. 54, pp2-19, January 1966.

[23] I.D. O'Donnell and R.W. Brodersen, "An ultra-wide band transceiver architecture for low power, low rate, wireless systems", IEEE Trans. On Vehicular Technologies, vol. 54, issue 5, pp.1609-1622, September 2005.

[24] J. Faran and R. Hills, "Correlators for signal reception", Acoustic Research Labs, Harvard University, Cambridge, MA, Tech. Memo No 27, 1952.

[25] M.C. Cheng, "The clipping loss in correlation detectors for arbitrary input signal-to-noise ratios", IEEE Trans. on Information Theory, vol. 14, no 3, May 1968.

[26] A. Wellig and Y. Qiu, "Trellis-based maximum-likelihood crystal drift estimator for ranging applications in UWB-LDR", IEEE Int. Conference on UltraWideBand (ICUWB), pp.539-544, September 2006.

[27] D. Blasi *et al.*, "Ad-hoc Wireless Sensors Networks: Challenges and Issues impacting on Networking", ST Journal of System Research, Special Issue on Wireless Sensor Networks, 2006.

[28] Langendoen and Reijers, "Distributed localization in wireless sensor networks; a quantitative comparison", Int. Journal of Computer and Telecommunications Networking, Elsevier Computer Networks, vol. 43, issue 4, pp. 499 – 518, 2003.

[29] M. Pelissier, "A new pulse detector based on super-regeneration for UWB low power applications", IEEE ICUWB conference, pp.639-644, 2006.

[30] L. Necchi *et al.*, "An ultra-low energy asynchronous processor for Wireless Sensor Networks", 12th IEEE Int. Symposium on Asynchronous Circuits and Systems (ASYNC), pp. 78-85, 2006.

[31] A. Wang and A. Chandrakasan, "A 180-mV Subthreshold FFT Processor Using a Minimum Energy Design Methodology", IEEE Journal of Solid-State Circuits, vol. 40, no. 1, pp. 310-319, January 2005.

■ Contact: st.journal@st.com ■

AD HOC WIRELESS SENSOR NETWORKING: CHALLENGES AND ISSUES

Danilo Blasi[1], Vincenzo Cacace[1], Luca Casone[1], Marco Rizzello[1], Salvatore Rotolo[1]
Luciano Bononi[2]

(1) STMicroelectronics
(2) Department of Computer Science, University of Bologna, Italy

Wireless Sensor Networks (WSN) constitute an emerging application field of microelectronics that promises wide support of the interaction between people and their surroundings.
Because of the extremely variable nature of this interaction, the topic of WSN is still in its early stages of development and is giving rise to several challenges, from the design of a smart device that allows easy and reliable access to the environmental characteristics, to the creation of a wireless network of devices that cooperate to perform complex tasks. In this field, the cooperation among nodes is the most peculiar aspect reflecting directly on the network operation.
This paper gives an overview of the topic from the Ad Hoc networking concept perspective. Although the seeds of these concepts can be found in some already existing implementations (e.g., ZigBee), we believe many further refinements could be introduced to fully exploit the potential of the widely investigated Ad Hoc approach and the strength of the mutual interaction of networked elements.

1. INTRODUCTION

It is easy to recognize the main changes of our daily life caused in the past decade by the effect of the 'marriage' of computation and communication technologies.

The synergy resulting from such a combination is producing a powerful technological push: the Wireless Sensor Networks (WSN) is emerging as the new revolution that will make a reality the vision of people like Gordon Bell and Mark Weiser [1], who envisaged the ability of microelectronics to form new computer classes and the capability of networking to support the Pervasive – or, Ubiquitous – Computing, i.e., the idea of integrating computation and communication into the environment, allowing the seamless interaction of computing entities with people in a 'natural' and 'automatic' way.

The potential of the WSN concept simply lies in the three words composing the acronym itself: '*Wireless*' puts the focus

on the freedom that the elimination of wires gives in terms of mobility support and ease of system deployment; '*Sensor*' reflects the capability of micro-/nano-technology to provide the means to perceive and interact – in a wide sense – with the world; '*Networks*' gives emphasis to the possibility of building systems whose functional capabilities are given by a plurality of communicating devices, possibly distributed over large areas.

In this paper, we concentrate directly on the keyword 'Network,' presenting a general overview of the topic and identifying the main challenges that still require solutions for the effective introduction of some real innovation. Given the WSNs' peculiarities, which make them quite different from traditional (wired and wireless) networks, real, effective solutions may only be designed by exploiting the *cooperation* among many elements, i.e., the potential resources of each single network's member can be aggregated and organized to implement multifaceted features and to perform complex operations. In the 'network' perspective, this design approach leads to the 'Ad Hoc' networking approach [2].

Ad Hoc networks are commonly defined as a kind of general, infrastructure-less, cooperation-based, opportunistic network, possibly customized for specific scenarios and applications, as the Latin expression 'Ad Hoc' indicates something which is tailored to a given matter. This networking approach has to face frequent and random variations of many factors (radio channel, topology, data traffic, etc.), implying the dynamic management of a large number of parameters in the most efficient, effective, and reactive way.

To this end, a number of key research problems have been studied (and solutions have been proposed) by Ad Hoc networking researchers:

- self-configuration and self-organization in infrastructure-less systems;
- support for cooperative operations in systems with heterogeneous members;
- multi-hop peer-to-peer communications among network nodes;
- network's self-healing behavior providing a sufficient degree of robustness and reliability; and
- seamless mobility management and support of dynamic network topologies.

In the following sections, we will discuss more details of the Ad Hoc networking concepts applied to WSNs. In Section 2, we sketch the architecture of a generic WSN node, just to highlight some peculiarities – quite different from typical elements of a data network – that directly impact the networking operation. In Section 3, we investigate the most important requirements emerging on the basis of the communication features of WSN networks, and we show how an Ad Hoc approach could apply to specific applications, in terms of communication model and traffic characterization. Section 4 identifies some 'added values' of the WSN networking features. Section 5 illustrates some of the state-of-the-art solutions and offers a short perspective on what the proposed standards have already captured. Conclusions are reported in Section 6.

2. DEVICE FEATURES

A possible simplified functional architecture of a generic WSN device can be found in [2] and is shown in Fig. 1.

Altogether, the illustrated components form a complex system which can be configured according to a wide set of application

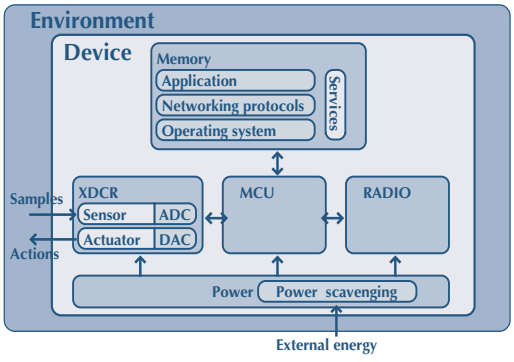

FIGURE 1: THE FUNCTIONAL ARCHITECTURE OF A WSN DEVICE.

requirements. By means of a *transducing unit* (XDCR), which may host multiple kinds of sensors and/or actuators [3], the device interacts with the surrounding environment by collecting 'samples' of environmental characteristics (e.g., temperature, humidity, pressure, etc.) and by causing 'actions' (e.g., air conditioning, light level adjusting, fire alarming, etc.). A *processing unit* (MCU) controls the overall system and manages the procedures that make the device collaborating with the other network members to carry out the assigned tasks.

A *memory unit* stores all processed data, networking control information, and the code to be executed by the MCU (e.g., application, networking protocols, operating system, services), thus providing the device with direct access to the information it needs. A *radio interface* unit connects the device to the network through the wireless channel and makes it able to share information with its peers via packets' exchange. The whole system is usually powered by a battery pack that might be recharged by additional ambient power scavenging units such as solar cells.

Besides flexibility, common WSN applications demand low cost devices, imposing severe constraints on the system design and configuration.

Typical WSN devices could be as large as coins, possibly resulting in limitations of processor power and memory. Moreover, they are often expected to be untethered, which emphasizes their capability of saving energy while allowing the completion of network operations. Furthermore, given the need for such devices to be used in hostile environments, the robustness of these devices becomes a central issue.

The kind of operations WSN devices may participate in when networked depends on the solutions to the above issues, as will be illustrated in the following sections. In this paper, we point out the two key factors that most differentiate WSNs from other data networks: the *application dependency*, which mainly impacts the communication features, and the *Ad Hoc networking approach*, which specializes in the monitoring and control contexts.

3. COMMUNICATION FEATURES

WSNs are operated for 'instrumenting' their surroundings [4]. They typically count on a large number of devices that are able to detect and/or react to external stimuli associated with events or objects of interest. As already mentioned when we introduced the concept of *cooperation*, these devices, when networked, may work together to perform complex monitoring and control tasks, depending on the intended application (e.g., the devices may be demanded to report simple detections of the presence of a moving object or the number of such kind of detections performed in a predefined time interval). The application dependency, combined with other peculiarities of WSNs (wireless channel usage, device heterogeneity, device redundancy, to name a few), impacts on the network communication model (i.e., the roles assumed by devices in the communication context, the rules adopted for information gathering and distribution) and on the characteristics of the data/control traffic injected into the network.

3.1 Communication model

Users of a WSN are likely interested in gathering information about environmental conditions, current events, objects, or living beings as well as in remotely enabling a conditioning system located inside some possibly hazardous place to regulate its environment's characteristics. Monitoring and control tasks can be *data centric*, that is, they could aim to react to the existence of a given set of information rather than being triggered by *who* produces some portion of the information set. This consideration becomes more relevant when taking into account that device redundancy is normally exploited to improve accuracy and reliability of the network operations; for instance, the correctness of a metering procedure can benefit from the collection of correlated data coming from multiple devices, in terms of both data availability (e.g., active originators may be replacing failed ones in producing data) and measure settling (e.g., mean of different samples). On the other hand, deploying a large number of transducing devices in a region for the purposes previously described would make the reference to each single network's member (e.g., address) less relevant. This is because it

contributes to operating the network without being essential for that; in other words, the single device may become 'anonymous' in the communication context.

This line of reasoning distinguishes WSNs from traditional networks in the characterization of the *communication paradigm*; the *address centric* approach used in end-to-end transmissions between specific devices, with explicit indication of both source and destination addresses in each packet, could be replaced with an alternative (and somewhat new) *data centric* approach [5], [6]. Perhaps, a straightforward effect of such 'address blindness' would be related to the choice of the most suitable data diffusion strategy for data centric networks. As an example, in a WSN, a great amount of similar data (i.e., highly correlated samples) might be produced by multiple sensor devices that are reporting the occurrence of a common phenomenon. In a well-known data centric technique called *in-network aggregation* [5], [6], rather than separately propagating these data items to data consumers, redundant data items may be aggregated, depending on the target application, as they are flowing through the network, so that considerable energy and bandwidth savings can be achieved.

These considerations may lead designers to model this kind of network as a sort of distributed data repository where anonymous devices may behave as data originators (also denoted as *sources*) and/or data recipients (also denoted as *sinks*), depending on many factors (e.g., application requirements, device capabilities, and device resources). The typical network deployment would consist of the sources placed around the areas to be monitored and the sinks near easily accessible places, the sinks provided with adequate storage capacity to hold the data from the sources. Sources may send information to sinks in accordance with different scheduling policies: periodic (i.e., *time-driven* mode), event specific (i.e., *event-driven* mode), a reply in response to requests coming from sinks (i.e., *query-driven* mode), or some combination thereof [7]. It is worth noting that, if the sources are unaware of identities of the sinks (as would be the case with the data centric communication paradigm), their data will be 'flooded' over the network, eventually reaching their recipients. Conversely, if the sources know the sinks' addresses, data packets can include such information to be forwarded only to their intended destinations.

Finally, a 'mixed' communication paradigm, which is data-centric (i.e., one based on anonymity) at the source side and address-centric at the sink side, can be defined for WSNs. To the best of our knowledge, such a mixed communication model has not been defined or investigated in the literature. A sink might publish its interest in data by issuing a query packet and by attaching its address so that every receiving source would be able to refer to it. Queries might also bring time or event indications along to give instructions about the data transmission scheduling to the receiving sources; in this way, time-driven and event-driven scheduling policies would be specializations of the query-driven one and thus allowed.

In the *mixed model*, the efficiency of data delivery obtained by limiting it only to the interested collectors would not be exploitable in any way for the query distribution, as it is not supposed to differentiate among anonymous sources. Nevertheless, for most applications, the sinks' interest should be directed towards sources located inside specified target regions, for instance, lamp switches in a building room [6], [8]. To allow this to occur, queries might somehow specify their target regions so that every receiver of a query packet might determine whether or not it belongs to the reported target region, or, in other terms, whether or not it is a candidate source with respect to that query. Obviously, with these assumptions, the adoption of a localization system that enables devices to know their current position (see Sec. 3.3) becomes mandatory.

3.2 Traffic characterization

One of the key aspects of WSNs is the possibility of easily tasking sensor devices to produce information about a certain phenomenon by giving them the necessary instructions about the type of information to be produced, the type of actions to be taken

for getting that information, and the time-plan for the reports' production. In addition, a general agreement among all network members on the coding of the above instructions/information would ease the system control and increase the efficiency; in other words, adopting the terminology used in [6], data and/or control messages in WSNs should be *named*, that is, structured by the aggregation of 'attribute-value' pairs that all network members can interpret. To this end, all devices should refer to the same *code-book*, locally stored as a reference, defining a list of attributes (e.g., type of event, position, time, intensity, accuracy, etc.) and their respective set of valid values. Besides defining a common 'language,' the naming techniques could be exploited for allowing in-network aggregation (introduced in the Sec. 2.1) of multiple data with respect to some common attributes.

The benefits of the in-network processing, the need for saving limited network resources, as well as the utility-based filtering of information to be transmitted would make preferable the extraction of essential features from detected phenomena, instead of detailed reproduction. For example, a video coming from a camera (e.g., for detecting intruders or fire) could be as helpful as a simple signal sent to the control system, but it would require more resources and a more complex transmission management.

Consequently, WSNs are not commonly expected to carry multimedia data flows or, in other words, sequences of time-correlated information that cannot be handled one at a time. Nonetheless, some kind of quality of service (QoS) support should be provided for managing 'time critical' data (e.g., fire alarms) whose delivery may require upper bounded time interval (i.e., latency). To meet such a requirement, transmission scheduling solutions for preventing the network congestion which would cause unacceptable delivery delays should be addressed, even by considering that congestion may be frequent in WSNs, due to several factors (e.g., use of a shared wireless medium, limited bandwidth and buffer size, high level of local contention, and the multi-hop nature of WSN) that are discussed in Sec. 4.

3.3 A key service: the Localization system

The location information management is an essential feature in most WSN applications [2]. In most cases, the data gathered by devices are really useful if and only if they are stamped with location and time information; for example, a temperature value without indication of where and when it was detected is considered meaningless.

The location information could be exploited for both application and networking scopes, for example, to gather data from a specified geographic zone, as already described in Sec. 3.1. According to the mixed communication model, an efficient routing protocol might introduce some smart rules for delivering queries only to nodes inside target zones, and corresponding data only to queries' originators, respectively [6]. Discovery and tracking services (e.g., people, assets) are possible examples of location-enabled application functionalities required in industrial scenarios.

The location information should be reliable and available, no matter where the device is placed. Technologies such as the Global Positioning System (GPS) [9] could be used to obtain the location information needed, but they may be sensitive to external disruptive factors or to physical obstacles, regardless of any considerations about costs and power consumption. A similar reasoning can be extended to the use of a centralized location information database, which may not be fault-tolerant enough when implemented by one single location server. Hence, the aim of achieving robustness may suggest building up a distributed or replicated localization system [10].

Besides reliability and availability, another main feature of the localization system is its accuracy. Again, a GPS without signal corruption provides quite accurate absolute location information. Alternatively, under less favorable conditions, the accuracy should be evaluated in relation to the *ranging* capability of devices. This consists of making devices able to estimate the distance from their *neighbors* (i.e., devices able to exchange data packets directly) on the basis of measurements of the

received signal strength and/or the propagation time (e.g., *time of arrival*, ToA) of predictable control signals. Ultra Wide Band (UWB) [11] – the impulse-radio version [12] – is an example of technology that can provide high ranging accuracy by estimating the round trip time between transmitter and receiver. A number of such technologies and solutions are presented in literature [10]; methods based on triangulation allow 'blind' devices to infer their position from the collection of measurements obtained by assuming the knowledge of reference points, called 'anchors', even for devices without ranging capability [13].

4. NETWORKING FEATURES

WSNs are a class of wireless networks that may be generally conceived in two different topologies, i.e., *star* and *mesh* [14]; the former is basically a one-hop topology, where all terminals communicate with a central node, while the latter refers to the most general multi-hop topology, where multiple nodes implement a distribution system for data communication among wireless terminals.

The focus of this paper is on mesh-configured networks. Such kinds of network can be 'self-contained' when information/commands generated by nodes/terminals are directed towards resident nodes/terminals; nevertheless, gateway nodes may be provided to allow connections with the 'external world' – e.g., Internet. Mesh-configured WSNs are supposed to have no need of any fixed, centralized management infrastructure, while they are instead to rely on the adoption of the cooperation-based *Ad Hoc* networking approach [2, 15]. Each node, besides being a sensing or actuating device, may act as router, at which point it is able to receive packets generated by other devices and forward them to the next hop along a multi-hop path towards their final destination.

Therefore, each device has to multiplex (share) the available link bandwidth with all active neighbors. Because the WSN density and utilization needs may be high, the bandwidth may be negatively affected. On the other hand, WSNs are not expected to be communication intensive at the application layer, as their devices are expected to have low communication duty cycles, reducing the average long-term per-device bandwidth requirement [16].

While the traffic bandwidth requirement is not the main WSN networking issue, the *reliability* is strongly expected to be fulfilled [2]. Any WSN is deeply involved in and related to the monitored environment, and any change occurring to the surroundings will significantly influence its performance; nevertheless, the network must be able to tolerate and 'survive' any change by implementing proper reactions and adaptation mechanisms sustaining communications for both sensed data and commands. In order to comply with the *self-working paradigm*, a WSN should implement a set of viable WSN management guidelines, including:

– *self-configuration*, i.e., the ability to automatically and autonomously set relevant parameters to operate according to some given specifications;
– *self-organization*, i.e., ability to detect the presence of the other network devices and to modify the working behavior, accordingly;
– *self-adaptation*, i.e., ability to automatically interpret feedback information and to adjust optimal settings to well operate in the environment; and
– *self-healing*, i.e., ability to detect devices' and links' failures, by providing autonomous reactions to restore operating conditions – *without human intervention.*

To obey the self-working paradigm, WSN protocols should be designed with strong attention to both *device coordination* and *redundancy exploitation* issues, both of which might have to cope with the network member resource heterogeneity.

4.1 Heterogeneity exploitation

Network node heterogeneity will involve both hardware and software.

Possible hardware differences include the transducing capabilities, computational power, storage room, power supply

(e.g., power line, battery), and many others. Being somehow complementary to the redundancy, the *hardware heterogeneity* can be exploited to improve the way the whole network's resources are used, thus helping to improve the network's resilience to failures and prolonging the whole system lifetime. The *resource awareness* can be a good guideline to follow in the protocol design process. Basically, it may govern the assignment of the most important and critical network management roles (e.g., router, gateway, coordinator, etc.) to the devices with the greatest number of resources; further, it may improve the energy consumption optimization by allowing dynamic transmission power adjustments or packet exchange rate minimization.

Greater availability of hardware resources on any device may be translated into the possibility of putting on that device more computation intensive codes (e.g., protocol stack, application programs) – i.e., the hardware heterogeneity may turn into *software heterogeneity*. For instance, similarly to what has been already proposed by the IEEE 802.15.4 group [14], WSN devices may be classified into *terminals* and *nodes*: terminals can be identified as the network's 'users', or 'end points', so that they can act as source and/or sink only and participate in a restricted set of network operations; conversely, nodes are meant to run the full set of networking functionalities (e.g., coordination, address assignment, packet relaying, etc.). Altogether, the nodes form a distribution system for packets coming from or directed to terminals, while simultaneously giving them the possibility to access the network services. Software heterogeneity should be considered while dealing with the devices' coordination, since it puts some ties on the way nodes can share their functionalities; for instance, any failed node can be replaced only by another one having similar resources. Moreover, it could require smart capabilities sharing policies, since the network may potentially provide several functions, though widely "scattered" throughout itself.

4.2 Device coordination

Device *coordination* is the way the network may achieve both self-configuration and self-organization. Basically, it means that devices have to work together to make the network effective. Even if coordination may be achieved through either centralized or distributed protocols, the latter are usually preferable; in fact, distributed approaches, despite being less efficient, are likely more fault-tolerant and, therefore, more suited to the typical WSN scenarios.

Distributed protocols could be partially supported by pseudo-centralized management and coordination schemes even in the absence of any fixed infrastructure, by making devices able to self-organize into some form of stable hierarchy. This is the case of the *clustering algorithms* (see, for instance, the ZigBee Cluster-Tree organization [17], or Ref. [18, 19]), where some nodes may be elected to play a leading role such as the cluster-head. The clustering concept is differently applied to WSNs with respect to address-centric networks, according to the used communication model (see Sec. 3.1). For instance, cluster-heads may be their clusters' representative (e.g., the cluster identifier is set to the cluster-head's one) [3] in networks where member addressing is needed, while they are often used to perform data in-network aggregation to decrease the number of transmitted messages in WSNs [20]. The advantage of clustering can be better exploited under the distributed management and coordination viewpoint when such algorithms are smart enough to assign leading roles to devices having more capabilities and resources than others (see Sec. 4.1). In case of node failures and/or topology changes, clustering algorithms must be able to dynamically perform 'cluster re-organization' procedures by adapting the role assignment. For example, [20] suggests a randomized rotation of the role of the cluster-head so that uniform energy dissipation due to data collection and merging is obtained. In addition, the clustering scheme should be properly assisted and considered in the cross-layered design of network communication protocols to cope with resources, stability, and performance issues [19]. Hence, WSN coordination functionalities must not rely exclusively on a single device (or a static group of devices) so that the effectiveness of the operations can be preserved also under conditions of failure.

Two factors have to be taken into account while dealing with the distributed protocol design process, namely the *network dimension* and the *network topology dynamics*. The former, easily appreciable through different system parameters (e.g., number of nodes, network extension, node density, etc.), is very important while assessing the protocols' *scalability*, i.e., their ability to tolerate performance degradation 'smoothly', and thus acceptably, when the network's dimension grows [21]. The latter tells us how fast the network topology is prone to change because of the occurrence of disruptive events (e.g., failures of nodes running out of energy, mobility, variable channel conditions, etc.), which impacts the ability of protocols to maintain their correct working behavior.

4.3 Redundancy exploitation

Redundancy exploitation means designing protocols able to take effective advantage of the existence of multiple information sources, network devices, routing paths, and so on. Networking protocols, to be properly deemed as self-healing and self-adapting, should be capable of smartly and dynamically choosing among several available resources; in case some resources become unavailable, the protocol must be able to select and to start using the remaining ones.

Many examples of redundancy exploitation can be found among routing solutions for WSNs. Reference [22] proposes an energy-aware routing protocol that maintains a set of paths alternately used to convey data packets so that the energy budget of any single path will not be depleted quickly. In [6], the Directed Diffusion strategy performs in-network aggregation of redundant data coming from different sources, building up a robust routing mesh that gets data from multiple sources to multiple sinks. Reference [23] proposes a variant of Directed Diffusion called Gradient-Based Routing. Its key idea is to allow network devices to know the hop distance from a common sink and to force data to pass through nodes with decreasing distances until the data is delivered to the sink; exploiting the typical high density of WSNs and the likely presence of multiple relayers with the same distance from the sink, the forwarding process becomes highly reliable. Device redundancy is exploited also by some MAC schemes. GeRaF [24] is a technique that integrates routing, MAC, location awareness, and topology management to forward packets just towards their intended destinations. In this scheme, a node calculates its priority in acting as a relayer of a received broadcast message on the basis of its position towards the final destination; thus, a set of best positioned nodes may volunteer to relay the message, possibly originating multiple paths to be followed.

As the resource unavailability is more likely to occur as a consequence of node failures – e.g., node breakages or energy depletion – the management of such failures is a main issue in WSNs. A failure classification is thus helpful here, as it allows for a better understanding of which management policies are required to handle them. A straightforward categorization can split them between *predictable* or *unpredictable* ones. The former require an *explicit signaling* approach to anticipate correct actions before the failure occurrence; the latter can be handled only through *reaction-based* (adaptive) mechanisms, as failures occur in an unpredictable way. Given that the unpredictable events in WSN scenarios are likely to occur more often than the predictable ones, the self-adapting and self-healing capabilities are expected to be more effectively supported by a reaction-based (i.e., adaptive) mechanism. This policy may then realize an acceptable trade-off between simplicity and efficiency issues with both kinds of events.

5. SOME 'AD HOC CONCEPTS' IN THE EXISTING STANDARDS

Today's users, industrial, professional, or consumer, are maturing in their awareness of and practice in the adoption of pervasive computation and communication systems like WSNs. The 'sensor and control arena' has been enriched by a wide range of possible choices, which unfortunately are mostly composed of many proprietary solutions that are almost incompatible. Standardization efforts are generally a positive factor in moving

producers and vendors to concrete actions for developing largely reusable and interoperable components and/or devices, with benefits in terms of time to market, reliability, and cost reduction for users. Without standards, which require us to put 'progress' ahead of 'self-interest', there would be no mass production or mass communication and, as such, no modern economy and progress.

In the WSN field, several standard efforts have been started recently; most of them have been in evolution for several years and are now beginning to have market impact that will grow over the next few years. There are standards related to wireless communications and sensor management promoted by the *Institute of Electrical and Electronics Engineers* (IEEE) and those focusing on the application scenarios promoted by *Industrial Alliances*. We will consider only the networking-related standards, starting from a short description of *IEEE 802.15.4* that addresses the lower layers of the ISO/OSI networking reference model for wireless personal area networks (WPAN). Then, we will introduce *ZigBee*, based upon *IEEE 802.15.4*, as an industrial standard technology for providing both network and application management with enlarged capabilities obtained by taking up some of the solutions to the typical challenges of the Ad Hoc networks; these will be highlighted along with a short description of ZigBee's main peculiarities.

5.1 IEEE 802.15.4

The IEEE 802.15 working group defines the physical layer (PHY) and the medium access sub-layer (MAC) for low-complexity, low-power consumption, low bit-rate WPAN connectivity; currently, among the four IEEE 802.15 different frameworks, the IEEE 802.15.4 is considered the most relevant for WSN scenarios.

A global RF standard for WSN is fundamental because it would accelerate the technology evolution by identifying leading directions, and it could yield benefits such as the lowering of design costs and the interoperability at the communication level. Moreover, at the raw physical level, standards have to cope with the issue of frequency band allocation. Different regulations in different regions of the world should not be a problem for most applications, whereas the most used wireless technologies, including all those suitable for WSNs, fall in the ISM bands; this could bring the advantage of interoperability but also a concentration of technologies, with possible problems for the coexistence of different wireless systems.

Nevertheless, the available frequencies are only the starting point for operating wireless links. In order to make equipment, software, protocols, and applications, all made by different manufactures, interoperable, the industry needs to set standards. The IEEE 802.15.4 Standard [25], approved in 2003 and amended in 2006 with a "b" version, is contributing to all of these aims, and several compliant products are already available on the market, even if more as development kits only than real end-products. The standard provides for a low bit-rate (i.e., up to 250kbps, or 1 Mbps in "b" version) connectivity in the *Personal Operating Space* (POS), typically 10/100 meters, through 27 RF channels in the ISM RF bands, in order to guarantee wide adoption in several market segments worldwide.

The most innovative IEEE 802.15.4 feature is the full support of *mesh networks* for battery powered nodes, through the classification of devices into two different types – i.e., Full Function Device (FFD) and Reduced Function Device (RFD). An IEEE 802.15.4 network should include at least one FFD operating as the PAN coordinator for special (but not centralized) functions, whereas all the other FFDs would contribute to realize the WSN backbone; RFDs, which are usually intended as the 'leaf' nodes of the WSN spanning tree, perform simple tasks more related to sensing than networking. An RFD can communicate only to one FFD, while an FFD can communicate to both RFDs and FFDs.

This asymmetry of roles arises from the fact that RFDs have minimal resources (energy and memory) and basically act as 'parasites' of FFDs functions by limiting working activity to (typical) 1% duty-cycle.

As anticipated, the Standard has recently come out of revision, introducing simplifications to the overall architecture and resolving some design ambiguities and inconsistencies, while improving interoperability worldwide. Concurrently, under development is IEEE 802.15.4a, which essentially regulates the use of UWB physical medium with enhanced technical and protocol specifications, thus allowing high bit-rate transmissions and new management features like ranging/localization (described in Sec. 3.3).

5.2 ZigBee

On the basis of the current IEEE 802.15.4 specifications, a consortium of more than 200 companies is negotiating and working on the adoption of an industrial standard called ZigBee, whose name and working principle is inspired by the social behavior of bees that work together to tackle complex tasks. ZigBee exploits cooperation to allow for the multi-hop exchange of messages, as described in Sec. 4, and adds the logical network, security, and applications management on top of the referenced IEEE 802.15.4 standard by defining the upper layers of the protocol stack, from network to application (see Fig. 2). In addition, ZigBee defines *application profiles* [17], which refer to a set of template-based description of device configurations, each one specialized for working to a common cooperative and distributed application. Aside from its technical aspects, one of the main tasks of the ZigBee Alliance is to certificate interoperability among devices made by different manufacturers, thus expanding their potential adoption.

Technically speaking, ZigBee is a fairly good standardization effort that is gaining wide acceptance of big players, but it is currently subject to several refinement attempts, which would risk reducing its market momentum. The first official release of the standard, known as "ZigBee v.1.0" and dated December

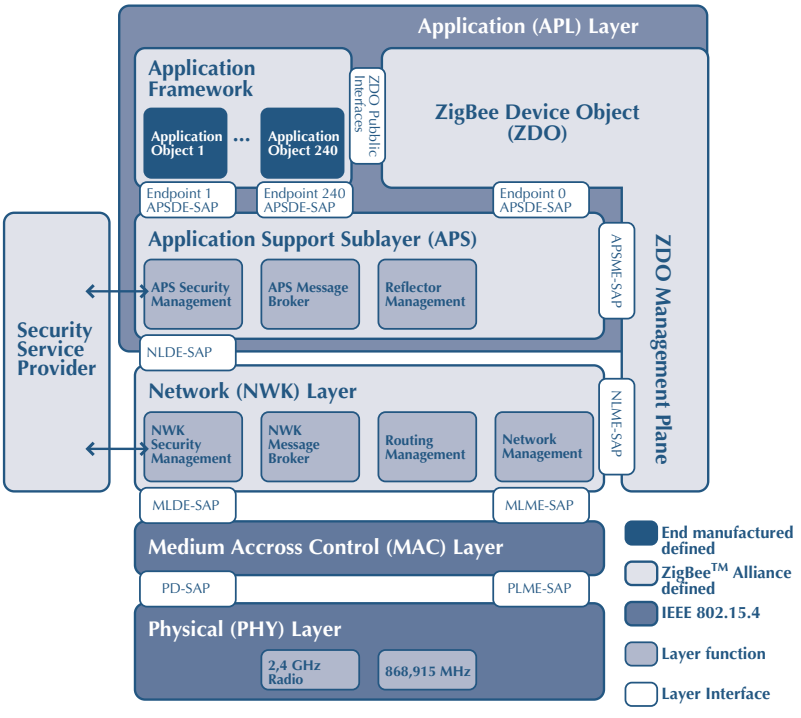

FIGURE 2: THE ZIGBEE PROTOCOL STACK.

2004, is currently shipped by many platform providers. It is typically referred to as the "Home Control" version of the stack. Unfortunately, its lifetime is marked, and it is changing in a non-backward-compatible way, while maintaining the overall skeleton. For the time being, the Alliance has decided to propose ZigBee in two flavors: a base version featuring the Home Automation market only, named "ZigBee", which is an enhancement of the "old v.1.0," and a "ZigBee Pro" version, most likely available starting in 2007, which will incorporate several features we listed in the previous sections.

The new releases improve ZigBee networks' scalability and performance through full support of mesh networking, abandoning the hierarchical organization of the "old v.1.0" known as "Cluster-Tree," and bringing in the so-called "many-to-one routing" feature, which allows a concentrator (a gateway or more generally a sink) to establish routes from all nodes in the network without large increases in route table size and without creating a broadcast storm within the network.

This is a first step in the direction of a data centric paradigm (see *mixed mode* in Sec. 3.1), which, together with the support for "multicast," will fill the gap with the 802.15.4b added support. Even more relevant is the new "alternate addressing scheme," which will replace the tree organization by stochastically assigning the network level addresses and resolving conflicts where originated. This will result in a lighter management of addressing issues – as the address of the node will no more depend on its physical position on the tree – enhancing the scalability and thus allowing pure mesh topologies as well as node mobility.

The concept of reliability is hence strengthened by mesh networking because its many different and dynamic paths could be exploited for routing. The "frequency agility" feature enhances reliability as well. Starting from the assumption that the channel and the entire RF band conditions can vary dynamically in both time and space, this feature will allow the operation to continue reliably and in an unattended way in the presence of well-known interference sources, like WiFi or Bluetooth, or in the presence of an unusable operating channel as well. This is a big step towards the *self-working paradigm* (see Sec. 4).

The heterogeneity exploitation is also being considered for the 2007-later on versions, by adding features like "end-device management" and "battery powered routers", allowing, respectively, the employment of battery-powered devices as leaf nodes - by designating the relevant parent router as a proxy - and even permitting the deployment of battery-powered router nodes, given that in the current version they are supposed to be 'mains' (i.e. line) powered.

If all these are networking add-ons that will differentiate "ZigBee Pro" from previous versions, two important enhancements will have been produced at the application level, starting from the "ZigBee" stack. The old mechanism used to describe and identify the application profiles has been updated with the new "ZigBee Cluster Library (ZCL)," which can be imagined as a code-book (see Sec. 3) where all devices, classified on the basis of functional domains, have been described in terms of their attributes/properties in an application-independent fashion; as an example, the same description for a temperature sensor can be adopted even if it will be used in a Home or Industrial Automation application.

All these are only a part of the discussed features in the ZigBee Alliance. This can be expected, given that such an organization counts several members with different strategies and requirements, such as OEMs as well as application-oriented and semiconductor companies. Now momentum has come, and 2007 will be the year expected for launching ZigBee to the mass market. At the same time, ZigBee will continue to enhance its capabilities; mobility management, location awareness, QoS, power awareness, network-wide and time synchronization, and new routing strategies will be the ZigBee response to the classical Ad Hoc paradigm challenges.

6. CONCLUSIONS

WSNs are an emerging application field that deserves the great attention of the microelectronics industry and research. This paper presents an overview of WSNs with special focus on networking challenges and issues due to the peculiarities of such networks, mainly identified as application-dependency and natural demand for the 'Ad Hoc' approach. ZigBee / IEEE802.15.4 promises this kind of network availability in a fairly short time, but, in our opinion, further steps still need to be taken before making WSNs truly user-friendly.

FIGURE 3: CHALLENGES AND ISSUES OF WSNS.

The bubble diagram in Fig. 3 provides a list of main issues that the natural technology evolution and the market raise. Some of them have already been tackled by ZigBee / IEEE802.15.4, while some others are still representing challenges for academic and industrial researchers, in both the fields of general and Ad Hoc WSNs.

According to recent marketing reports such as those found in [26], today's wireless sensor industry is being driven by customer demand for reliability, simplicity, price and availability, independent of the standardization processes, whose advancement is often limited by external influences and political or strategical issues. This can be seen today in the home automation market, where leading companies are simply and readily reacting to customer needs, while the standards often adapt to the market-based process. The companies with the products that solve real customer needs and desires in a timely manner will likely become the standard in the home market.

The recent partnership between ST and Ember Corporation [27] will enable ST to become one of the leading providers of the ZigBee networking systems. However, ST has also started thinking about the next generations of WSN devices in order to preserve its competitiveness while gaining a competitive advantage by adding improvements at the device and software levels. In this sense, STMicroelectronics contributes to the growing knowledge of WSN solutions, continuously fed by relationships with the academic world. Distributed localization algorithms, geographical routing, and fairness and congestion control at MAC layer are only some of R&D areas under investigation, areas whose importance has been extensively pointed out in the paper.

REFERENCES

[1] M. Weiser, "THE COMPUTER FOR THE 21ST CENTURY," Scientific American, 265(3), pp. 94-104, 1991.

[2] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, "A SURVEY ON SENSOR NETWORKS," IEEE Communications Magazine, 40(8), pp. 102-114, 2002.

[3] D.J. Baker, A. Ephremides, and J.A. Flynn, "THE DESIGN AND SIMULATION OF A MOBILE RADIO NETWORK WITH DISTRIBUTED CONTROL," IEEE J. Sel. Areas Commun., Sac-2(1), 226, 1984.

[4] D. Estrin, L. Girod, G. Pottie, and M. Srivastava, "INSTRUMENTING THE WORLD WITH WIRELESS SENSOR NETWORKS," Proc. of ICASSP, 2001.

[5] B. Krishnamachari, D. Estrin, and S. Wicker, "MODELING DATA-CENTRIC ROUTING IN WIRELESS SENSOR NETWORKS," Proc. of IEEE Infocom, 2002.

[6] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, "Directed Diffusion for Wireless Sensor Networking," Proc. of IEEE/ACM Transactions on Networking, 11(1), 2003.

[7] Y. Yao and J. Gehrke, "The Cougar Approach to In-Network Query Processing in Sensor Networks," Proc. of SIGMOD, 31(3), 2002.

[8] Y. Yu, D. Estrin, and R. Govindan, "Geographical And Energy-Aware Routing: A Recursive Data Dissemination Protocol For Wireless Sensor Networks," Technical report UCLA-CSD TR-01-0023, 2001.

[9] "Navstar Global Positioning System Surveying," ASCE Publications, Technology & Industrial Arts, 2000.

[10] N. Patwari, J.N. Ash, S. Kyperountas, A.O. Hero III, R.L. Moses, and N.S. Correal, "Locating the Nodes - Cooperative localization in wireless sensor networks," IEEE Signal Processing Magazine, pp. 54-69, 2005.

[11] M.G. Di Benedetto and G. Giancola, "Understanding Ultra Wide Band Radio Fundamentals," Prentice Hall Communications Engineering and Emerging Technologies Series, 2004.

[12] IEEE 802.15 WPAN Low Rate Alternative PHY Task Group 4a (TG4a)

[13] D. Niculescu and B. Nath, "DV Based Positioning in Ad Hoc Networks," KLUWEIR Journal, 2003.

[14] http://standards.ieee.org/reading/ieee/std/lanman/restricted/802.15.4-2003.pdf

[15] S. Rotolo, D. Blasi, V. Cacace and L. Casone, "From WLANs to Ad Hoc Networks, a new Challenge in Wireless Communications: Peculiarities, Issues and Opportunities," in Handbook Of Wireless Local Area Networks: Applications, Technology, Security, and Standards, CRC Press, pp. 2.15-154, 2005.

[16] E.H. Callaway, **Wireless Sensor Networks – Architectures and Protocols**, Auerbach Publications (CRC Press), 2004.

[17] ZigBee Alliance internal documents (http://www.zigbee.org)

[18] D. Simplot-Ryl, I. Stojmenovic and J. Wu, "Energy-Efficient Backbone Construction, Broadcasting and Area Coverage in Sensor Networks," in Handbook of Sensor Networks: Algorithms and Architectures, John Wiley & Sons Inc., pp. 343-380, 2005.

[19] L. Bononi, M. Di Felice, L. Donatiello, D. Blasi, V. Cacace, L. Casone, S. Rotolo, "Design and Performance Evaluation of Cross Layered MAC and Clustering Solutions for Wireless Ad Hoc Networks," Performance Evaluation 63 (2006), Elsevier, pp. 1051-1073

[20] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Micro sensor Networks," Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS '00), 2000.

[21] F. Martincic and L. Schwiebert, "Introduction to Wireless Sensor Networking," in Handbook of Sensor Networks: Algorithms and Architectures, John Wiley & Sons Inc., pp. 1-40, 2005.

[22] R. C. Shah and J. Rabaey, "ENERGY AWARE ROUTING FOR LOW ENERGY AD HOC SENSOR NETWORKS," IEEE Wireless Communications and Networking Conference (WCNC), 2002.

[23] C. Schurgers and M.B. Srivastava, "ENERGY EFFICIENT ROUTING IN WIRELESS SENSOR NETWORKS," MILCOM Proceedings on Communications for Network-Centric Operations: Creating the Information Force, 2001.

[24] M. Zorzi and R. R. Rao, "GEOGRAPHIC RANDOM FORWARDING (GERAF) FOR AD HOC AND SENSOR NETWORKS: ENERGY AND LATENCY PERFORMANCE," IEEE Transactions on Mobile Computing, 2(4), 2003.

[25] IEEE Std 802.15.4™-2003 Standard: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs).

[26] K. West, "WIRELESS SENSOR TECHNOLOGY AND MARKET TRACKING SERVICE: ZIGBEE, ZWAVE, INSTEON, RFID, IEEE 802.15.4 AND THEIR COMPETITION," Report abstract WTRS, 2005.

[27] http://www.ember.com/

PICO RADIO: FROM VISION TO REALITY

M. Sheets[1], B. Otis[1], H. Qin[1], N. Pletcher[1], F. Burghardt[1], J. Ammer[1], T. Karalar[1], P. Monat[1], Y. Cao[1], D. Markovic[1], A. Vladimirescu[1], J. Rabaey[1], S. Cervini[2]

(1) Berkeley Wireless Research Center, University of California, Berkeley
(2) STMicroelectronics

The vision of ubiquitous, dense, heterogeneous, ad-hoc Wireless Sensor Networks (WSNs) can only become a reality through the combined efforts of system miniaturization and power/cost minimization.
In turn, these efforts open up a variety of interrelated issues, including protocol stack design, energy scavenging, cost/precision tradeoff of components, and low-power CMOS design.
At BWRC all of these issues have been addressed in a harmonized way, and a WSN node (Quark), incorporating the outcome of multidisciplinary research, has been designed. This article presents an overview of the current state of the Quark node. STMicroelectronics has manufactured many of the chips developed at BWRC under the Pico Radio project.

1. INTRODUCTION

This paper offers an overview of the efforts undertaken at BWRC (Berkeley Wireless Research Center) in the field of Wireless Sensor Networks (WSNs) by a research group led by Prof. Jan Rabaey.

Novel techniques in the areas of protocol stack algorithms, energy trains, node locationing, digital processing, and RF communication circuitry have contributed to making the implementation of a WSN node more energy efficient. At BWRC a WSN node that integrates these technological advances into a self-contained energy-scavenging sensor node (Quark Node) has been developed.
The overall energy consumption of the node is attained by means of a system level codesign of the protocol stack and circuit-level low-power design of two ASICs. The size of the node has been minimized through the design of a custom PCB with an integrated planar inverted-F (PIFA) antenna and the design of the energy train so as to allow the node to be self-contained with batteries recharged by solar energy.

The Quark node is implemented as a printed circuit board that contains two custom chips (the digital Charm and the RF Strange), a PIFA antenna, and a power train, described respectively in

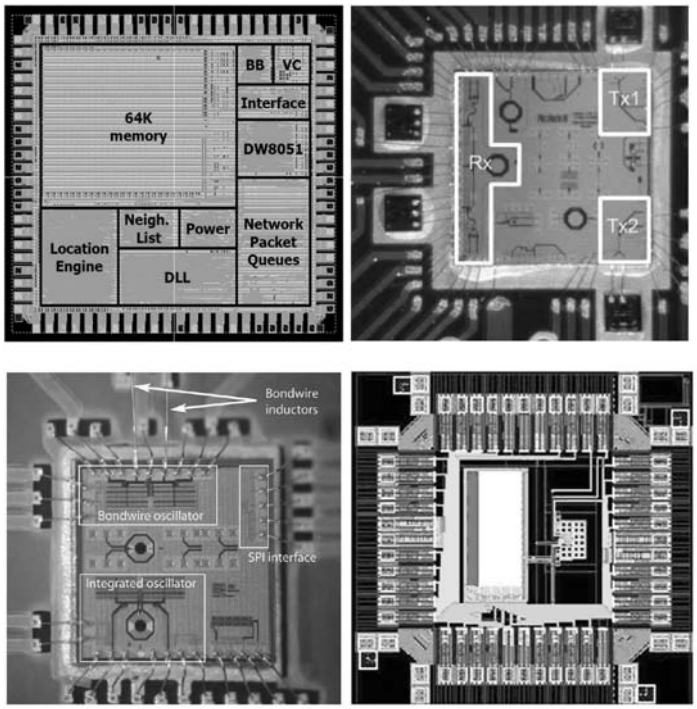

Figure 1: Four chips designed at BWRC and manufactured by STMicroelectronics. a) Charm: Digital Protocol Processor, © 2004 IEEE; b) Strange: Analog Two-Channel Transceiver, © 2004 IEEE; c) 1.9 GHz Digitally Controlled Oscillator, © 2005 IEEE; d) SRAM Leakage-Control Test Chip, © 2005 Elsevier.

subsections 3, 4, 6 and 7. Off-the-shelf analog components for the baseband complete the system. The block diagram of the Quark node is shown in Fig. 2. This paper is organized as follows: section 2 summarizes the functionalities and algorithms used in the protocol stack; section 3 presents the digital processor and how SRAM power leakage can be minimized by setting the VDD at the minimum value as predicted by a DRV (Data Retention Voltage) Model; section 4 describes the design of the transceiver; section 5 presents a wake-up receiver architecture that can be used to reduce power consumption further by allowing the node to remain in sleep mode as long as possible, waking up only when another node is transmitting to it; section 6 presents the main features of the planar inverted-F antenna (PIFA); section 7 describes the energy scavenging and storage subsystems; section 8 shortly discusses optimizations of the clock subsystem; and section 9 presents the power consumption breakdown among the various system components.

2. PROTOCOL STACK

The node implements a wireless protocol stack that is tailored for sensor network applications. Subsystems generally follow the OSI reference model [1] and include application, network, data link, and physical layers. The protocol stack is augmented to include a locationing subsystem because a location is required to associate the sensed data with the source or destination.

2.1 Application Layer

Node application functions fall into three categories: controllers, sensors, and actuators. A sensor application waits passively until it receives instructions from a controller application. Controllers can request a one-shot measurement or schedule measurements to be performed periodically. Upon receiving a request, the sensor application queries the appropriate external sensor and returns the data. Controller applications can also instruct actuators to perform an action, such as toggling an external

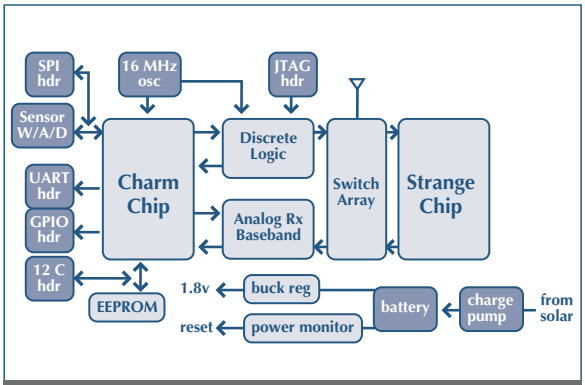

FIGURE 2: QUARK BLOCK DIAGRAM, © 2004 IEEE.

relay. Flexibility in the application protocol layer allows any combination of these three operations to be performed through programmable parameters and detachable user interface, such as a laptop or PDA.

2.2 Locationing

In a self-configuring sensor network, most sensor nodes are deployed without any presumed position. Only a small number of anchor nodes are given a priori information about their positions with respect to a global coordinate system. The rest of the nodes then calculate their positions using the position and distance to these anchors.

The locationing subsystem performs 3-D triangulation using at least four anchor node reference points. If more anchor nodes are present in the network, the error in the calculation is reduced [4]. During triangulation, inter-nodal distances are expressed in terms of the reference and unknown position coordinates. These equations are then cast into an over-determined linear system, which is solved using a least squares algorithm. The hop count between an anchor and the node is used as a substitute for the real Euclidean distance. This algorithm, known as Hop-TERRAIN, was initially proposed by Savarese, et.al. [5]. The sensor network is assumed to be mostly static, so only slow position changes are expected and tracked. The location update rate is programmable and usually operates on an order of minutes.

2.3 Network Layer

The network protocol layer is responsible for implementing the multi-hop routing protocol. The protocol in use is a form of the directed diffusion algorithm in [2]. In order to make the network truly ad-hoc, there are no absolute node addresses. Packets are routed to a 3-D box-shaped region defined by the two locations of its opposite corners. The nearest neighbor links are distinguished by dynamically assigned node identifiers (IDs) exported from the data link layer (DLL).

In this implementation, the controller can initiate an interest packet, which is then broadcast to all nearest neighbors. Neighbors that are spatially closer to the destination repeat the broadcast. Each intermediate node that sees an interest packet keeps a record of its controller source location and the ID of the node from which it came. After the interest packet reaches its destination, a data packet may be returned to the controller location in response. At each intermediate node, one of the saved paths back to the controller location is selected, and the packet is forwarded down that link. Although our current selection metric selects a path with the lowest hop count to the destination, recent results [3] indicate that a stochastic selection based upon energy metrics can increase network survivability. Since the network layer is implemented in software, this change can be made in future revisions. A flexible platform for further exploration of such alternative routing protocols is one of the main benefits of this node implementation.

2.4 Data Link Layer

The data link layer performs ID assignment and maintains link-to-link connectivity. Network layer routing depends on the local uniqueness of IDs within every neighborhood. When a node is activated, it requests the list of used IDs from each nearest neighbor to create a list of used IDs within two degrees of separation (two-hop neighbors). It then selects an unused ID and notifies its neighbors. In certain situations, such as those involving a lost packet or hidden nodes, the assumption of

local ID uniqueness may be violated. A detection mechanism is used to identify and correct this situation.

2.5 Physical Layer

The node physical layer consists of a 1.9GHz two-channel transceiver and a mostly analog baseband. The air interface is a 50kbps OOK modulation scheme, and the symbol pulse shape is determined by the oscillator startup and dampening. Received symbols are matched and filtered in the analog domain and then sliced using an adaptive threshold determined using the packet header. The synch header is 23 bits and consists of a 16-bit string of alternating ones and zeroes, used for bit timing synchronization and threshold-setting, followed by a 7-bit PN sequence used for packet synchronization. The synchronization and detection algorithm is designed to achieve 10^{-4} BER at 13dB input SNR. The maximum packet length is 1024 bits, which is dictated by the 50ppm clock accuracy and the Doppler effects in the channel.

3. PROTOCOL PROCESSOR (CHARM CHIP)

The digital processing chip (Fig. 1a) contains the logic for the upper levels of the protocol stack, including the digital portion of the baseband, and the locationing subsystem. The application and network layers run on an integrated 8051-compatible microcontroller and can be programmed using C code. The DLL, locationing, and digital baseband subsystems are implemented in direct-mapped hardware. The chip is clocked at a relatively slow 16MHz. This relaxed timing requirement was exploited to save power and hardware area. For example, in the case of the locationing subsystem, area savings was achieved by performing the least squares computation in a time serial fashion using a single set of computational units along with a register file.

An aggressive power controller selectively gates the power rails to the eight internal subsystems. An internal voltage converter generates a 300mV data retention voltage used to reduce leakage during idle modes (more on this on 3.1). The Charm chip is implemented in a 0.13μm process with six metal layers. The chip is (2.7x2.7) mm^2 and integrates 3.2M transistors.

3.1 SRAM Leakage Power Minimization

Continuous technology scaling over the past four decades has enabled higher speed and higher integration capacity in VLSI designs. However, the chip leakage power increases about five times each technology generation and becomes one of the main challenges in future CMOS design [22]. In battery-supported applications with low duty-cycles, such as the Pico-Radio wireless sensor nodes [23], cellular phones, or PDAs, in most situations, active power only accounts for a small portion of the system power consumption, with leakage power ultimately determining the battery life. On the other hand, microprocessor designs of today incorporate large memory components, which consume a significant portion of system power budget. For instance, 30% of Alpha 21264 and 60% of StrongARM power are dissipated in cache and memory structures [23].

While activity factor is usually small in memory structures, leakage causes a major part of the memory power consumption. In 70 nm technology, it has been projected that 70% of the cache power budget will be the leakage power [24]. As a result of both the technology scaling and large leakage power dissipation in memory structure, memory leakage suppression is critically important for the success of power-efficient designs, especially ultra-low power (ULP) applications. While the leakage of logic modules in a chip can be effectively controlled by gating off these paths at standby mode, the leakage suppression of memories is especially difficult due to the data retention requirement in such structures. An effective scheme for SRAM leakage suppression in battery powered applications such as wireless sensor nodes is needed. Since leakage power of the peripheral circuits during an idle period can be eliminated by turning off these leakage paths with switched source impedance (SSI) [25], only the leakage suppression of SRAM cells is considered. In the proposed approach, the standby supply voltage (VDD) of the whole memory is minimized with the memory states preserved. As a result of reduced voltage, all the leakage components in an

SRAM cell are effectively minimized. This analysis of ultra low voltage reliable data retention and its results can also be used for future SRAM design in ULP applications with aggressively scaled operational VDD.

Presently, at the circuit level, the most effective low power design methods in minimizing SRAM cell leakage power involve lowering supply voltage and increasing transistor threshold voltage (Vth), both of which are detrimental to the speed of memory read/write operations. For this reason, leakage reduction techniques at this level typically exploit dynamic control of transistor gate-source and substrate-source bias to enhance driving strength in active mode and low leakage paths during standby periods [26].

For example, the driving sourceline (DSL) scheme connects the source line of the cross-coupled inverters in an SRAM cell to negative voltage VBB during read cycle and leaves the source line floating during write cycle. As a result, the cell read access time is improved with boosted gate-source voltage and forward bias of the source–substrate junction of the transistors. The write cycle is also improved since the NMOS transistors in the cross-coupled inverter pair are inactive [27]. Another technique is the negative word-line driving (NWD) scheme. It uses low Vth access transistors with negative cut-off gate voltage and high Vth cross-coupled inverter pairs with boosted gate voltage to achieve both improved access time and reduced standby leakage [28]. The dynamic leakage cut-off (DLC) scheme biases the substrate voltages of non-selected SRAM cells ~2VDD for VNWELL and -VDD for VPWELL [29]. A concern in some of these schemes is that the gate voltages of transistors far exceed VDD, which raises reliability issues [29]. All of these techniques achieve enhanced memory operation speed and suppressed standby leakage current at sub-1V supply voltage compared to those achieved in conventional cell implementation.

At the architectural level, leakage reduction techniques include gating-off the supply voltage (VDD) of idle memory sections or putting less frequently used sections into drowsy standby mode. To achieve optimal power-performance tradeoffs, compiler-level cache activity analysis is employed to balance the potential for saving leakage energy against the loss incurred in extra cache misses. As an example, the cache decay technique applied adaptive timing policies in cache line gating, achieving 70% leakage savings at a performance penalty of less than 1% [30]. To exploit leakage control in caches with large utilization ratio even further, the approach of drowsy caches allocates inactive cache lines to a low-power mode, where VDD has been lowered while preserving memory data [24].

With a conservatively chosen standby VDD in the drowsy caches approach, leakage energy savings of over 70% in a data cache can be achieved [24]. However, the question still remains on the lower bound of standby VDD that still preserves the data, namely DRV. Knowledge of DRV allows a designer to exploit the maximum achievable leakage reduction for a given technology. Furthermore, in the sub-1V low power VLSI designs of today, the reliability requirement on memories has become the bottleneck in further reducing the system VDD. To enable even more aggressive memory supply voltage minimization, an understanding of low voltage SRAM data preservation behavior is required to quantitatively evaluate the SRAM data retention reliability under low VDD and optimize the future SRAM designs for ULV operation conditions.

To obtain silicon verification of the DRV model and explore the potential of SRAM leakage suppression with ultra low standby VDD, a 4 KB SRAM test chip with dual rail standby control was implemented in a 0.13mm technology. Designed for ultra low-power applications, this scheme puts the entire SRAM into a deep sleep mode during the system standby period. The test chip consists of a standard 4 KB SRAM module and a custom on-chip switch-capacitor (SC) converter that generates the standby VDD with 85% conversion efficiency. Compared to the existing SRAM leakage control techniques, the simplicity of this scheme leads to minimized design effort and, therefore, the minimum extra power necessary to support control circuitry.

When designing for an ultra low standby VDD, reliability of the SRAM data retention at idle mode is the top design concern. Besides process variations, the other factors that may disturb the memory state preservation are mainly noises on the standby supply rail and radiation particles. In this scheme, power supply noise is mostly caused by the output voltage ripple of the SC converter. Thus, an appropriate noise margin needs to be provided in order to achieve the desired reliability. Assigning a guard band of 100 mV above DRV for standby VDD gives about 50 mV SNM in an SRAM cell. With the 20 mV peak-to-peak ripple on the SC converter output, a guard band of 100 mV provides worst case SNM of 45 mV, which is sufficient for state preservation. In comparison to power supply noise, the radiation particle events pose a more serious hazard. With parasitic capacitance at the data storage node of about 1fF in a 0.13 um technology SRAMcell, the critical charge (Qcritical) for a 1 V VDD is simulated to be approximately 3 fC. This is the minimum amount of charge injection on the storage node needed to disrupt the state preserved in this cell. For a reduced VDD at 100 mV above the DRV, Qcritical is reduced to 0.5 fC. Considering the danger of data loss (i.e., soft error), a larger guard band is needed. Other options to ensure reliable state preservation include additional storage capacitance [34] or implemenation of error-correction schemes.

For a dual supply scheme, other design considerations include the operation delay overhead due to the power switch resistance, memory wake up delay, and the power penalty during mode transition. Targeted for ultra low power applications, the system requirements of this design are much more stringent with respect to power than performance [22]. In this context, the concern of the operation delay overhead is not crucial. A 200-um-wide PMOS power switch with 30 Ω conducting resistance is used to connect the memory module to a 1 V active mode supply voltage. With the same switch, the memory wake up time is simulated to be within 10 ns, which is typically a small fraction of the system cycle time in battery-operated applications [35].

The wake up power penalty incurred during switching from the standby mode to the full-VDD mode determines the minimum standby time for the scheme if net power savings over one standby period is to be achieved. This break-even time is an important system-design parameter, as it helps the power control algorithm to decide when a power-down would be beneficial. With the parasitic capacitance information attained from the process model, the minimum standby time in this design is estimated to be several tens of microseconds, which is much shorter than the typical system idle time in a battery-supported system.

Layout of the 0.13 um SRAM test chip is shown in Fig. 1d. The two main components are a 4 KB SRAM module and SC converter. This memory is an IP module with no modifications from its original design.

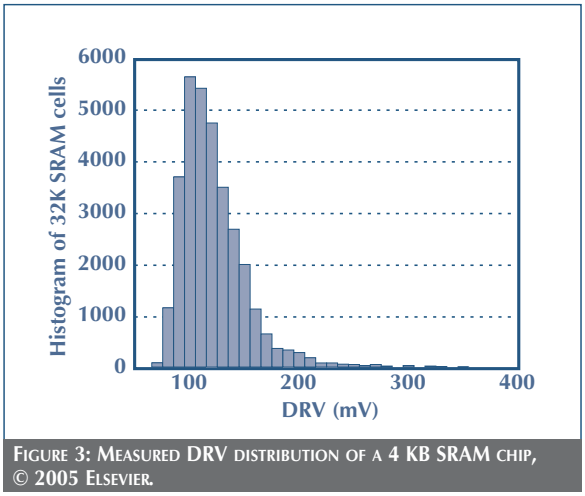

FIGURE 3: MEASURED DRV DISTRIBUTION OF A 4 KB SRAM CHIP, © 2005 ELSEVIER.

Using automatic measurement with a logic analyzer, the DRV of all 32 K SRAM cells on one test chip was measured. Fig. 3 shows the distribution of the 32 K measurement results. The DRV values range from 60 to 390 mV with the mean value around 122 mV. Such a wide range of DRV uncertainty reflects the existence of considerable process variations during fabrication. Due to global variations, the lower end of measured DRV is slightly lower than the 78 mV ideal DRV, assuming

perfect process matching. As a result of large process variation, the long DRV tail at the higher end reduces the leakage reduction achievable by minimizing the SRAM standby VDD.

To improve the gains in leakage power, more advanced techniques, such as error tolerant schemes, are required to cope with this situation. Temperature dependency of DRV was investigated experimentally. When the test chip was heated up to 100°C, a 10 mV increase in DRV is observed. The proposed analytical DRV model not only predicts the ideal DRV values but also fully captures the impact of process and temperature variations.

Thus, it can serve as a convenient base for further design optimizations. Furthermore, Fig. 4 shows the first presented spatial distribution plot of DRV on the measured SRAM chip. From the plot, it can be observed that the on-chip DRV distribution is the combination of random within-die mismatches and systematic deviations on the boundaries of SRAM sub-array blocks. The pattern of SRAM DRV spatial distribution can be exploited in the future work of designing effective error tolerant scheme for even more aggressive SRAM voltage scaling.

Leakage measurement result of the 4 KB SRAM is shown in Fig. 14. The leakage current increases substantially when VDD is high. This phenomenon reflects the impact of process variations

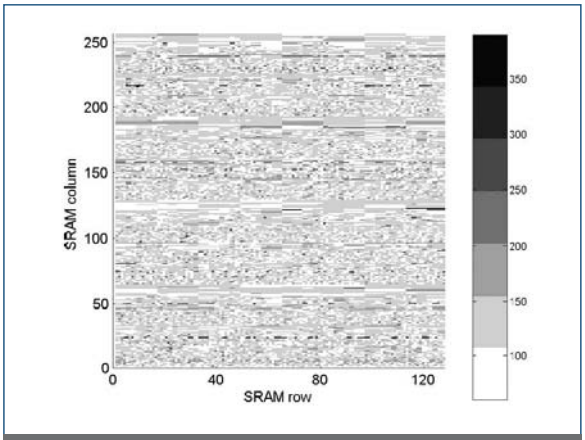

FIGURE 4: DRV SPATIAL DISTRIBUTION OF A 4-KBYTE SRAM CHIP, © 2005 ELSEVIER.

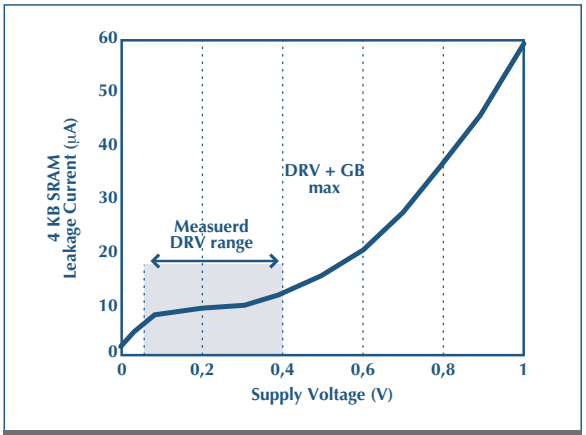

FIGURE 5: MEASURED SRAM LEAKAGE CURRENT, © 2005 ELSEVIER.

on SRAM leakage, more specifically the fluctuations in channel length and Vth. For short channel transistors, a drain-induced barrier-lowering (DIBL) effect causes Vth degradation, resulting in even higher leakage in high VDD conditions. The shaded area in Fig. 5 indicates the range of measured DRV (60–390 mV). Although the memory states can be preserved at sub-400 mV VDD, adding an extra guard band of 100 mV to the standby VDD enhances the noise robustness of state preservation as discussed in Section 3.1. With the resulting 490 mV standby VDD, the SRAM leakage current can still be reduced by over 70%. Subsequently, the leakage power, as the product of VDD and leakage current, is reduced by about 85% compared to 1V operation.

The dual rail scheme is shown to be fully functional through the DRV measurements. With a 10 MHz switch control signal, the SC converter generates the standby VDD with less than 20 mV peak-to-peak ripple. A wake up time of 10 ns is observed during mode transition, while the sleep time spans around 10 us. The delay overhead in the SRAM read operation is measured to be about 2X, which is reasonable for an ultra low-power application where the system clock period is typically 10 times the operation cycle of a low leakage SRAM.

Transistor sizing is an important factor that determines the DRV of an SRAM cell. While sizing has long been an effective technique in conventional power and speed optimization, taking

DRV into account is important for future ULP SRAM designs. In conventional performance-optimized SRAM cells, the pull-down NMOS devices are sized about 2 times larger than the PMOS devices. These NMOS transistors are also typically designed with a smaller L to minimize the cell area. Although providing good stability at high-VDD, this imbalance in the pull-up and pull-down leakage paths leads to exacerbated VTC deterioration at low VDD and degrades DRV. The minimum L of NMOS transistors is also highly sensitive to process variations, which leads to an increase of DRV. Therefore, it is of interest to investigate the impact of each of the sizing variables on DRV.

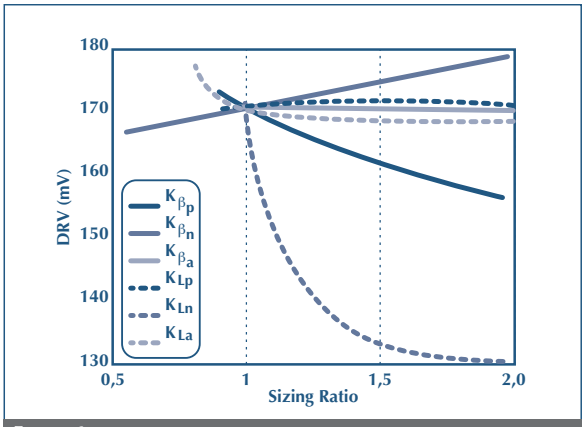

FIGURE 6: DRV AS A FUNCTION OF SIZING PARAMETERS β AND L, © 2005 ELSEVIER.

Fig. 6 shows simulated DRV over the sizing variables βi and Li for different transistors. For each curve, all the other sizing variables are fixed at their nominal values from an industrial SRAM cell design. These simulations assume 3σ local process variations in SRAM transistor channel length and Vth. The X-axis of the plot is the sizing ratio that each variable (β, L) is scaled by. Parameter βi represents (W/L) ratio of the transistor: the pull-up PMOS (βp), pull-down NMOS (βn), and access transistors (βa). The range of sizing ratios in Fig. 6 is constrained by βi >1 and Li > Lmin. From the plot in Fig. 6, it can be observed that DRV can be reduced only by increasing βp or Ln, with the impact of Ln being much stronger. This strong DRV dependence on Ln is the result of its small nominal value (chosen for minimum cell area), which is sensitive to the process variations. Lp has a much smaller impact on DRV due to its larger nominal value.

In summary, the following are the guidelines for DRV-aware ULV SRAM cell optimization for applications in which the read/write performance is not a major concern:

- Increasing Ln SRAM cell reduces DRV most effectively, followed by increasing βp.
- Reducing both βn and Lp improves DRV, but the improvement space is very limited.
- The sizing of access transistors has a negligible effect on DRV.

As an example of power and area tradeoff, Fig. 7a plots the leakage power and SRAM cell area when tuning Ln and βp for the minimized DRV. In this analysis, the SRAM cell transistor area is simply modeled as the sum of the transistor gate areas. A 30% increase in SRAM cell transistor area brings about 30 mV reduction in DRV and almost 70% additional leakage power savings. In Fig. 7b, Ln first exploits the increase of area budget due to its effectiveness in reducing DRV. The effectiveness of increasing Ln is utilized until Ln is about 25% larger than the nominal value, where its impact on DRV drops and, from this point on, βp can be used to further reduce DRV under given area constraint. Although an increase of βp continuously reduces DRV, no more savings in leakage power is attained due to the positive correlation between PMOS leakage and its sizing ratio βp.

4. TRANSCEIVER

The transceiver specifications for wireless sensor networks demand extremely low power consumption, very high levels of integration, and a fast turn-on time for both the transmitter and receiver [9]. These requirements ensure that the transceiver and its energy source will be small and cheap, enabling deployment in quantities sufficient to create a truly ambient intelligent environment. Due to the heavily duty cycled nature of these systems, long transceiver turn-on times dramatically degrade the global efficiency of the system. The level of radio integration

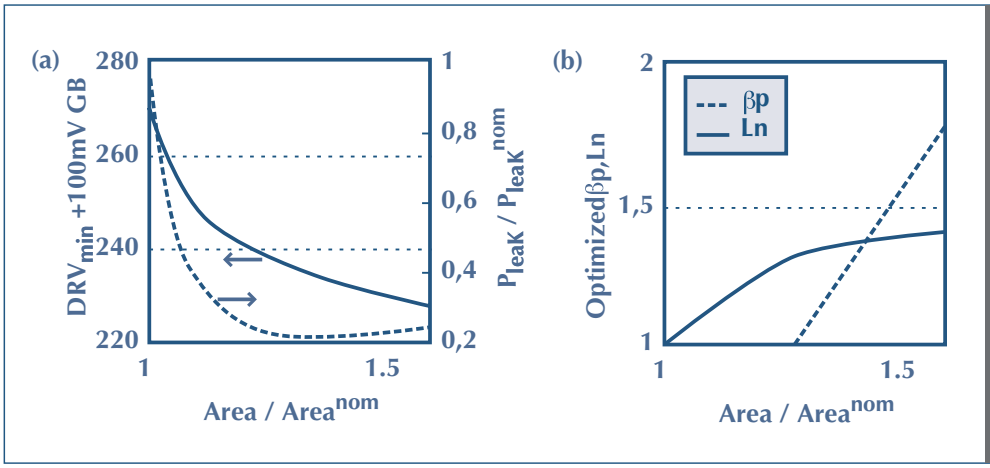

FIGURE 7: DRV-AWARE SRAM OPTIMIZATION WITH βP AND LN. (A) AREA TRADEOFF WITH DRV AND LEAKAGE POWER; (B) OPTIMIZED βP AND LN; © 2005 ELSEVIER.

must be high, eliminating the possibility of some standard radio components such as quartz crystals.

To address these stringent requirements, a new and emerging technology was investigated: RF-MEMS (RF Micro-Electro Mechanical Systems) resonators. Recently, much progress has been made on GHz-range MEMS resonators, typically for use in bandpass filters and duplexers [10][11]. Since these resonators exhibit quality (Q) factors greater than 1000, these devices have the potential to facilitate the design of low power RF transceivers. The presence of an RF frequency reference can eliminate the need for quartz crystals in the system, greatly increasing the level of integration while decreasing the cost. For this implementation, Agilent Technologies' Thin-Film Bulk Acoustic Wave (FBAR) resonators [11], which exhibit resonant frequencies at approximately 1.9GHz, were used,. This relatively high carrier frequency allows the use of small chip antennas, further increasing system integration [12].

4.1 Transmitter

In a typical sensor network, the transmitter sends out sporadic bursts of short data packets to neighboring sensor nodes (< 10m). For a receiver sensitivity of –60dBm and indoor multi-path fading conditions, the application of the Friis wave propagation equation [14] shows that a transmit power of about

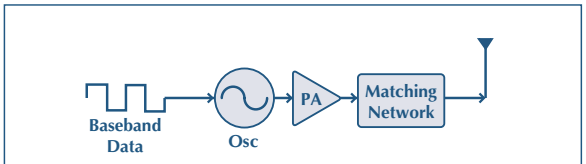

FIGURE 8: DIRECT MODULATION TRANSMITTER ARCHITECTURE, © 2004 IEEE.

0dBm is required. The transmitter must exhibit fast turn-on time and high efficiency. The transmitter architecture shown in Fig. 8 is well-suited for these requirements. The MEMS-based oscillator is directly modulated by the baseband data, and the power amplifier efficiently boosts the RF signal power. Direct modulation eliminates power hungry mixers and PLLs. Multiple channels can be implemented by tuning the oscillation frequency or by adding oscillators/ transmit chains in parallel.

In this implementation, carrier generation is accomplished with an oscillator co-designed with an FBAR resonator [8]. This architecture provides a low power 1.9GHz frequency reference without a PLL and quartz crystal reference. In addition, the start-up time is approximately 1μs, enabling the oscillator to be power cycled between transmitted bits, thus increasing the overall efficiency of the transmitter.

The schematic of the low-power amplifier is shown in Fig. 9. Maximizing the drain efficiency requires the voltage swing Vds to

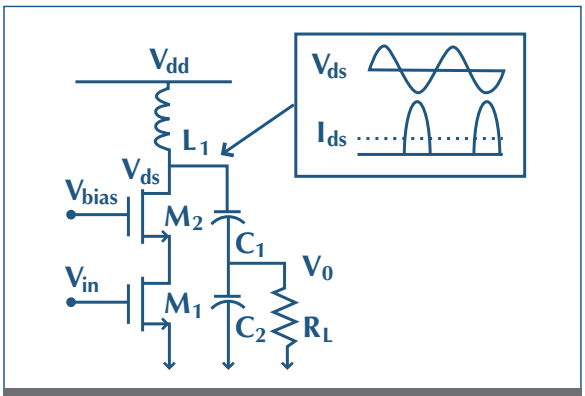

FIGURE 9: SCHEMATIC OF THE LOW POWER AMPLIFIER, © 2004 IEEE.

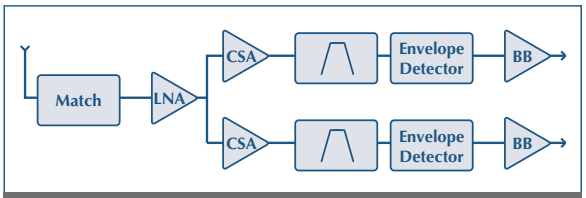

FIGURE 10: RECEIVER ARCHITECTURE, © 2004 IEEE.

be maximized. For 1.2mW output power with a 1.2V supply and 100mV knee voltage, a 500Ω load resistance at the drain of M2 is required. To achieve this, the 50Ω antenna (RL) is transformed to the required impedance using C1 and C2. Inductor L1 tunes out the transformed capacitance and the parasitic drain capacitance. Capacitive transformers are preferred over LC matching networks or inductive transformers because on-chip capacitors have much higher Q (>50) than on-chip inductors (Q of 5–10), resulting in less loss. In the current prototype, the required L1 is 1.2nH and is implemented using an off-chip inductor. However, a short bond wire or an on-chip inductor can be used for a fully integrated solution. A cascode transistor M2 ensures that the drain voltage does not exceed the low gate breakdown voltage for deep sub-micron CMOS. Cascoding increases the isolation between the input and output and improves the efficiency by boosting the drain resistance of M2.

4.2 Receiver

The receiver block diagram is shown in Fig. 10. The two-channel tuned radio frequency (TRF) architecture was chosen to demonstrate the effectiveness of RF-MEMS resonators in low power transceivers. The antenna feeds a 50Ω impedance presented by the LNA. The LNA drives a tuned load consisting of two channel select amplifiers (CSAs). Although two channels were used in this implementation, this architecture is scalable to larger numbers of channels. Each CSA is tuned by an FBAR resonator, which performs receiver channel selection. The CSAs drive an envelope detector, which acts as a self-mixer to perform signal downconversion. Baseband buffers are included to drive test instrumentation. Since the frequency stabilization is performed entirely by the MEMS resonators, no quartz crystals are used in this receiver architecture. The absence of a phase-locked loop (PLL) ensures a much faster receiver start-up time than that of a traditional radio.

The two-channel embodiment displays flexibility in terms of modulation schemes: the receiver can detect two unique OOK data streams at two carrier frequencies or it can detect FSK. For dense wireless sensor networks, it is anticipated that two separate OOK channels will be used, with one reserved for beaconing. Changing between these two modulation schemes can be accomplished with no receiver modifications and can be performed dynamically in either the analog or digital baseband detection circuitry.

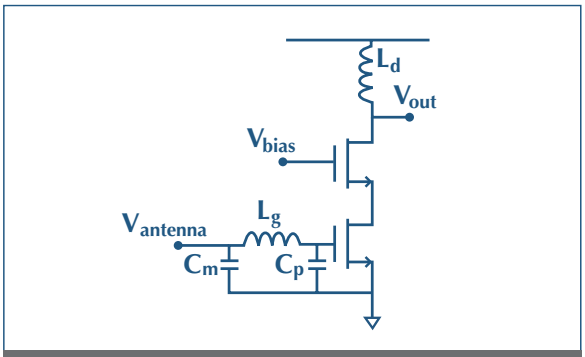

FIGURE 11: LNA SCHEMATIC, © 2004 IEEE.

The low noise amplifier must achieve high gain and moderate noise figure while simultaneously providing input impedance matching. In order to satisfy these requirements, the architecture

shown in Fig. 11 was implemented. The nonquasi static (NQS) gate resistance of the input transistor was transformed via a passive matching network to match the 50Ω source resistance. The NQS gate resistance is inversely proportional to the device transconductance and is especially pronounced at low current levels [13]. An off-chip matching inductor (Lg) was used for maximum flexibility in the prototyping phase, but it could easily be integrated on-chip. The parasitic capacitance Cp results from ESD diodes, bond pads, and board and layout traces. For a given bias current and fixed Cp, an optimal device size exists that maximizes the cut-off frequency of the transistor (ft).

The output tank inductor was implemented on chip by shunting the two top-level metal layers in order to increase the quality factor. It was characterized using a 3D electromagnetic solver and experimentally verified.

The channel select amplifier (Fig. 12) provides high RF gain to overcome the high noise figure of the detector as well as to interface the electrical signal with the acoustic resonator in order to perform high-Q filtering of the signal. The amplifier must exhibit high RF gain with a low power consumption while limiting the extent to which the resonator is de-tuned.

The amplifier consists of an NMOS cascode transconductance stage and a tuned load. The amplifier input capacitance is absorbed into the LNA load. A cascode transconductor structure is used to increase reverse isolation, thereby ensuring amplifier stability. The tuned load consists of an FBAR resonator to perform channel selection and an inductive PMOS-RC structure, which stabilizes the low-frequency bias point of the amplifier. The loaded quality factor of the tuned load is approximately 600, yielding an RF bandwidth of 3MHz.

The on-chip baseband circuitry downconverts the OOK or FSK modulated data to DC, performs filtering, and buffers the signal for driving subsequent instrumentation or circuitry. The first baseband block, shown in Fig. 13, is a non-linear low pass

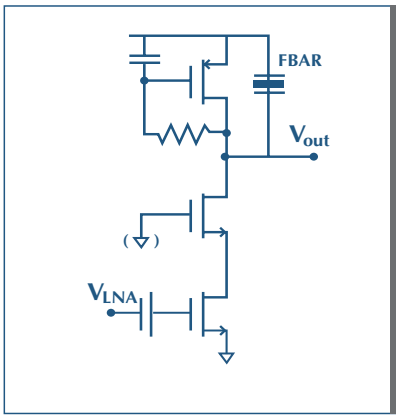

FIGURE 12: CHANNEL SELECT AMPLIFIER SCHEMATIC, © 2004 IEEE.

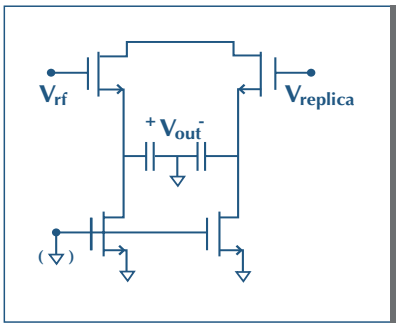

FIGURE 13: ENVELOPE DETECTOR SCHEMATIC, © 2004 IEEE.

filter, or envelope detector. This circuit, composed of MOSFETs biased in the deep subthreshold regime, performs a self-mixing operation on the RF signal that has been filtered by the channel-select amplifiers.

The low-pass filter has a 300kHz cutoff frequency that attenuates all fundamental tones passing through the receive chain. A replica envelope detector provides a reference DC level, producing a pseudo-differential baseband output. The current consumption of each detector is 200nA.

Each channel in the receive chain also contains a low power buffer to drive off-chip instrumentation. The buffers are capable of driving a 20pF off-chip load while consuming 50μA. On-chip threshold-referenced bias circuits provide a moderate level of power-supply independence.

5. WAKE-UP RECEIVER

In order to enable fully asynchronous communication between nodes, even under heavy duty cycling, each transceiver should be able to monitor the channel continuously and wake up only when data is being received. To enable this behavior, a carrier sense receiver is needed to monitor the channel for activity. Since this receiver is always listening, its power consumption should be on the order of 100μW.

A digitally controlled oscillator to be used as the frequency synthesizer of such a carrier sense receiver has been designed. By employing digital tuning, the oscillator may be periodically calibrated to an accurate reference such as a crystal or MEMS-based oscillator [17]. The calibration aligns the oscillator with the system reference, performing channel selection and tuning for the carrier sense receiver. Between calibration cycles, the accompanying tuning loop circuitry is turned off to save power, and the oscillator runs free until the next tuning cycle. Re-calibration could occur each time the main transceiver turns on to send and receive data.

The frequency tuning specifications for the digitally controlled oscillator (DCO) are derived from the intended application with periodic calibration, where the MEMS reference tolerance is 500ppm [17]. Accordingly, the target frequency resolution is set at 500kHz, which translates to approximately 1fF of switched capacitance at 1.9GHz. In addition, the total tuning range is chosen to be 200MHz (10%) in order to cover uncertainty in the tank inductance value and other process variations. In order to minimize the DCO power consumption, two major techniques have been utilized. At the circuit level, the DCO core transistors are designed to run in weak inversion, where more device transconductance is available for a given bias current. The circuit is also designed to run on a supply voltage well below the technology's nominal value to further minimize the power consumption. Recently, there has been interest in operating synthesizer VCOs at low supply voltages. In [18], transformers were used to enable the VCO internal nodes to swing above Vdd and below ground to maximize voltage headroom. In the design presented here, subthreshold device operation simultaneously enables low current and low voltage operation while still achieving the desired performance.

The use of subthreshold transistor biasing has been successfully employed in low power analog circuits for many years in order to take advantage of the high transconductance efficiency available in this regime [19]. Unfortunately, weak inversion operation usually results in large devices with substantial parasitics, so this technique has traditionally not been attractive for RF designs.

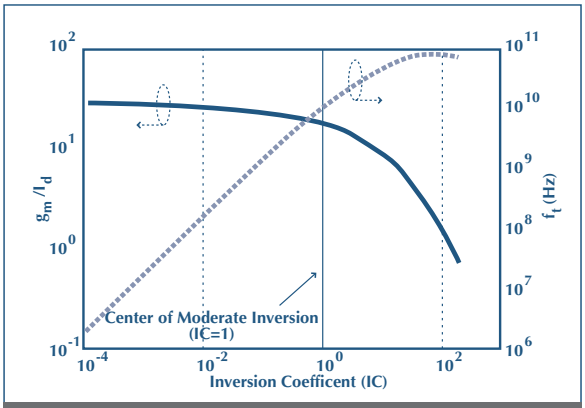

FIGURE 14: TRANSCONDUCTANCE EFFICIENCY AND FT FOR 0.13 μM CMOS, © 2005 IEEE.

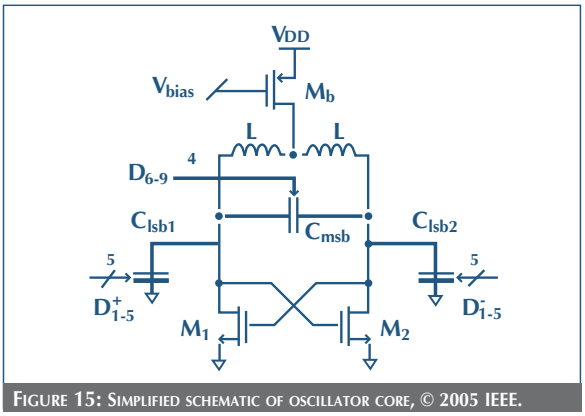

FIGURE 15: SIMPLIFIED SCHEMATIC OF OSCILLATOR CORE, © 2005 IEEE.

The conflicting trends of transconductance efficiency and device fT are illustrated in Fig. 14, where simulated values of

these parameters are plotted across the entire range of device operation for a 0.13λm technology. The parameter IC represents normalized current density with IC=1 indicating the center of moderate inversion. In this scaled technology, peak ft is close to 100GHz, meaning that transistors no longer need to be biased for optimal ft when low power consumption is one of the chief design goals. The weak inversion regime (IC ≈ 0.1) may then be used to take advantage of high transconductance efficiency, which may be more than a factor of two higher than in strong inversion.

The simplified schematic of the oscillator core is shown in Fig. 15. In order to reduce power consumption, the oscillator is designed to operate on a 0.5V supply, exploiting the trend towards low supply voltages in scaled CMOS technologies. A differential NMOS-only topology is adopted because it provides higher output swing for low supply voltages. With only two stacked transistors, the chosen oscillator topology is able to operate on very low supply voltages. Although the low supply reduces the maximum possible voltage swing (2Vdd), this is not a large disadvantage because the ultra low power oscillator is designed to run in the current-limited regime. Devices M1 and M2 are sized with large W/L =300λm/0.13λm for a design bias current of 100λA in each device. At this bias point, the device inversion coefficient is approximately 0.1, providing gm/Id ≈ 25, while still yielding manageable parasitics. An additional benefit of subthreshold operation for this design is that Vdsat and Vgs are both much lower than they would be in strong inversion, which conserves valuable voltage headroom. The current source PMOS transistor Mb provides bias current to the core and is sized to operate in saturation for supply voltages of 0.5V and below.

6. PLANAR INVERTED-F ANTENNA (PIFA)

The antenna of the Quark node is a planar inverted-F antenna (PIFA). The PIFA structure is low-profile, reliable, cheap, and has a nearly omni-directional radiation pattern. This type of antenna is printed directly on the circuit board during the copper etching process. The circuit board area for the antenna is small and therefore adds minimal cost to the node. Like most single ended antennas, the PIFA needs a ground plane to which it can couple, in this case, the PCB used for the rest of the node).

In a wireless network, a transceiver must be able to receive and transmit a signal in any direction because it is not aware of the location of other nodes. Because the PIFA is a type of a monopole antenna, it has a relatively omni-directional (donut shaped) radiation pattern that enables it to communicate with other PIFA.

Additionally, the PIFA avoids the use of inefficient matching networks. In low-voltage IC design, the efficiency of a power amplifier increases with output resistance typically above 50Ω. Most of-the-shelf antennas have a radiation resistance of 50Ω, which requires the use of a matching network. Matching networks comprised of discrete parts are expensive, and on-chip matching networks are inefficient. A PIFA, however, allows the designer to choose almost any radiation resistance (preferably between 50Ω to 500Ω) by simply changing the geometrical dimensions. Optimization of the geometrical dimensions is performed using electromagnetic optimization software.

Although the PIFA is a narrowband antenna, it still provides more than 250MHz of bandwidth at 1.9GHz with a Voltage Standing Wave Ratio (VSWR) < 2. This is more than sufficient for the low data rate, narrowband communication of a PicoRadio wireless network.

7. POWER TRAIN

The PicoRadio node must be self-powered since physical access to nodes in a deployed network may be difficult and the sheer number of nodes makes changing batteries or manual refueling unrealistic. At the present level of integration for Quark, the best option for powering the board is solar scavenging augmented by rechargeable batteries. Solar scavenging provides more energy per unit area than most alternatives, and the cells are readily available commercially. For the Quark node, there are three main challenges: utilizing solar power with strict array size

restrictions, storing energy, and regulating voltage to the Quark circuitry. Since minimizing board size is a primary goal, any viable solution must be small in terms of area or volume.

Obtaining sufficient power from solar energy is the first challenge. Solar panels are characterized by open circuit voltage and short circuit current in full sunlight. However, most nodes will be deployed indoors where the cell efficiency drops off dramatically under fluorescent lighting. For instance, a (2x2.5)cm^2 single crystal cell has an open circuit voltage of 530mV and a short circuit current of 80mA when measured outside on a typical day. Under standard fluorescent lighting, however, this same cell yields half the voltage and a small fraction of the current 240mV, 360μA, respectively. To determine how much power is available between open and short circuit conditions, a series of measurements were taken under fluorescent lighting conditions. Peak power for a single crystal cell is 32μW (125mV, 250μA), but, when three additional cells are wired in series, power does not quadruple as might be expected. Instead, peak power is only 2.3 times the single cell case (74μW). For this test, the combined open circuit voltage is 700mV, and the short circuit current is 290μA.

Due to the use of several off-the-shelf components, the power budget of the current Quark version is in the single-digit milliwatt range. An array sufficient to power a milliwatt device or to produce supply voltages directly usable by the Quark circuitry would be unacceptably large. Thus, solar power is insufficient to power the node in typical indoor conditions.

The second challenge is the efficient storage of energy in a rechargeable battery. While there are multiple efforts to develop thin film battery technologies, only commercially available parts were considered for this design. Coin cells are desirable from a size standpoint but are typically primary (non-rechargeable) technologies. Lithium-ion coin cells that are attractive for their nominal voltage (3.6V)are available but have complex charging requirements. Nickel Metal Hydride (NiMH) is also available in a variety of capacities. These batteries can be trickle-charged at about 1/10 capacity without damage but only produce 1.2V. Therefore, the choice of NiMH requires a series connection or boost regulation to reach the voltage requirement.

The third power train challenge is the conversion of the battery voltage into DC levels appropriate for the Quark circuitry. The original Quark design called for three supply voltages: 1.8V for the Charm I/O interface and most support chips, 1.2V for the Strange chip, and 1.0V for the core Charm logic. There are a number of tradeoffs between nominal battery voltage, battery charge/discharge characteristics, minimum input voltage for a regulator chip, efficiency of regulation, and the complexity and size of the circuit.

There are three generally available methods for regulation: switched capacitor charge pump, linear regulation, and switching DC-DC regulation. The charge pump is the simplest in terms of external components but is limited mainly to unregulated voltage inverters or doublers. Linear regulators are also relatively simple but efficiency is related to V_{out}/V_{in}. Switching regulators (buck converters) are most efficient within their design output current range, but the circuit board layout is relatively large, and the efficiency drops sharply at low current levels. Unfortunately, there are no existing off-the-shelf power supply components specifically designed for a low milliwatt assembly with severely limited available power.

One way to overcome input voltage minimums (enabling use of a single 1.2V battery) is first to increase the input voltage using a switching boost regulator. However, efficiency losses are multiplicative, so if the boost stage and step-down stages are each 80% efficient, the combined efficiency will be 64%. In practice, these numbers are optimistic, as a test circuit with a boost-buck topology that converted 1V to 2.75V to 1.2V was less than 50% efficient at 1mA.

Another option is to tolerate an unregulated supply and source power directly from the battery. This is reasonable for some

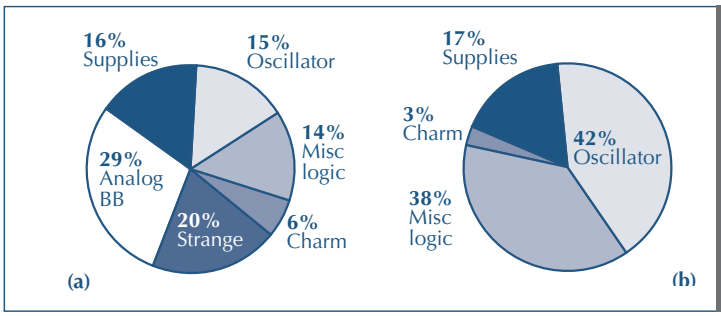

FIGURE 16: NODE POWER CONSUMPTION BREAKDOWN IN RX (A) AND WITH RADIO OFF (B), © 2004 IEEE.

devices; for instance, the custom Strange and Charm chips are tolerant to supply variations. Variation on the battery voltage, however, can violate the V_{in} specifications for some of the commercial parts driven by the Charm, which usually specify a supply voltage above the 1.2V nominal of a single NiMH cell. Two NiMH cells in series raise the battery voltage to 2.4V, which is higher than nominal for devices on the Quark board but generally within operating range.

Given the above analysis, a solar module with five single-crystal cells in series and two 15mAh NiMH coin batteries for power collection and storage was selected. The module produces sufficient voltage in full or partial sun to charge the batteries at an acceptable rate, and, with the addition of a voltage doubling charge pump, can charge at lower light intensities (cloudy days or indoors) or at a higher rate. The Charm core and the Strange radio are powered directly from a single NiMH battery. The result is a power train that will perform with reasonable efficiency and recharge the batteries when there is enough available light.

8. CLOCK

The Quark node uses three primary clocks: a digital logic system clock (16MHz), a sampling baseband clock (1MHz), and an idle mode clock (160kHz). Currently, the three clocks are derived from a single external crystal oscillator running at 16MHz. The accuracy of the idle and system clocks are not tightly constrained, but the accuracy of the baseband clock must be tight enough to hold synchronization during the reception of a complete packet (currently 50ppm). Derivation of the baseband clock from the main clock thus overconstrains the accuracy of the other two clocks. If this accuracy requirement is relaxed, alternative lower-power clock generation solutions such as an integrated micro electro-mechanical systems (MEMS) oscillator or a simple free-running inverter ring can be considered.

The baseband clock may still require an external crystal to achieve the accuracy requirements, but it could be turned off entirely when the radio is not in use. This same idea can be extended to disable the system clock, since only the low speed idle clock is required during idle mode. Although we currently save significant power in the Charm by clock gating, separating the clocks into lower-power sources and controlling them individually will reduce total power.

9. POWER CONSUMPTION BREAKDOWN

Clearly, further reduction of the node's power requirements will increase achievable duty cycles. Once the power requirements are reduced below the amount provided by the solar cells, the number or size of the cells can be reduced.

In Fig. 16, the power consumption breakdown of the entire node during reception (a) and idle mode (b) is shown. An interesting observation is that the custom chips together consume only about 25% of total power when receiving (transmit is comparable) and only 3% with the radio off. The four largest power sinks are the inefficient voltage regulators, the external clock oscillator, miscellaneous external logic, and the external analog components in the baseband block.

The commercial voltage regulators are inefficient because they are not designed for low-voltage, low-current operation. Custom power regulation circuitry would address this problem.

All of the miscellaneous logic can be integrated into either the Charm or Strange chip, where power consumption would be reduced by perhaps two orders of magnitude.

Lastly, almost 30% of the node power in Rx mode is dissipated in the analog baseband circuitry. Although disabled when the node is not receiving, it is still a significant drain on the power train when operating. In addition, it consumes nearly 10% of the board area. Preliminary work shows that integrating this circuitry into the RF chip will reduce power consumption by a factor of twenty.

10. CONCLUSION

This article summarizes the research and design efforts of the BWRC towards the realization of the PicoRadio vision. The confluence of progress made in the fields of protocol stack algorithm design, energy train, node locationing, digital processing, and RF communication circuitry has been crucial to the results obtained. These technologies were integrated into a self-contained energy scavenging sensor node (Quark Node). Two custom CMOS chips (manufactured by STMicroelectronics) have been designed and implemented for the Quark node, and a custom PCB with integrated PIFA antenna has been developed to reduce the size of the node.

The main contributions that the system-level integration of the Quark node has given to the field of WSNs are a custom protocol stack using energy-efficient algorithms implemented mostly in low-power custom ASICs and an energy train designed to allow the node to be self-contained with batteries recharged by solar energy. Future versions of the node would decrease the node size and increase the duty cycle by integrating custom baseband circuitry, custom circuits for the power train, and MEMS-based clock references.

A third chip in 0.13 m CMOS (STMicroelectronics) implements a 100μW digitally controlled oscillator suitable for reactive receivers. Bondwire inductors are employed for high Q, and no external components are used. Subthreshold device operation is shown to be a feasible method of increasing transconductance efficiency at RF.

The ever increasing problem of SRAM power leakage has been extensively addressed by exploring the limit of SRAM data preservation under ultra-low standby VDD. A commercial SRAM module (from STMicroelectronics) with high-Vth process is shown to be capable of sub-400 mV standby data preservation. With additional 100 mV guard band to account for power supply ripple and cosmic particles, leakage power savings of more than 85% can be achieved with an SRAM module under 490 mV standby VDD, compared to 1V active mode. The DRV is observed to be a strong function of process variation and also SRAM cell sizing. With proper sizing optimization, an additional 70% leakage power savings can be achieved with only a 30% SRAM cell transistor area increase. Besides sizing, more variables are being investigated for their impacts on ULP SRAM cell design. Such variables include the transistor Vth and body bias voltages. With the control of body bias voltages, the SRAM DRV and leakage current can be dynamically adjusted in different operation modes.

REFERENCES

[1] Reprinted from Huifang Qin, Yu Cao, Dejan Markovic, Andrei Vladimirescu, Jan Rabaey, "STANDBY SUPPLY VOLTAGE MINIMIZATION FOR DEEP SUB-MICRON SRAM," Microelectronics Journal, Vol. 36, 2005, pp. 789-800, Elsevier, with permission from Elsevier.

[2] Nathan M. Pletcher and Jan M. Rabaey , "A 100 μW, 1.9 GHZ OSCILLATOR WITH FULLY DIGITAL FREQUENCY TUNING," IEEE Proceedings of ESSCIRC, Grenoble, France, 2005.

[3] M. Sheets, B. Otis, F. Burgardt, J. Ammer, T. Karalar, P. Monat, and J. Rabaey, "A (6x3)cm2 Self-Contained Energy-Scavenging Wireless Sensor Network Node," IEEE Wireless Personal Multimedia Communications (WPMC), 2004, Abano Terme, Italy

[4] "Information Technology – Open Systems Interconnection – Basic Reference Model," International Organization for Standardization (ISO). Standard number ISO/IEC 7498-1:1994, 1994.

[5] C. Intanagonwiwat, R. Govindan and D. Estrin, "Directed Diffusion: A scalable and robust communication paradigm for sensor networks," IEEE/ACM Mobicom, 2000, pp. 56-67.

[6] R. Shah, J. Rabaey. "Energy aware routing for low energy ad hoc sensor networks." 2002 IEEE Wireless Communications and Networking Conference Record. WCNC 2002 (Cat. No.02TH8609). IEEE. Part vol.1, 2002, pp.350-5 vol.1. Piscataway, NJ, USA.

[7] C. Savarese, K. Langendoen, J. Rabaey, "Robust positioning algorithms for distributed ad-hoc wireless sensor networks," USENIX Technical Annual Conference, Monterey, CA, 2002.

[8] C. Savarese, J. Rabaey, J. Beutel, "Locationing in distributed ad-hoc wireless sensor networks," Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP), pp 2037-2040, Salt Lake City, UT, May 2001.

[9] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, J. Rabaey, "SRAM Leakage Suppression by Minimizing Standby Supply Voltage," ISQED 2004.

[10] B. Otis, Y.H. Chee, R. Lu, N. Pletcher, J. Rabaey, "An Ultra-Low Power MEMS-Based Two-Channel Transceiver for Wireless Sensor Networks," IEEE VLSI Circuits Symposium, 2004.

[11] B. Otis, J. Rabaey. "A 300μW 1.9GHz CMOS Oscillator Utilizing Micromachined Resonators," IEEE J. of Solid State Circuits, Vol 38, p. 1271-1274, July, 2003.

[12] J. Rabaey, J. Ammer, T. Karalar, S. Li, B. Otis, M. Sheets, T.Tuan, "PicoRadios for Wireless Sensor Networks: The Next Challenge in Ultra-Low Power Design," IEEE ISSCC, pp. 200-1, Feb 2002.

[13] B. Bircumshaw et. al, "The Radial Bulk Annular Resonator:Towards a 50Ohm RF MEMS Filter," Tech. Dig., 12th Int. Conf. on Solid State Sensors, Actuators and Microsystems, Boston, pp. 875-878, June 8-12, 2003.

[14] R. Ruby et. al, "Ultra-miniature high-Q filters and duplexers using FBAR technology," IEEE ISSCC, pp.120-1, Feb. 2001.

[15] S. Roundy, B. Otis, Y.H. Chee, J. Rabaey, P. Wright, "A 1.9GHz RF Transmit Beacon using environmentally Scavenged Energy," Dig. IEEE Int. Symp. on Low Power Elec. And Devices, Seoul, Korea, 2003.

[16] Y. Tsividis. **Operation and Modeling of the MOS Transistor**. McGraw-Hill, Boston, 1998.

[17] C. A. Balanis, **Antenna Theory – Analysis and Design**, 2nd ed. New York, Wiley, 1997.

[18] P. Choi, et. Al, "An Experimental Coin-sized Radio for Extremely Low Power WPAN (802.15.4) Application at 2.4 GHz", IEEE ISSCC, pp. 92-3, Feb 2003.

[19] T. Melly, et. al, "An Ultralow Power UHF Transceiver Integrated in a Standard Digital CMOS Process: Transmitter", IEEE J. Solid State Circuits, vol. 36, no. 3, Mar 2001.

[20] B. Otis, Y. Chee, and J. Rabaey, "A 400λW-RX, 1.6mW-TX super-regenerative transceiver

FOR WIRELESS SENSOR NETWORKS," in IEEE ISSCC Digest of Technical Papers, Feb. 2005, pp. 396–397.

[21] K. Kwok and H. Luong, "ULTRA-LOW-VOLTAGE HIGH PERFORMANCE CMOS VCOS USING TRANSFORMER FEEDBACK," IEEE J. Solid-State Circuits, vol. 40, no. 3, Mar. 2005.

[22] E. Vittoz and J. Fellrath, "CMOS ANALOG INTEGRATED CIRCUITS BASED ON WEAK INVERSION OPERATION," IEEE J.Solid-State Circuits, vol. SC-12, pp. 224–231, June 1977.

[23] R. Staszewski, C.-M. Hung, D. Leipold, and P. Balsara, "A FIRST MULTIGIGAHERTZ DIGITALLY CONTROLLED OSCILLATOR FOR WIRELESS APPLICATIONS," IEEE Trans. Microwave Theory Tech., vol. 51, no. 11, Nov. 2003.

[24] J. Craninckx and M. Steyaert, "A 1.8-GHz CMOS LOW-PHASE-NOISE VOLTAGE-CONTROLLED OSCILLATOR WITH PRESCALER," IEEE J. Solid-State Circuits, vol. 30, no. 12, Dec. 1995.

[25] S. Borkar, Design challenges of technology scaling, IEEE Micro 19 (4) (1999) 23–29.

[26] S. Manne, A. Klauser, D. Grunwald, "PIPELINE GATING: SPECULATION CONTROL FOR ENERGY REDUCTION," International Symposium Computer Architecture (1998) 132–141.

[27] N. Kim, "DROWSY INSTRUCTION CACHES: LEAKAGE POWER REDUCTION USING DYNAMIC VOLTAGE SCALING AND CACHE SUB-BANK PREDICTION," Proceedings of the 35th Annual Int'l Symposium Microarchitecture (MICRO-35), IEEE CS Press, 2002. pp. 219–230.

[28] M. Horiguchi, T. Sakata, K. Itoh, "SWITCHED-SOURCE-IMPEDANCE CMOS CIRCUIT FOR LOW STANDBY SUBTHRESHOLD CURRENT GIGA-SCALE LSI'S," IEEE Journal of Solid-State Circuits 28 (11) (1993) 1131–1135.

[29] K. Itoh, "LOW VOLTAGE MEMORIES FOR POWER-AWARE SYSTEMS," Proceedings of the ISLPED (2002) 1–6.

[30] H. Mizuno, T. Nagano, "DRIVING SOURCE-LINE (DSL) CELL ARCHITECTURE FOR SUB-1-V HIGH-SPEED LOW-POWER APPLICATIONS," Digest of technical papers. Symposium on VLSI circuits (1995) 25–26.

[31] K. Itoh, A.R. Fridi, A. Bellaouar, M.I. Elmasry, "A DEEP SUB-V, SINGLE POWER-SUPPLY. SRAM CELL WITH MULTI-VT, BOOSTED STORAGE NODE AND DYNAMIC LOAD," Digest of technical papers. Symposium on VLSI circuits (1996) 132–133.

[32] H. Kawaguchi, et al., "DYNAMIC LEAKAGE CUT-OFF SCHEME FOR LOW VOLTAGE SRAM'S," Digest of technical papers. Symposium on VLSI circuits (1998) 140–141.

[33] S. Kaxiras, Z. Hu, M. Martonosi, "CACHE DECAY: EXPLOITING GENERATIONAL BEHAVIOR TO REDUCE CACHE LEAKAGE POWER," Proceedings of the ISCA (2001) 240–251.

[34] E. Seevinck, F.J. List, J. Lohstroh, "STATIC-NOISE MARGIN ANALYSIS OF MOS SRAM CELLS," IEEE Journal of Solid-State Circuits SC-22 (5) (1987) 748–754.

[35] J. Rabaey, A. Chandrakasan, B. Nikolic, "DIGITAL INTEGRATED CIRCUITS: A DESIGN PERSPECTIVE," second ed., Prentice-Hall, 2002.

[36] J. Lohstroh, E. Seevinck, J.D. Groot, "WORST-CASE STATIC NOISE MARGIN CRITERIA FOR LOGIC CIRCUITS AND THEIR MATHEMATICAL EQUIVALENCE," IEEE Journal of Solid-State Circuits SC-18 (6) (1983) 803–807.

[37] C. Lage, et al., "SOFT ERROR RATE AND STORED CHARGE REQUIREMENTS IN ADVANCED HIGH-DENSITY SRAMS," Proceedings of IEDM (1993) 821–824.

[38] M.J. Ammer, et al., "A LOW-ENERGY CHIP-SET FOR WIRELESS INTERCOM," Proceedings of DAC (2003).

[39] K.D.T. Ngo, R. Webster, "STEADY-STATE ANALYSIS AND DESIGN OF A SWITCHED-CAPACITOR DC-DC CONVERTER," Proceedings of PESC (1992) 378–385.

[40] Y. Cao, T. Sato, D. Sylvester,M. Orchansky, C. Hu, "NEW PARADIGM OF PREDICTIVE MOSFET AND INTERCONNECT MODELING FOR EARLY CIRCUIT DESIGN," Proceedings of CICC (2000) 201–204.

"CUPOLA-SENSING": AN INDOOR CLIMATE MONITORING TEST CASE

Max Cortiana,
Andrea Labombarda,
Laura Vanzago

STMicroelectronics

In this paper, we describe a test case for indoor climate monitoring deployed in our laboratory. The test case represents a significant class of Wireless Sensor Network (WSN) monitoring applications. It was developed to practice with data gathering and in-network reprogramming and to identify the basic requirements for interfacing a WSN within Internet to enable network analysis. The system architecture includes a WSN physically implemented with COTS (Components Off The Shelf), a gateway device connecting the network to Internet, and client tools running on PCs connected via Ethernet to the gateway. The client tools allow a user to observe the WSN behavior at run time for network debugging and to analyze the sensing data stored in a database integrated in the gateway. The SW platform running on sensor nodes is built on top of TinyOS, a widely used open source operating system specifically designed for sensor network applications.

1. INTRODUCTION

A large number of Wireless Sensor Networks have been proposed in many different application scenarios, from environmental monitoring to asset tracking to health monitoring. WSN applications can be grouped into three main categories: monitoring, tracking, and control, differing in terms of system requirements such as lifetime, cost, ease of deployment, and response time constraints. For example, in most control scenarios, real-timeness is a fundamental constraint, while, in general, it is not a significant requirement for the monitoring applications. Therefore, different protocols and architecture solutions can be defined to optimize the behavior of a WSN in each category.

Environmental monitoring applications are widely accepted as a common driver for WSN research. A significant number of deployments have been documented in the literature, describing specifications and constraints for several indoor and outdoor scenarios [1], [2], [3].

In general, environmental monitoring applications are characterized by long lifetime requirements, low data rates, and relatively static topology features. Environmental data collection applications typically use hierarchical - often tree-based - routing topologies, where each data packet is rooted at

high-capability nodes storing sensed data (sink). With a tree-based data collection, each node is responsible for forwarding the data to all its descendants. Once the network is configured, data is periodically transmitted from child nodes to parent nodes up the tree structure until it reaches the sink. For many scenarios, the interval between these transmissions can be on the order of minutes. In addition to large sample intervals, environmental monitoring does not have strict latency requirements, as, in general, data is collected for future analysis and not for real-time operation. In order to meet lifetime requirements, each communication event is often required to be precisely scheduled, allowing the sensor nodes to remain in sleep mode most of the time, waking up only to transmit or receive data. If the precise schedule is not met, the communication event will fail, highlighting the role of distributed time synchronization as an important service to be implemented in this type of network [19].

Besides having the capability to collect data, an additional advantage of a WSN is its ability to operate for extended periods of time without human intervention. In addition, many environmental monitoring applications make physical interaction with the nodes for maintenance purposes unacceptable because nodes can only be accessed with difficulty or because, within the network, the nodes can be very numerous. End users must be able to add or change the functionalities of a deployed network remotely to fully utilize its capabilities and to dynamically adapt to different environmental conditions or application requirements. Therefore, in-network reprogramming is a key feature for the success of a WSN and, due to limited node resources, a very active research field spanning innovative middleware and operating system capabilities, high-level programming, and abstraction models for heterogeneous distributed systems. Some approaches that have been investigated consist of injecting the network with queries that change the working state of the node [4], with a binary image of the entire program [5], or with a script requiring a virtual machine within the software platform of the node [6].

Implementing both data gathering and in-network system functionalities for indoor climate observation, the test case we designed represents the class of monitoring applications introduced above. The high-level application is indoor climate observation. The main objective of the test case was to experiment with those system functionalities by means of a real deployment. We also outlined problems of WSN integration within Internet and WSN network analysis by means of client utilities running on any PC connected on the Internet network.

Therefore, the test case includes:

- a WSN that uses COTS, namely TelosSky [7] motes platforms, while running TinyOS [8] operating system; we used TelosSky because of its solid representation of sensor nodes capabilities as well as the support offered by the TinyOS community at the time we started the test case development;
- a set of SW components conceptually referred to as gateway services. They include a database which stores sensed data, a web server, and Java applications which interact with the database and create the packets used to reprogram the operational conditions of the nodes remotely; and
- client tools which analyze the data stored within the database and which interact with the network by observing its behavior and topology at run time or which select the parameters of in-network reprogramming commands.

Our design choices have been strongly influenced by the additional objective of using the test case to identify a fast prototyping framework to quickly evaluate performances of new HW platforms or components in WSN applications. To reach this goal, we decided to base our developments on open source libraries and free design environments. The advantages of open source solutions are well known. The most notable of them fitting our needs is the fact that an open source enables reuse, improvements, and unlimited tuning. This makes it possible to port the code to new hardware, adapt it to changing conditions, and reach a detailed understanding of how the system works, taking full advantage of previously developed solutions.

Therefore, our test case is based on the adoption of:

- a TinyOS library environment;
- a Linux machine working as gateway server;
- a database built on top of MySQL [10]; and
- an Apache web server [11].

For implementation, we used NesC [9] for node applications, PHP [12] for dynamic web pages, and java for gateway and graphical utilities.

The paper is organized as follows: in section 2, we introduce the TinyOS framework which constitutes the heart of our SW platform. In section 3, we describe the test case architecture by giving implementation details of its main components: the sensor network, gateway services, and client utilities. In section 4, we discuss some limits of the test case and we introduce recent developments. Section 5 is dedicated to conclusions.

2. TINYOS

An extended class of nodes typically used in many WSN applications is characterized by limited HW resources [13]. Therefore, an operating system framework is needed to manage those resources efficiently while supporting concurrency-intensive operation in a manner that achieves efficient modularity and SW robustness [14].

Popular embedded operating systems do not meet the size, power, or efficiency requirements. Therefore, several operating systems have been developed specifically for tiny networked devices [15],[16],[17], to be used within WSN applications.

We decided to select TinyOS, the most adopted open source operating system in the WSN research community, an operating system which is very well supported through a mailing list and a working group which has joined development efforts from both industry and academia. In our implementation, we used version 1.1.15 of the TinyOS package including an extended SW library for sensor nodes, utilities implementing gateway, and client services. Addressing the goal of a fast prototyping framework design, we were confident the popularity of TinyOS would enable us to reuse a large number of components already available. Moreover, the involvement of leading research groups in the field of WSN, would also allow our reusing of components representing the state of the art in MAC and networking protocols, components we were not interested in developing further.

TinyOS was designed by targeting Wireless Sensor Network requirements. It implements an event-driven execution paradigm where every execution is triggered by some external event representing hardware interrupts coming from the radio, timers, or sensor interfaces. TinyOS provides physical device abstractions such as a conventional operating system. To be portable on architectures with very limited resources, the TinyOS approach defines a simple component model and develops a set of components that are statically wired together to support a specific application.

Therefore, a TinyOS application consists of a scheduler and a graph of components, each described by interfaces and internal implementation. An interface can declare commands and events. The first ones are implemented by the components that use that interface, while the second ones are implemented by the components providing it. Then, a wiring specification describes how the interfaces are wired together to form the overall application.

TinyOS takes a language-based application development approach, implemented in the NesC language, an extension of C language. The NesC compiler consists of a preprocessor translating a NesC application into a C module where only the necessary parts of the operating system are compiled with the application. The programming model exposed by NesC incorporates the event-driven execution, the flexible concurrency model, and the component-oriented application design of TinyOS. Restrictions on the programming model allow the NesC compiler to perform whole-program analyses, including data-race detection, which

improves the reliability of the SW by safely checking the concurrency introduced by the programmer.

3. CUPOLA-SENSING TEST CASE

In this section, we describe the main features of Cupola-Sensing, which we used to monitor the indoor climate within our laboratory for about three months. During this time, we validated and refined our design, and we collected a significant amount of data, which is now accessible through our web site. The test case name derives from the shape of the roof of the building where the laboratory is located.

At the system level, the test case consists of the multi-tiered architecture represented in Figure 1.

The first tier consists of a WSN built with COTS platforms running TinyOS and integrating a 802.15.4 radio, a processor, and sensors for climate (humidity, temperature and light) monitoring. The WSN implements two high level functionalities: data gathering and command broadcasting. Data gathering means that motes periodically send packets containing the values collected by the sensors and information about their working status (i.e., voltage level and parent node address). Data are routed through a multihop network to a node, called a "basestation", identified with a "0" network address, and physically connected to the gateway with a USB port. Commands Broadcasting is a kind of in-network reprogramming based on the usage of commands sent through the gateway to modify some variables characterizing the node working conditions.

The second tier is a Linux PC implementing gateway functionalities. The gateway allows us to interface the WSN with Internet, and it is responsible for collecting the data coming from sensors, parsing and storing them in a database. It is also responsible for creating the command packets broadcasted to the network.

The third tier includes WSN client tools that allow an end user or a network developer to interact with the WSN with a network visualizer, a web browser (Explorer), and DB clients. Client tools run on PCs linked with the gateway through a TCP/IP connection.

3.1 WSN

The WSN was constituted by 7 nodes deployed in the laboratory, as showed in Fig. 1. The rationale behind selecting the node position was to cover the lab with a reasonable granularity and to monitor "significant" areas such as entrance and meeting rooms to maximize the differences among values sensed by different nodes. Two nodes were always powered with batteries: one located at the main entrance door and one in a closet where

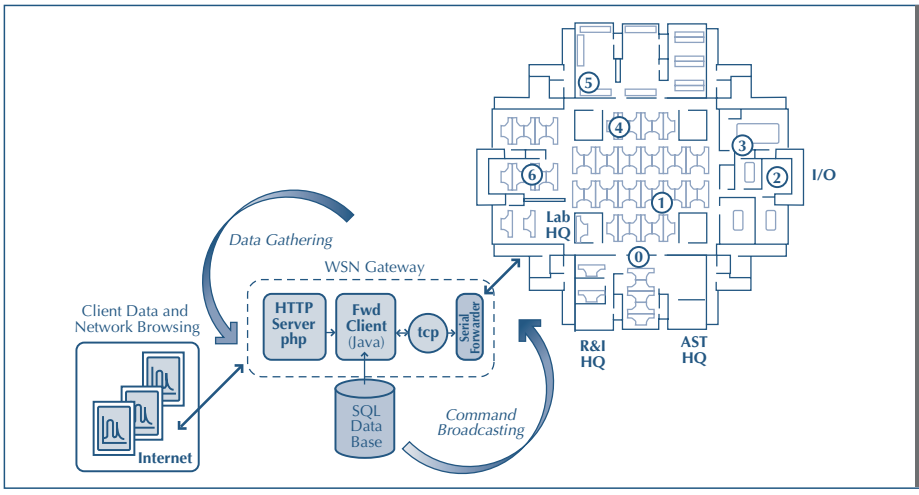

FIGURE 1:
CUPOLA SENSING TESTBED SYSTEM ARCHITECTURE.

power line sockets were not available. The other network nodes were connected to PCs in the lab through their USB ports, allowing them to be powered with batteries only when the PCs were switched off during the night or when the owner was absent. Therefore, our network replicated mixed power-supply conditions for a realistic indoor scenario by which to analyze distributed power management network requirements.

Fig. 2 outlines the main components of Cupola Sensing and their interfaces, where components are grouped in three layers. While we reused most of the Driver and Networking modules from the TinyOS package, the Cupola Sensing Component

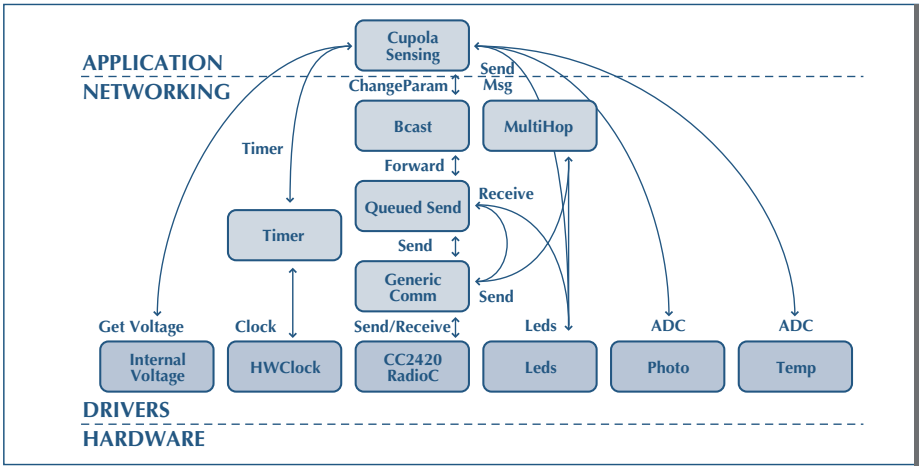

FIGURE 2:
SIMPLIFIED VIEW OF CUPOLA SENSING. BOXES REPRESENT COMPONENTS, AND EDGES REPRESENT INTERFACE WIRING.

and global wiring have been specifically developed for this test case. Cupola Sensing includes all the payload composition and command execution functions. The binary image of the full application (including drivers, networking and application components) was about 22 kBytes.

3.1.1 Nodes

To implement the WSN, we decided to use TelosSky. The following reasons motivated our choice.TelosSky well represents the popular MICA family originally defined by UC Berkeley [18] for WSN prototyping and it is well supported within the TinyOS development community. The architecture integrates a ChipCon radio implementing the IEEE 802.15.4 standard adopted by the Zigbee Alliance and extensively promoted in many WSN scenarios requiring low data rate communications. TelosSky also integrates both temperature and light sensors that we wanted to use for climate monitoring. The reprogramming of the on-board processor is efficiently done through a USB port that allows powering the motes, as an alternative to a battery-based solution.

3.1.2 Data Gathering

Data gathering implies an unidirectional data flow towards the basestation and includes the operations of periodic sampling of the node sensors, dispatching of packets to the gateway, and tree-based routing maintenance. The network protocol implements multihop routing policies, meaning that not every node directly communicates with the basestation, thus requiring the implementation of a routing mechanism on each node of the network.

For the implementation of the routing protocol, we reused a module (LQIMultiHopRouter) already integrated in the TinyOS library package. It allows for the building of the routing tree and the routing of messages towards the basestation connected to the gateway, which represents the root of the tree. The tree is built and maintained by means of beacon messages periodically

broadcasted by each node within the network. The beacon contains information related to its depth within the tree. When a node is powered on, it selects an initial parent by listening to messages and choosing the node corresponding to the smallest depth within the tree; to seed the creation of the tree, the basestation periodically broadcasts beacon messages with depth "0." Multiple networks can be active in the same physical spaces by setting different group IDs in a group field of radio Message Header.

Beacon packets fields are represented in Fig. 3. The cost field indicates the total cost to reach the root node and is calculated by adding the cost indicated within the beacon of the parent with the parent link quality derived from the channel estimation hook provided by the radio transceiver. The hopcount field represents the total number of hops needed to reach the basestation. The cost and hopcount fields are used by the routing algorithms to dynamically determine the best way to reach the root node and consequently the parent to be used. Thus, by allowing each node to change the current parent, the routing protocol network is robust against events that can create bad reception on the channel link or against events caused by possible parent failures.

The CupolaSensing application module on each node samples the sensors by using a programmable rate, builds the payload by using those data, and then sends the packet to the network layer that appends its multihop header. The MAC modules complete the packet before sending it to the transceiver by also adding the corresponding message header. The resulting packets and field details are represented in Fig. 3.

The radio module within the node board consumes 20mA in reception mode and 17mA in transmit mode. By keeping the radio always on, the batteries lasted only three days. Thus, to augment the lifetime of the node, we implemented a power management policy to reduce the radio duty cycle. Key consideration is that the radio can be switched off when communication is not occurring. By considering that potentially incoming packets were a) beacons, b) sensing data from other nodes, c) commands coming from the basestation, we empirically evaluated a minimum active timing window of about two minutes. For example, by considering a duty cycle of 13% (sampling data every 15 minutes) the lifetime of battery-supplied nodes was extended to 15 days. Of course, our policy implies

MsgHeader	8 bits	16 bits	8 bits	16 bits	16 bits	8 bits	8 bits
	Length	Frame Control	Dsn	Destpan	Addr	AM type	Group
MultiHop	16 bits	16 bits	16 bits	16 bits	16 bits		
	Source Addr	Origin Addr	Seqno	Origin SeqNumber	HopCount		
Beacon Payload	16 bits	16 bits	16 bits				
	Parent	Cost	HopCount				
Data Payload	16 bits	16 bits	16 bits	16 bits	16 bits	16 bits	32 bits
	Type	Light	Temperature	Humidity	Parent Address	Voltage	Sequence Number
Data Packet	MsgHeader	MultiHop	Data Payload				
Beacon Packet	MsgHeader	MultiHop	Beacon				

Figure 3: Data and Beacon packets.

that commands are sent from the gateway in a synchronous way. While this was not a limitation for the application, the need to align commands broadcasting with the network sampling rate was annoying for the WSN user. Hence, we implemented a function CommandSync in the gateway that simply delays the delivery of a command asynchronously requested by a user to meet the radio active time slot of the nodes.

and their payload fields are logically correlated with the state defined for each node established by:

- the LED status (RED, Green, Blue on/off);
- the sensing rate, representing the time interval between two successive sensor readings;
- the radio transmit power, ranging in the interval -25 dB and 0 dB as allowed by the radio component; and

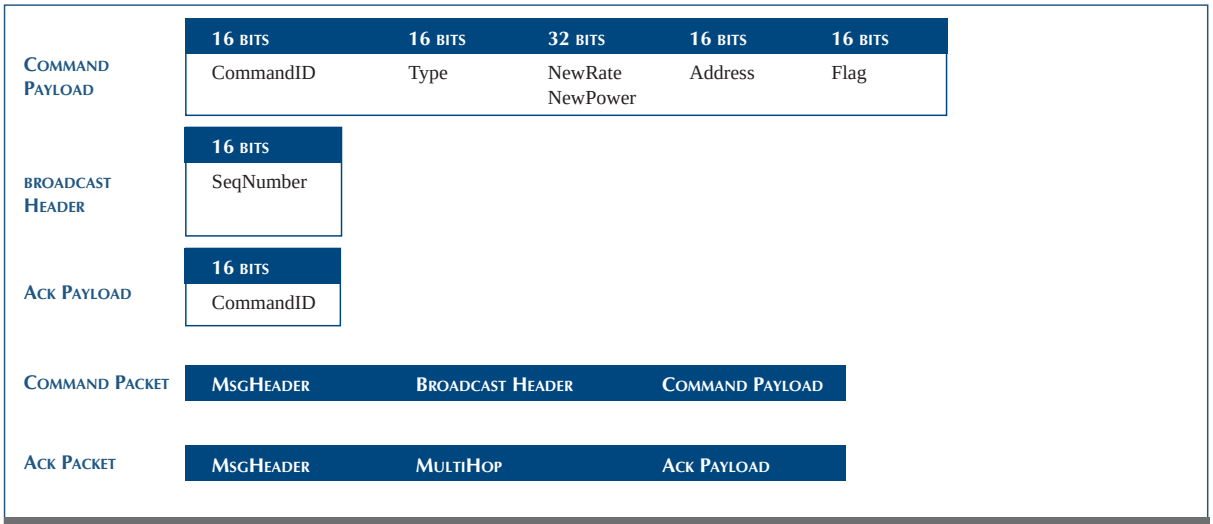

FIGURE 4: COMMAND AND ACK PACKETS.

3.1.3 In-Network Reprogramming

To enable the post-deployment working condition update of sensing nodes, we implemented a policy for remote reprogramming based on diffused commands. The commands, unicast or broadcast, are composed on the client side, injected within the network through the gateway, and flooded by each node. For protocol implementation, we relied on components already integrated within TinyOS library package (BCastM).

The command packet format is represented in Fig. 4. The sequence number is the only information contained in the broadcast header and is used by the protocol to establish if the packet received was new or previously received. The sequence number is filled by the basestation node before sending the message. Five different kinds of commands have been defined,

- the power save (on/off) that activates the radio duty cycle policy.

In the data gathering operation for this application scenario, message loss has no significant consequence on the global application results. Instead, a command message delivery requires reliable communication, as the user must be notified if the command has been correctly received by the target node. Therefore, the CupolaSensing component on each node implements an acknowledgement policy by sending the Ack packets represented in Fig. 4 after the command has been executed. The Ack packets are sent to the basestation using the multihop protocol described in the previous paragraph. The only payload field is the command sequence number, which allows the gateway to associate the Ack packet to the right command.

3.2 Gateway

Within a WSN system, the gateway plays the fundamental role of interfacing a WSN with other networks. In our case study, the gateway has been mapped on a PC running Linux allowing us to interface the WSN with Internet to use client tools for data analysis and network maintenance. The main features implemented enable us (a) to read the packets received by the basestation through the USB port, (b) to interpret their content, (c) to store data in a database, (d) to compose command packets used to reprogram the nodes, and (e) to provide basic services to the clients connected, for example, via Web or via a TCP socket. The overall gateway functional architecture is represented in Fig. 5.

The Serial Forwarder module in Fig. 5 is an application belonging to the TinyOS package. Its main function is to read, without computation, ack and data packets coming from the USB port and, vice versa, to deliver command packets generated on the gateway side to the USB port. The Serial Forwarder opens a TCP connection to which every client can connect. While the physical connection between the PC and Telos uses the USB port, the communication protocol transfers the packets serially with a micro dedicated to the conversion on the Telos and of a SW driver on the PC side. The gateway implements a database service based on MySQL. The database is used by the gateway applications and by the web server integrated within the gateway. Moreover, the database can be accessed by other tools on the client side. In the database, we implemented five tables, which are represented in Fig. 6.

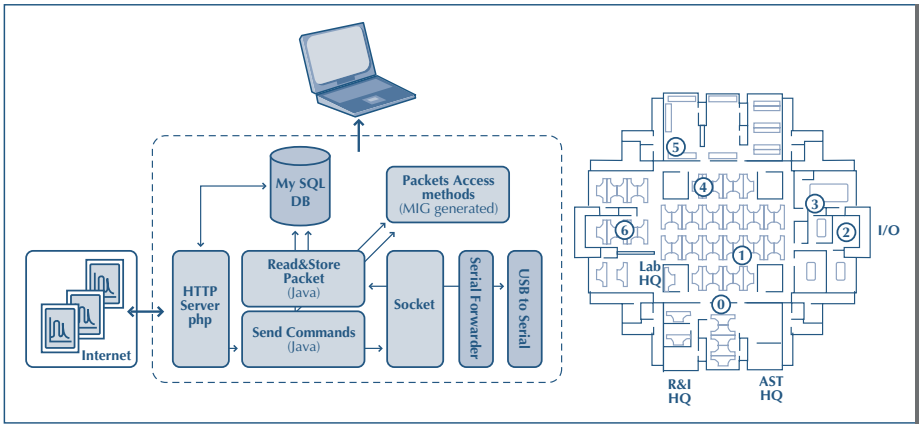

FIGURE 5: GATEWAY ARCHITECTURE AND INTERCONNECTIONS.

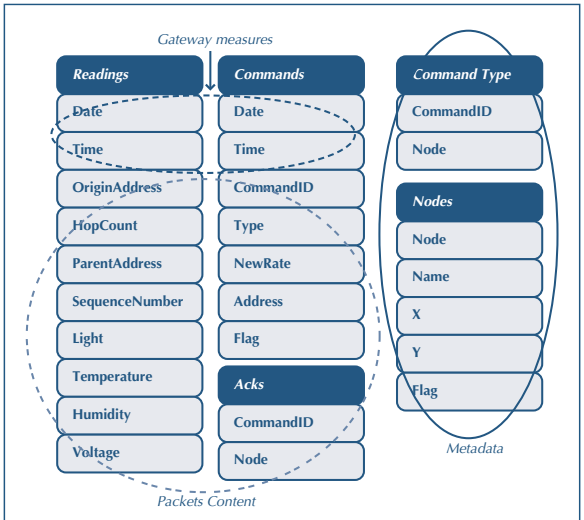

FIGURE 6: MYSQL DATABASE TABLES.

Readings contains data collected by the network. Its fields correspond to the data payload, multihop header, and additional information identifying the time where the packet is received by the gateway.

Commands contains the list of all the commands injected in the WSN with related parameters.

Acks contains the list of commands for which an ack signal has been received.

Nodes and **CommandType** are additional tables with some descriptive info about command types and node attributes, such as their position within the network.

Table content is used by most of the gateway functions providing services to the clients. For example, information stored in Acks and Commands are composed to derive the network status corresponding to the last acknowledged commands for each node or to identify the beginning of the radio active timing window of nodes in power saving. This allows for the synchronization of the command injections within the network.

The gateway modules have all been written in Java. To parse or generate the bit streams going through the serial forwarder, all the applications use the methods generated by the Message Interface Generator (MIG) tool, included in the TinyOS package. MIG is a tool which automatically generates Java classes corresponding to the message format used in a given application. MIG reads the NesC file where the structure of application packet is specified as a *struct* definition and generates a Java class that takes care of the details of packing and unpacking fields in the message byte format. Therefore, the use of MIG is valuable in saving a programmer from parsing message formats for each given application.

The last module implemented on the gateway is an Apache web server providing web pages for client use. The web pages have been written in PHP and can execute simple queries on the MySQL or run the Java applications described above.

3.3 Clients

In the overall WSN system structure, a significant role is played by client tools that are fundamental for network topology visualization, data analysis, network debugging, and network programming. Therefore, we believe the availability of suitable client tools can greatly facilitate the investigation and even the adoption of WSN solutions. In our test case we relied on graphical utilities to interact with the database, to compose commands for network reprogramming, and to analyze at run time the topology and behavior of the sensor network. Most of those utilities are accessible through a WEB browser interacting with the gateway web server or are available as Java services. The choice of a web browser as user interface has clearly been motivated by the popularity within the PC community.

The MySQL database can be queried through MySQL clients, as showed in Fig. 7b, or predefined queries can be submitted through the Web browser, Fig. 7a.

An additional client utility we customized from an existing application is a network visualizer that was a key debugging tool during the experimentation of the test case. As represented in Fig. 7c, the GUI serves two purposes. First, it provides a graphical interface for the user to explore, at run time, the various services offered by the WSN such as temperature and light intensity measures. Then, it ensures the validation of the proper network operation via augmented visualization of its topology, routing tree, and node health status. The network visualizer is accessible from the browser as a java Web service. When activated, it connects with the TCP/IP socket opened by the Serial Forwarder on the gateway, thus receiving a copy of each packet coming from the sensor network. By analyzing the payload and header with MIG-generated methods, the network visualizer can graphically show the network topology and the sensing values of each node.

4. RESULTS AND CURRENT WORK

In this section, we discuss the limits of the current implementation, and we introduce new developments working towards the definition of a general framework for fast prototyping, which we are defining to evaluate STM platforms and sensor components in WSN applications.

Many of the TinyOS modules used to build the Cupola Sensing application and running on sensing nodes have been reused from

the original TinyOS package. Besides fixing the bugs within those components, specific developments were focused on their integration and on the design of dedicated procedures for command parsing, power management, and ack generations, as described in the above sections. Concerning gateway and client services, instead, most of the procedures have been developed from scratch for the specific needs of the test case.

The entire development lasted about two months, during which a significant effort was also carried out to practice with TinyOS existing libraries. The validation of the test case was perfomed for three more months, during which time we further refined and fixed the system architecture and its components.

During the testing period, we stored a significant quantity of data accessible through the database tables. Fig. 8 represents typical graphs retrieved with the GNU plot tool activated by the PHP applications running on the Web server. We precisely quantified the climate conditions of the lab as expected by showing daily and monthly cycles, as indicated in Fig. 8a and 8b, where measures captured with temperature sensors are represented. Furthermore, we roughly analyzed the network behavior by verifying packet (data and ack) loss and some network dynamics such as parent identity changes for each node due to channel interference or node failures, as shown in Fig. 8c.

We believe client tools are a critical component of the WSN system architecture from both a user and developer perspective. The solutions implemented so far were characterized by low flexibility, and current developments are directed towards the integration of more powerful environments such as Excel and Matlab. We believe the integration of these tools is well

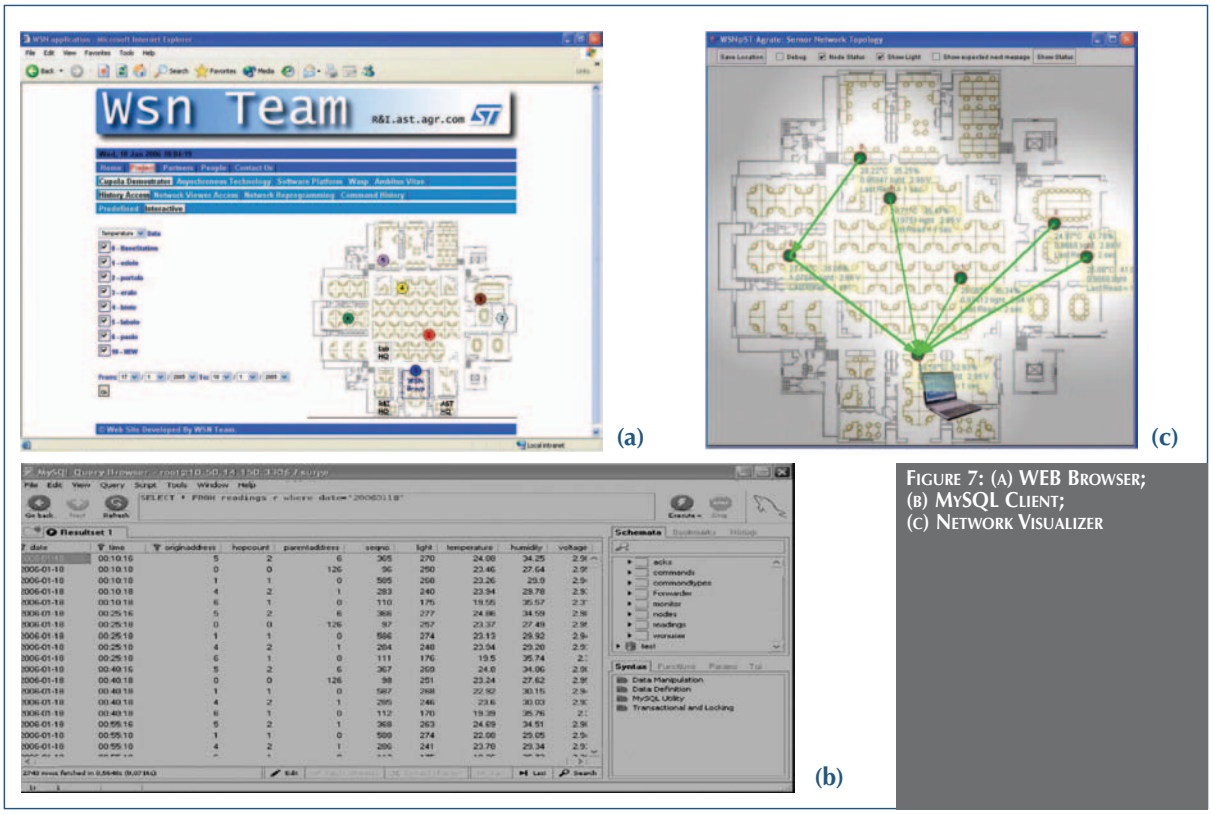

FIGURE 7: (A) WEB BROWSER; (B) MYSQL CLIENT; (C) NETWORK VISUALIZER

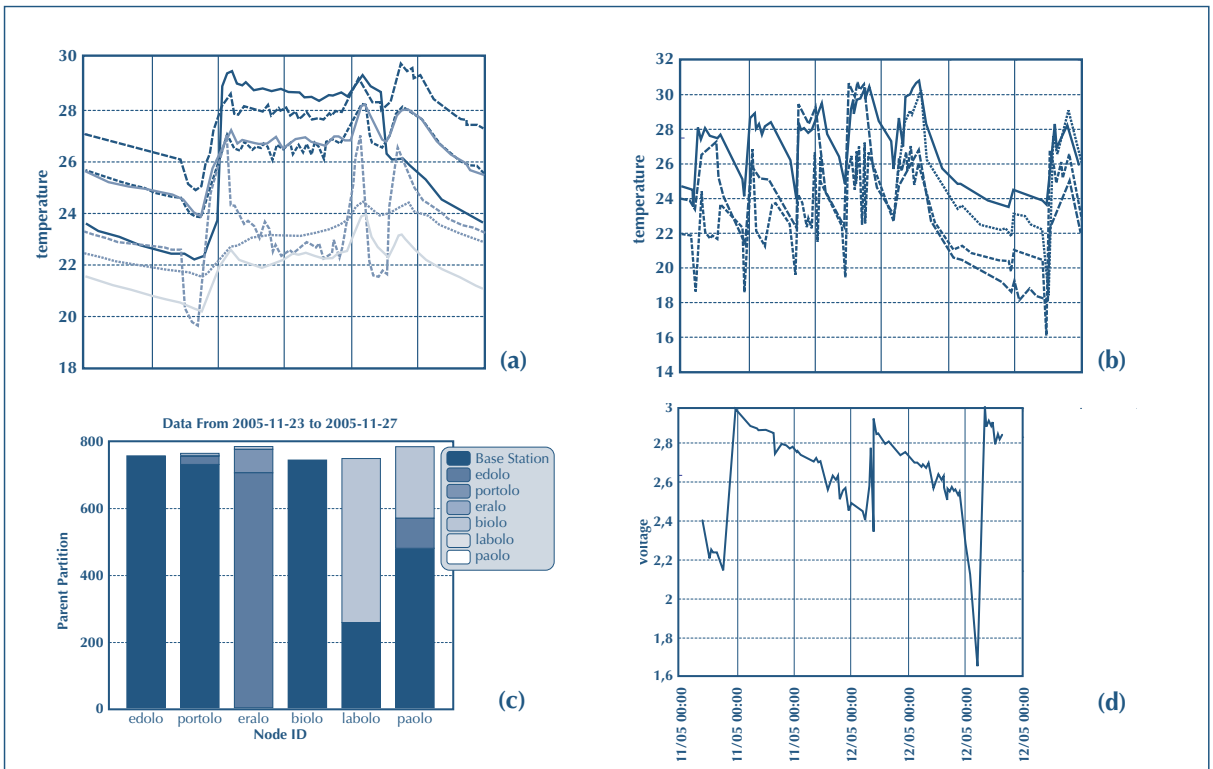

FIGURE 8: EXAMPLES OF SAMPLED DATA: (A) TEMPERATURE DATA IN A DAY WINDOW; (B) TEMPERATURE DATA IN A WEEK WINDOW; (C) PARENT CHANGES FOR EACH NODES DUE TO LQI EVALUATION; (D) BATTERY VOLTAGE MEASURES IN A MONTH TIME WINDOW.

motivated by the advantages they provide in terms of data analysis, data programming and processing, and run time system management and support, all widely used among engineering groups.

With respect to the gateway, we plan to increase its flexibility. We are developing an XML translator that transforms raw data coming from the Serial Forwarder into XML packets that are made available on an additional connection socket. In this way, the clients would not connect to the Serial Forwarder socket, thus avoiding misalignments among MIG-generated libraries working on different client machines. The implementation will be traded off against the run time monitoring requirements of the client tool. We are also defining an API library which will facilitate the client access to other gateway services. Our next step will be porting the developed gateway architecture on mobile HW platforms such as a smartphone or a PDA.

On the network side, one of the main problems we experienced was network debugging, as node and network failures may happen in very unpredictable ways. Additionally, because nodes do not provide a rich machine user interface, debugging is made only more difficult. Wherever possible, a solution is included within the data payload information about current working conditions. For example, we found the use of in-node voltage measurements (Fig. 8d) to verify when the battery was approaching low charge thresholds very useful. When the battery was under that threshold, we also observed unrealistic values of the temperature sensor measurements, and we could easily assess when this was related to the battery voltage. On the other

hand, this solution cannot always be applied, as many times unexpected node behavior consists of the suspension of data flow. We were not always able to understand the reason for these failures. In the test case we did not develop a dedicated in-node solution for auto-diagnosis and auto resetting by using watchdog timers or by logging code to motes flash memory. Therefore, we could only send a reset command remotely or push the reset key of the Telos board.

The power management policy implemented to reduce the radio duty cycle significantly increased the lifetime of the batteried nodes, as described in section 3.1.2. Actually, our current implementation is limited by the lack of distributed synchronization algorithms integrated within the nodes. Hence, we had to be very conservative in the evaluation of active radio timing window (two minutes) to ensure the arrival of beacons, commands, and data to their destination. In addition, time stored in the database corresponds to arrival time and not sampling time, as we do not stamp the data before sending them through the radio. Given the limited network size and our application target, the introduced approximations did not significantly impact the accuracy results. Nevertheless, a detailed analysis of synchronization within the network is one of the on-going developments addressing applications with tighter synchronization constraints (i.e., event localization).

5. CONCLUSION

In this paper, we have described a test case we deployed in our laboratory for climate monitoring. We used this case study to deepen our understanding of the engineering challenges faced by WSN system designers. The functionalities for data gathering, in-network reprogramming, and Web interfacing make this test case highly representative of a large class of applications for indoor monitoring. The implementation was based on open source libraries and tools that allowed for rapid development; it also improved our expertise with respect to state-of-the art solutions in the area of OS for embedded systems and ad-hoc networking protocols. Current developments are focused on the generalization of the test case features which will identify a framework for rapid WSN prototyping.

ACKNOWLEDGEMENTS

The authors want to thank Andrea Serlini for the important contributions to the development and testing of the Cupola Sensing test case.

REFERENCES

[1] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, J. Anderson, "WIRELESS SENSOR NETWORKS FOR HABITAT MONITORING," WSNA 2002.

[2] G. Werner-Allen, J. Johnson, M. Ruiz, J. lees and Matt Welsh, "MONITORING VOLCANIC ERUPTIONS WITH A WIRELESS SENSOR NETWORK," EWSN 2005.

[3] G. Werner-Allen, Pat Swieskowski, and M. Welsh, "MOTELAB: A WIRELESS SENSOR NETWORK TESTBED," IPSN 2005.

[4] Y.Yao, and J.E. Gehrke, "THE COUGAR APPROACH TO IN-NETWORK QUERY PROCESSING IN SENSOR NETWORKS," ACM Sigmod Record, 2002.

[5] J. Hui and D. Culler, "THE DYNAMIC BEHAVIOR OF A DATA DISSEMINATION PROTOCOL FOR NETWORK PROGRAMMING AT SCALE," International Conference on Embedded networked Sensor Systems, ACM 2004.

[6] P. Levis and D. Culler, "MATE: A TINY VIRTUAL MACHINE FOR SENSOR NETWORKS," ACM ASPLOS 2002.

[7] Moteiv Corporation, telos (Rev B) datasheet http://www.moteiv.com , Dec 2004.

[8] TinyOS Home Page http://www.tinyos.net.

[9] NesC Reference Manual http://nesc.sourceforge.net/papers.

[10] MySQL Home Page http://www.mysql.com.

[11] Apache Home Page http://www.apache.com.

[12] PhP Home Page http://www.php.net.

[13] J. Hill, M. Horton, R. Kling, L. Krishnamurthy, "THE PLATFORMS ENABLING WIRELESS SENSOR NETWORKS," Communications of the ACM, 2004.

[14] J. Hill, R. Szewczyk, A. Woo, S. Hollar, D. Culler and K. Pister, "SYSTEM ARCHITECTURE DIRECTIONS FOR NETWORKED SENSORS," ASPLOS, 2000.

[15] D. Culler, J. Hill, P. Buonadonna, R. Szewczyk, and A Woo, "A NETWORK-CENTRIC APPROACH TO EMBEDDED SOFTWARE FOR TINY DEVICES," Intel Research Report, 2002.

[16] C-C Han, R. Kumar, Roy Shea, E. Kohler, and M. Srivastava, "A DYNAMIC OPERATING SYSTEM FOR SENSOR NODES," MobySys 2005.

[17] A. Dunksels, B. Gronvall, T. Voigt, "CONTIKI – A LIGHTWEIGHT AND FLEXIBLE OPERATING SYSTEM FOT TINY NETWORKED SENSORS," First IEEE Workshop on Embedded Networked Sensors, 2004.

[18] J. Hill, D. Culler, "MICA: A WIRELESS PLATFORM FOR DEEPLY EMBEDDED NETWORKS," IEEE Micro 2002.

[19] J. Elson, K. Romer, "WIRELESS SENSOR NETWORKS: A NEW REGIME FOR TIME SYNCHRONIZATION," HotNets-I 2002.

ZIGBEE – THE JOURNEY TOWARD MASS MARKET ADOPTION

Gilles Thonet, Marc Bruel

Schneider Electric

ZigBee has recently been introduced as a promising technology for low-cost, low-power wireless networks requiring flexibility in node placement. Relying on the use of application profiles, ZigBee has been specifically devised for such scenarios as home and building automation as well as sensor networking and automated metering. Although a first specification was completed in December 2004, few products are found today on the market. One reason is that the standard is still being amended and application profiles are not all finalized yet. Another reason is that large-scale deployments in commercial and industrial environments are perceived to be hampered by potential hurdles. This article revisits the path toward ZigBee market introduction in various industries. Objectives are twofold. The first one is to identify technical issues that may constitute an impediment to bringing ZigBee products to life. In particular, the following concerns expressed by industry leaders and end customers, namely, the potential coexistence problems between ZigBee and other wireless technologies such as WiFi, are discussed and illustrated by experiments carried out at Schneider Electric. The second objective of the paper is to describe some of the most promising application areas and what needs to be acted upon to complete the journey toward mass market adoption for each.

1. INTRODUCTION

Wireless sensor networks have gained much attention in the last years. Responding at first to communication needs in military operations, the concept of wireless sensor networking has progressively been permeating most infrastructures and environments we interact with daily. Homeland security and surveillance have been strong catalysts for advancing such a technology, especially since 9/11. Today, business drivers are being found in many other markets. Home and building

automation, industrial control, environmental monitoring, and automated meter reading (AMR) are among those most often cited. The bottom line is deploying cheap, low-power, easily-packageable radio-frequency nodes whose typical tasks range from reporting relatively simple measurement data to participating in more sophisticated control processes. The added value of a wireless sensor network clearly does not reside in high bandwidth but rather in cost efficiency, low power consumption, and the ability to accommodate changing topologies.

Until recently, most wireless sensor network developments have relied upon using proprietary technologies. In addition to disallowing vendor and device interoperability, legacy systems entail higher development costs and riskier sourcing for original equipment manufacturers (OEMs). In 2005, the ZigBee Alliance, an international consortium of major industrial and consumer market leaders, publicly released the first version of an open standard for wirelessly networking control and monitoring devices. The cornerstones of ZigBee are flexibility in node placement, low power consumption, and vendor interoperability through the definition of public application profiles. A new, enhanced version was announced in September 2006, and work on specifying additional features is ongoing.

Manufacturers coming from a broad range of industries have expressed high interest in ZigBee's enabling some of their products. Early introductions have occurred in markets such as building automation, home control, and AMR. The focus, however, has been more on making press releases and investigating the respective markets than building up consistent product ranges. According to many industry observers, ZigBee clearly is on the rise, but adoption, and, to a further extent deployment, are taking more time than initially expected.

Our article aims to revisit the journey toward ZigBee deployment in residential, commercial, and industrial environments. To that end, we have identified some potential obstacles. For instance, a chief concern of building and home product manufacturers seems to be possible interference with other wireless networks, in particular WiFi (i.e., commodity IEEE 802.11b/g). Some experimental results are presented here to support improvements in the standard and motivate the adoption of deployment and installation guidelines. Then, based on this discussion, we further discuss application roadmaps, along with some of the issues to be addressed before product commercialization can begin. The conclusion summarizes the main steps that should guide both the ZigBee Alliance and OEMs in speeding up market introductions.

2. THE ZIGBEE STANDARD

ZigBee Alliance member companies come from a large spectrum of industries ranging from consumer electronics to industrial automation. They share one common goal: enabling interoperability among wireless sensing and actuating devices in various control environments. The ZigBee specification defines a communication stack, several public application profiles (including device and interface definitions pertaining to a given application space, e.g., lighting), several stack profiles (including stack configuration options relevant to a subset of applications, e.g., tree addressing/routing for light home automation), and means for ensuring standard compliance and platform certification. Detailed information is available directly from the ZigBee Alliance website [1]. The rest of this section provides a brief overview of the standard.

2.1 Communication Stack

As shown in Fig. 1, the ZigBee stack architecture primarily relies on three pillars: standard IEEE lower layers, ZigBee layers that define networking and application-level communication mechanisms, and a commissioning framework that will provide reference procedures for initializing and configuring ZigBee networks.

2.1.1 IEEE 802.15.4 Layers

Both physical (PHY) and medium access control (MAC) layers of ZigBee correspond to the IEEE 802.15.4 standard [2]. Based

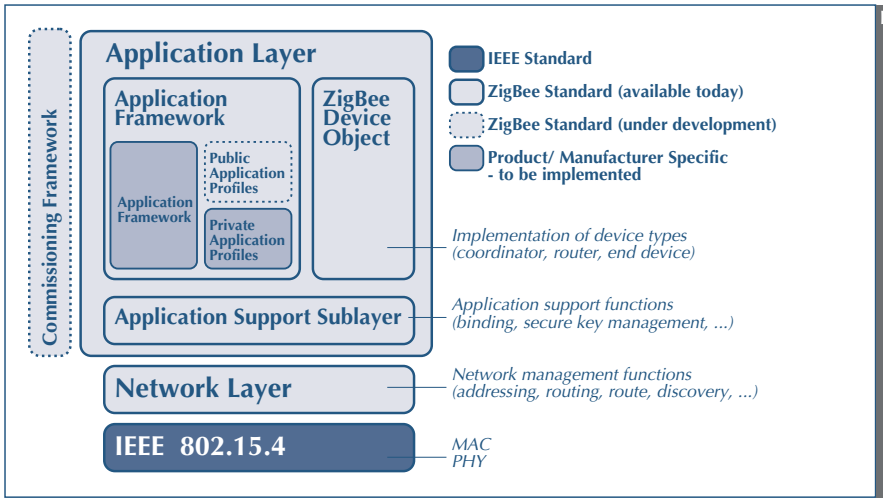

FIGURE 1: ZIGBEE STACK ARCHITECTURE

on direct sequence spread spectrum (DSSS) modulation, the PHY layer has been designed to accommodate low-cost and high-integration constraints. It is actually a dual layer operating in two sets of unlicensed frequency bands, as shown in Table 1. Deployment, however, is expected to occur mainly in the 2.4 GHz band because of worldwide availability and higher data rate. Transmission range typically covers between 10 and 50 m indoor, depending on the surrounding environment.

The MAC design has been essentially driven by the capability of supporting multiple network topologies with minimal complexity. Based on carrier sense multiple access with collision avoidance (CSMA/CA), it includes several features that facilitate power management while reducing functionality whenever desirable.

Two principal IEEE 802.15.4 evolutions are already in the pipeline:

- **IEEE 802.15.4a** addresses amendments to IEEE 802.15.4 for an alternative PHY layer. Currently an IEEE task group, this initiative aims at providing high throughput, ultra-low power, and high-precision location capability to enable new application and market opportunities. The baseline specification includes two optional PHY layers consisting of an ultra-wideband (UWB) impulse radio and a chirp spread spectrum radio. First specification release had been so far planned for the end of 2006.

- **IEEE 802.15.4b** targets a refinement of the current IEEE 802.15.4 specification to clear up ambiguities, resolve

Frequency Bands	Industrial, Scientific, and Medical 2.4 GHz (Worldwide)	Sub-GHz 868 MHz (Europe) 915 MHz (North America)
Channels	16	1 (Europe) 10 (North America)
Modulation	Offset Quadrature Phase Shift Keying (O-QPSK)	Binary Phase Shift Keying (BPSK)
Maximum Raw Data Rate	250 kbps	20 kbps (Europe) 40 kbps (North America)

TABLE 1: DUAL IEEE 802.15.4 PHY LAYER

inconsistencies, and make specific extensions such as faster sub-GHz data rates and time synchronization. The standard has been approved in June 2006, and the first products are expected by the end of 2007.

2.1.2 ZigBee Layers

The ZigBee specification actually takes place above the PHY and MAC layers. The network layer as defined by the ZigBee Alliance [3] provides primitives for managing star, cluster-tree, and mesh topologies (Fig. 2). Star networks are the simplest ones and provide for very long battery life. Mesh (or peer-to-peer) networks form self-healing topologies, enabling higher levels of reliability and scalability by offering more than one routing path. Cluster-tree networks are based on a hybrid star/mesh topology combining part of the benefits of both.

ZigBee nodes belong to one of the following three device types:

- The **coordinator** is responsible for initializing the network and discovering other nodes. There is one and only one coordinator per ZigBee network. When using a star topology, the coordinator routes all data from one node to another since there are no peer-to-peer communications.
  ZigBee coordinators are assumed to be mains-powered.
- **Routers** participate to routing and node discovery tasks.
  They generally are expected to be mains-powered to cope with frequent data transmissions.
- **End devices** typically are battery-powered sensing and actuating nodes. They will be in sleep mode most of the time.

Although the three topologies were introduced in the standard from its inception, support for full meshing has been quite limited so far. This capability has until now been supported by "enhanced implementations" of the specification, such as Ember's EmberZNet stack. Security services such as encryption mechanisms are provided at network and application levels. On top, the application layer supplies primitives to develop interoperable ZigBee products through the use of public application profiles. If interoperability is not a requirement, private application profiles may be implemented and used instead.

2.2 Application Profiles and Stack Profiles

To ensure product and vendor interoperability, public application profiles are being developed by the ZigBee Alliance for areas such as home automation, commercial building automation, and industrial process monitoring. The first one to have been completed was Home Controls Lighting. Based on a rather rigid stack configuration, this profile now serves illustration purposes. Since then, newer application profiles have been developed to address a much larger range of deployment environments (buildings, homes, plants, sensor networks, etc.).

In order to bring further consistency into the choice of a complete communication stack, the ZigBee Alliance has

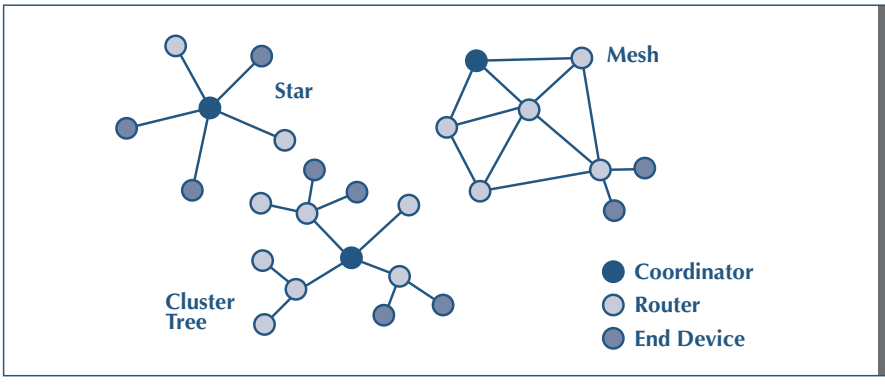

Figure 2: ZigBee Network Topologies

introduced the concept of a stack profile. Whereas an application profile specifies device descriptions and standard interfaces, a stack profile defines various stack configuration options (e.g., addressing mechanism, routing mode, security level, etc.) that apply to a subset of application profiles. There currently are two stack profiles:

- *ZigBee*, previously called *Home Controls (HC)*, which targets moderate-size networks for simple residential applications. This stack profile supports the Home Automation (HA) application profile.

- *ZigBee Pro*, previously called *Commercial, Industrial, and Institutional (CII)*, which focuses on larger-size networks or applications requiring advanced functionality, such as frequency agility, i.e., the possibility of changing channels in case of persistent interference. This stack profile will be working with all application profiles.

3. TECHNICAL HURDLES

As mentioned at the beginning of the paper, some claim that ZigBee adoption is taking more time than originally expected. Although a first version of the specification has been made publicly available since 2005, there still are some uncertainties and technical pitfalls that have slowed down mass adoption. This section aims at discussing two of them and pinpointing possible mitigation strategies.

3.1 Coexistence with Other Wireless Networks

While ZigBee can operate in three frequency ranges, 2.4 GHz is expected to become the widest used band since it simplifies worldwide deployment and provides higher data rate. The increasing presence of other wireless technologies in the same band (WiFi, Bluetooth, cordless phones in some regions of the world) makes, however, network coexistence a source of concern. In particular, WiFi concentrates the largest part of these worries because of its wide spectrum spread, high output power level, and growing presence in residential and office premises. Since these environments are likely to be the footstep of massive wireless sensor network deployment, WiFi appears today as the biggest obstacle when it comes to coexistence with ZigBee.

Although built-in mechanisms help mitigate the impact of WiFi interference, ZigBee networks may be significantly affected by tedious communication exchanges like downloads and media streaming. It is therefore critical to characterize interference levels accurately in order to devise appropriate mitigation strategies and network installation recommendations.

3.1.1 Background

Coexistence performance is affected by a large number of technological, environmental, and application-related factors. Typical ones include:

- distance of the desired transmitter to the desired receiver;
- distance of the interfering transmitter to the desired receiver;
- frequency offset between desired and interfering devices;
- transmitter output power of desired and interfering devices;
- transmitter duty cycle; and
- channel conditions such as path loss, multi-path, fading, and polarization loss.

Channel conditions are usually not controllable, whereas physical distances, carrier frequencies, and output power levels are typical setup or installation parameters. In general, a smaller distance between desired transmitter and desired receiver and a greater distance between desired and interfering devices mean higher coexistence performance. A similar reasoning applies to output power levels. Regarding carrier frequencies, a greater frequency offset from the desired carrier generally results in better interference rejection.

One approach to characterizing ZigBee/WiFi coexistence is superimposing spectrum usage of both technologies, as described in [4]. For illustration purposes, Fig. 3 shows the alignment between the 2.4 GHz European non-overlapping IEEE 802.11b

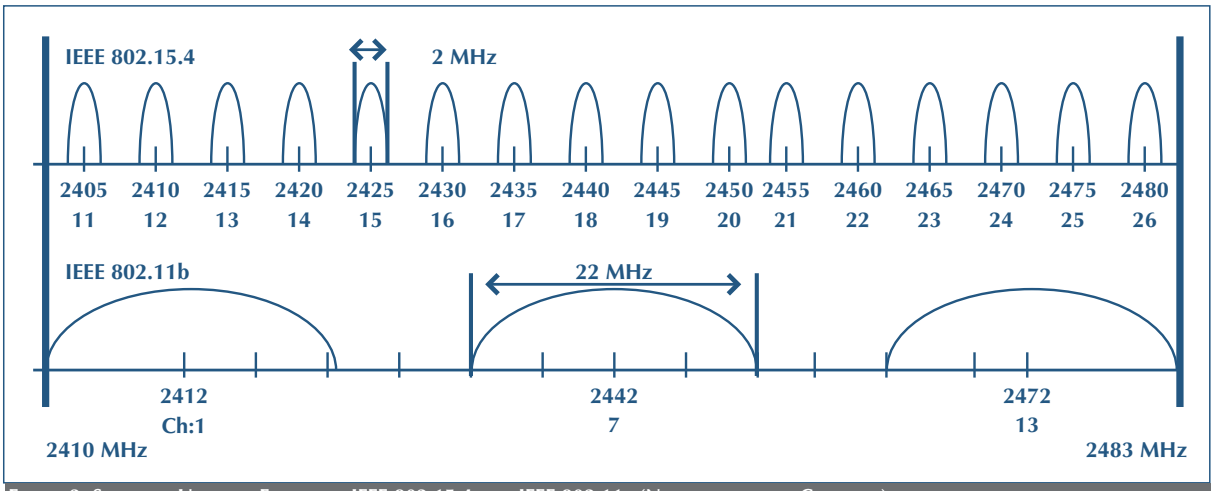

FIGURE 3: SPECTRUM USAGE IN EUROPE OF IEEE 802.15.4 AND IEEE 802.11B (NON-OVERLAPPING CHANNELS)

channels and IEEE 802.15.4 channels. A slightly different figure applies to the North American situation.

There are exactly four IEEE 802.15.4 channels that fall in the guard bands between or above the non-overlapping IEEE 802.11b channels:

- channels 15, 16, 21, 22 in Europe, and
- channels 15, 20, 25, 26 in North America.

Although energy is much lower within these channels, it will not be zero, and interference with close WiFi channels is likely. This is why this approach, while providing some guidance on how to position ZigBee channels with regard to WiFi channels, requires additional experiments to better assess overall coexistence performance.

Some insights into the ZigBee/WiFi interference problem may be found in the literature. Study [5], for example, reports the following results:

- In worst-case conditions with respect to frequency overlap, local distance and high interferer traffic load, more than 90% of the ZigBee frames can be lost. However, a few time slots subsist for successful transmissions. This may not prove sufficient for all applications, but makes a channel switchover possible.

- A minimum frequency offset of about two IEEE 802.11b channels is required to obtain a negligible interference on IEEE 802.15.4 transceivers.

Application note [6] provides a theoretical and experimental investigation of the same issue for an off-the-shelf IEEE 802.15.4 transceiver. Frequency offset, channel bandwidth, duty cycle, and transmitter output power of the desired and interfering signals are all shown to impact coexistence performance. To keep smooth coexistence, the frequency offset between ZigBee and WiFi networks is recommended to be greater than 25 MHz.

3.1.2 Experimental Setup

To better characterize coexistence in real-world environments, a MODBUS (widely spread industrial automation fieldbus) serial line application exchanging data between IEEE 802.15.4 transceivers in the presence of IEEE 802.11b interference has been run. Fig. 4 shows the corresponding test block diagram. Two PCs act as FTP server and FTP client, respectively, to send

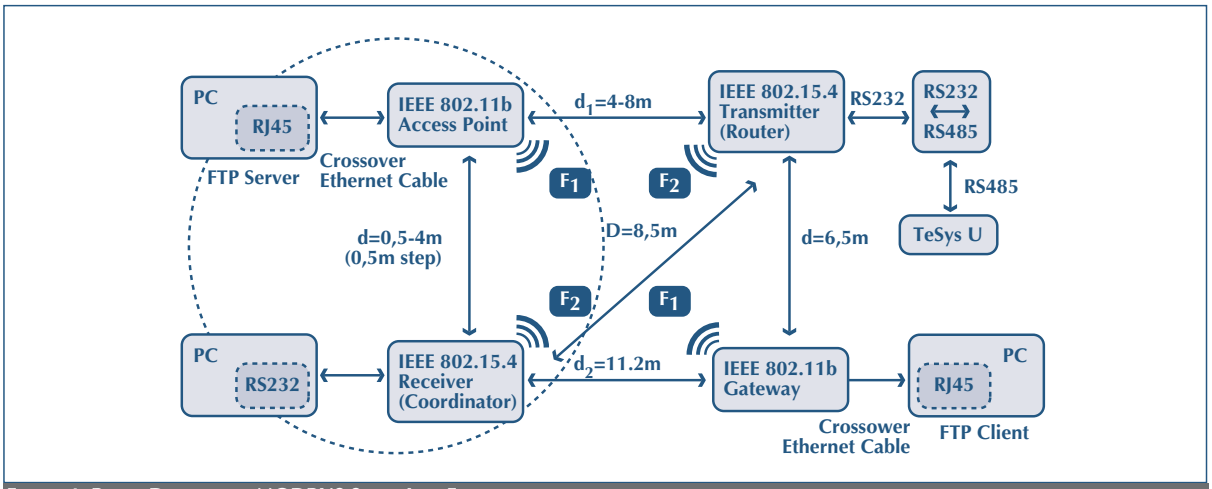

Figure 4: Block Diagram of MODBUS Serial Line Experiment

and receive pseudo-continuous WiFi frames. The serial line application consists of a Telemecanique TeSys U programmable logic controller generating MODBUS frames that are sent through an IEEE 802.15.4 transmitter (acting as ZigBee router) to a remote IEEE 802.15.4 receiver (acting as ZigBee coordinator). ZigBee devices in this setup are off-the-shelf Texas Instruments CC2420 transceivers.

Experiments are carried out as follows:

1. Interfering transceivers are turned off. Desired signal carrier frequency F_2 is set to a given ZigBee channel. N_0, the number of ZigBee frames received in one minute, is calculated.
2. Interfering transceivers are turned on with signal output power of 20 dBm. Raw data rate is set to 11 Mbps. Interfering signal carrier frequency F_1 is fixed at 2442 MHz, a predetermined channel for all tests. WiFi transceivers start an FTP exchange.
3. Desired signal carrier frequency F_2 is set to a given ZigBee channel. For a given distance d (ranging from 0.5 m to 4 m) between the IEEE 802.11b access point and the IEEE 802.15.4 receiver, the number of ZigBee frames received in one minute, N, is calculated.
4. Distance d is incremented by 0.5 m, and step 3 is run again.
5. Ratio N/N_0 is computed for each set of parameters (F_2, d). An overall ratio is finally determined as the average of 15 one-minute tests.
6. For all tests, a calibration phase using a spectrum analyzer allows for the determination of the interfering received power ($P_{r,1}$) and the desired received power ($P_{r,2}$).

Table 2 exhibits IEEE 802.11b received power levels ($P_{r,1}$) with respect to IEEE 802.11b access point/IEEE 802.15.4 receiver distance (d), as estimated during calibration phase.

d [m]	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
$P_{r,1}$ [dBm]	-16	-17	-21	-25	-26	-30	-32	-32

Table 2: IEEE 802.11b Received Power Levels.

Table 3 lists IEEE 802.15.4 received power levels ($P_{r,2}$) with respect to IEEE 802.15.4 transmitter/IEEE 802.15.4 receiver distance (D). The actual physical value is D = 8.5 m. Values D > 8.5 m have been simulated by varying the measured power output levels $P_{r,2}$ at IEEE 802.15.4 receiver during calibration phase.

d [m]	8.5	20	30	90
$P_{r,2}$ [dBm]	-61	-65	-71	-81

Table 3: IEEE 802.15.4 Received Power Levels.

Using this experimental setup, the success rate of ZigBee frame reception in the presence of WiFi interference can be calculated with respect to the frequency offset and physical distance between ZigBee and WiFi nodes. For illustration purposes, Fig. 5 presents results obtained for a ZigBee received power level of –65 dBm (corresponding to a physical distance of 20 m). Although a clear performance decrease is observed whenever ZigBee and WiFi channel frequencies are nearby ($\Delta F < 10$ MHz), coexistence may still be an issue for larger frequency offsets when ZigBee and WiFi nodes are physically close to each other.

Based on this set of results, physical distances and frequency offsets leading to smooth coexistence have been determined in Table 4, Table 5, and Table 6. Values are to be interpreted as follows. For instance, to guarantee successful delivery of 80% of the packets, two ZigBee nodes can be 30 m apart in free space if the WiFi interferer is at least 2 m apart and the frequency offset is greater or equal to 25 MHz.

D_{max} [m] ($P_{r,2}$ [dBm])	d_{min} [m] ($P_{r,1}$ [dBm])	ΔF_{min} [MHz]
90 (-81)	0.5 (-15)	0

Table 4: Coexistence Parameters for $N/N_0 > 5\%$.

D_{max} [m] ($P_{r,2}$ [dBm])	d_{min} [m] ($P_{r,1}$ [dBm])	ΔF_{min} [MHz]
20 (-65)	2 (-25)	20
30 (-71)	2 (-25)	25
90 (-81)	2 (-25)	30

Table 5: Coexistence Parameters for $N/N_0 > 80\%$.

D_{max} [m] ($P_{r,2}$ [dBm])	d_{min} [m] ($P_{r,1}$ [dBm])	ΔF_{min} [MHz]
30 (-71)	2 (-25)	30
90 (-81)	3 (-30)	30

Table 6: Coexistence Parameters for $N/N_0 = 100\%$.

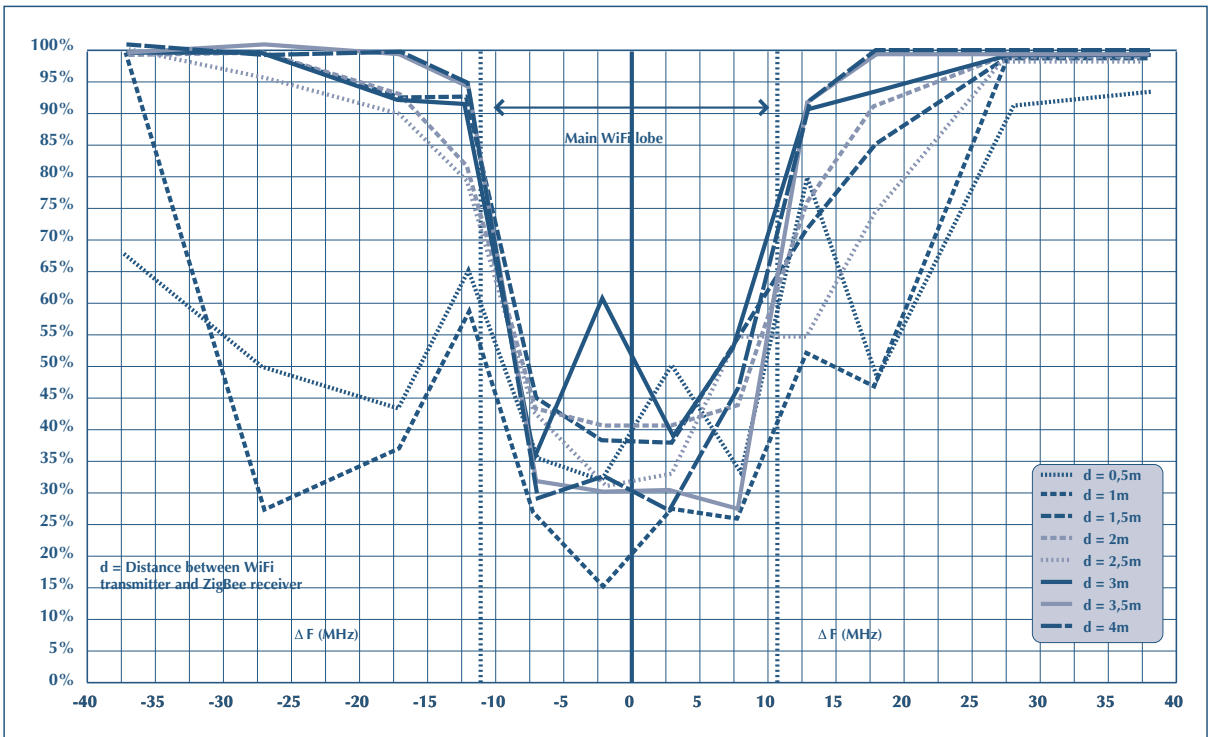

Figure 5: Results of MODBUS Serial Line Application for D = 20 m

3.1.1 Mitigation and Recommendations

Analyzing previous experimental results leads to two basic installation recommendations to guarantee an 80% packet delivery rate:

- Distance of the WiFi interferer to the ZigBee nodes should be at least 2 m.
- Frequency offset between both networks should be at least 30 MHz.

These thresholds are formulated as "safe-side" values, i.e., it is likely that many situations and environments could afford more relaxed recommendations. They should be considered as upper bounds, ensuring smooth coexistence of both networks.

Experimental results also showed that, thanks to the CSMA/CA mechanism implemented in the 802.15.4 MAC layer, ZigBee transmissions are never completely interrupted, even in presence of heavy WiFi traffic. Other research groups obtained similar results [5]. This suggests that some mitigation mechanisms can be introduced in the ZigBee specification to detect interference and move to another channel distant enough to prevent performance degradation due to WiFi proximity. This feature is currently under investigation in the ZigBee Alliance. Devising such a functionality is obviously a difficult task given the many technical issues to be solved (e.g., how to accurately detect an interference, what this means for very large networks, and how to deal with sleeping devices).

3.2 Power Consumption Issues

Another important issue for wireless sensor network applications is power consumption. Current ZigBee deployment models assume that both the coordinator and routers are permanently mains-powered, while end devices can be battery-powered. Typical battery usage scenarios reach a 1-to-10 year lifetime, depending on duty cycling. Whether power consumption is a limiting factor highly depends on application requirements.

At least three scenarios will require improvements in power consumption to come to maturity:

- Residential applications such as lighting and temperature control are asking for very low energy consumption. Although sensors and actuators are expected to be in sleep mode most of the time, changing batteries is always a cumbersome operation for the end customer.
  Batteryless devices like switches powered via piezoelectric or electromagnetic means are extremely attractive solutions for this market segment. The availability of such technologies in the future could significantly boost ZigBee deployments.
- Large-scale wireless sensor networks deployed for environmental or structural health monitoring will obviously not accommodate mains-powered supply.
  This means that energy must be provided to all ZigBee nodes (including routers) through batteries or external sources such as solar cells. A Profile Task Group within the ZigBee Alliance, Wireless Sensor Applications, aims at specifying modifications or improvements in the current stack to fulfill these objectives. New energy saving or routing strategies are to be worked out.
- Industrial applications in which sensors are hardly reachable or attached to moving or rotating parts are natural candidates for alternative powering models.
  Enhancements in battery technologies and power scavenging techniques are two relevant research directions with respect to that.

4. THE PATH TOWARD PRODUCT DEPLOYMENT

Although some technical hurdles are still to be cleared, ZigBee is on its way toward reaching mass market deployment. This section browses through the most promising application segments and, for each one, describes some of the steps required to eventually achieve this goal.

4.1 Building Automation

4.1.1 Business Drivers

Today's commercial buildings are managed by large sets of sensors and actuators which control such functions as lighting, heating, air quality and security. Optimizing energy efficiency

and increasing the comfort of occupants are two strong drivers for using complex building management systems. Tertiary and commercial premises typically rely on standardized communication protocols (such as BACnet and LonWorks) to transport command and sensing data across building networks. To date, such communication networks have been almost exclusively deployed on wired media.

Although the current cost of a wireless solution may not be always lower for new constructions, it is definitely a strong driver for retrofit installations or facilities in which it is highly expensive to install a wired communication network (e.g., buildings with concrete walls, museums, and architectural or historical sites that cannot be disturbed). In these scenarios, adding wired controls and sensors may end up being much more costly than deploying a wireless solution. ZigBee chips can be embedded in devices like controllers, switches, and sensors for light, temperature, air quality, or presence detection, as illustrated in Fig. 6. With no wiring required, these devices can be placed almost everywhere and easily moved when the building organization or functionality need to evolve. As underlined by research company Frost & Sullivan in [7], "Whether used in new constructions or retrofit, flexibility is the ultimate benefit in deploying a wireless system as opposed to a wired network. These sensors can be located – or relocated – to optimize system performance, increase customer comfort, and adapt to changing floor plans."

Usually, deploying more sensors and actuators means obtaining more energy-efficient buildings. By enabling the collection of a much larger amount of data, wireless systems pledge even greater energy and cost savings associated with better optimization of lighting and HVAC (Heating, Ventilation, Air

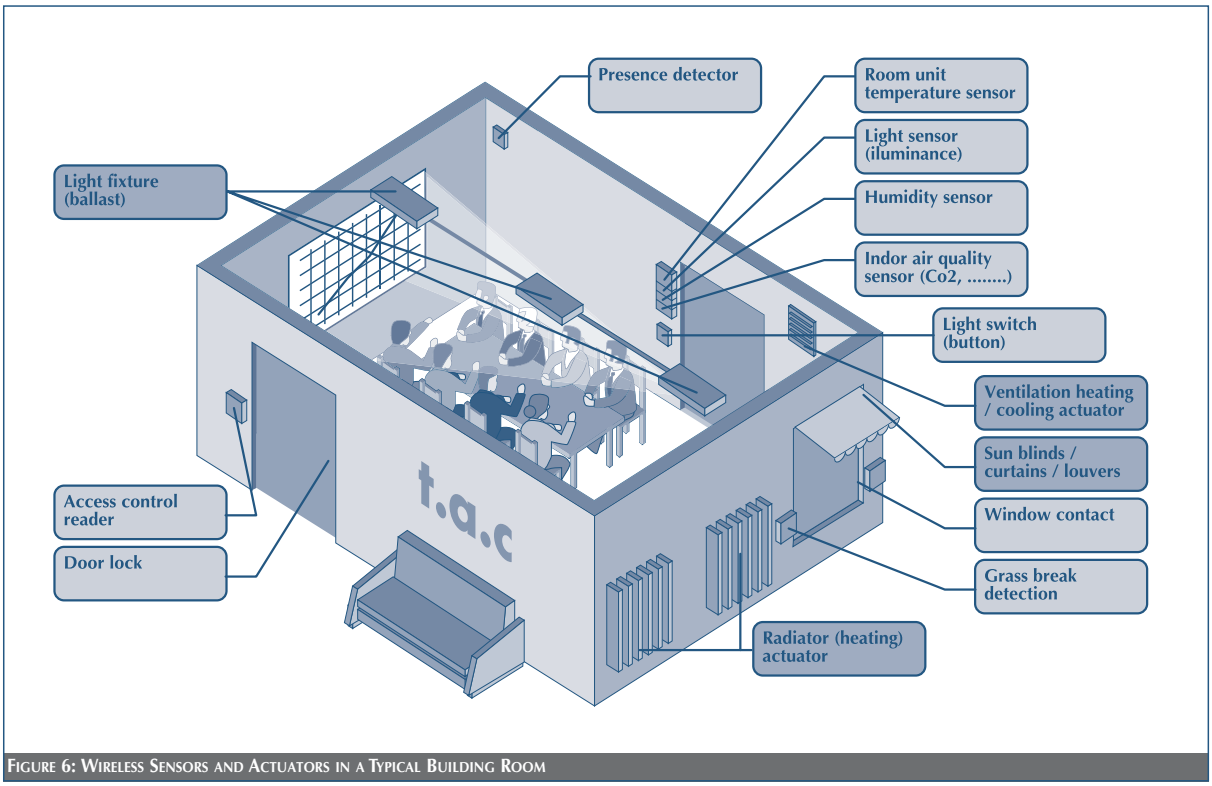

FIGURE 6: WIRELESS SENSORS AND ACTUATORS IN A TYPICAL BUILDING ROOM

Conditioning) functions. The increasing pressure to improve building efficiency by even a few percentage points makes such solutions worthwhile.

One may argue that these benefits are not particular to ZigBee but to wireless technologies in general. The added value of a ZigBee solution resides in vendor and product interoperability. An open international standard that fosters competitive sourcing and the availability of multi-vendor platforms holds the promise of bringing new control options that previously were only available to high-end markets. The influential presence in the ZigBee Alliance of leading building automation suppliers (such as Honeywell, Mitsubishi Electric, Schneider Electric/TAC, and Siemens) and the recent release of early ZigBee products (such as TAC's Andover Continuum Wireless and Siemens' APOGEE Wireless) put this market segment at the forefront of upcoming deployments. Business drivers are summarized in the table below:

Drivers for Adopting ZigBee in Building Automation

- Vendor interoperability ensured by an open international standard
- Competitive sourcing market and availability of multi-vendor platforms
- Flexibility to adapt to changing building arrangement or functionality
- Strong enabler for increased energy efficiency in large facilities
- Lower installation and maintenance costs (first in retrofit installations, then in new constructions)

4.1.2 Next Steps

The following issues need to be engaged with in order to accelerate the deployment of ZigBee for building automation:

- From a standardization point of view, completion of the appropriate application profile work (CBA, BACnet study group) is a prerequisite to holding the promise of interoperability. Common application bricks will deliver better interfacing capabilities with other equipment manufacturers, allowing customers to shift to other product or service providers without additional expenses in reworking the entire solution.
- Integration with existing building automation protocols is also desirable. For example, liaison with BACnet is being investigated by the ZigBee Alliance. Appropriate gateways also need to be developed to connect ZigBee networks to existing infrastructures such as Ethernet, LonWorks and KNX, for instance.
- Coexistence issues as described in section 3.1 have frequently been utilized by opponents to wireless technologies to promote traditional wired or powerline carrier solutions. In addition to frequency agility mechanisms to be implemented in the ZigBee stack, building management applications require installation guidelines to address concerns expressed by end customers. One can imagine devising ZigBee/WiFi node distance and frequency separation recommendations (based on conclusions like the ones presented in section 3.1) that would translate into specific rules for hotels, malls, or other commercial premises.
- Eventually, deployments in real and large-scale environments are needed to assess architectural choices (e.g., single large multi-hop network covering a significant part of the building vs. several smaller ZigBee networks with a limited number of hops and interconnected with each other).

4.2 Home Control

4.2.1 Business Drivers

Along with building automation, home control is expected to become one of the top market segments in terms of product deployment. The type of use foreseen for ZigBee in this area encompasses everything from domestic TV remote controls and central heating to lighting, rolling shutters, and alarm systems. Developing such devices requires great OEM involvement, which in turn needs large markets to generate adequate return on investment.

Although vendor and product interoperability represents a strong confidence factor in gaining access to large markets, this

is not the principal driver for adopting a standard in the home automation segment. Cost issues are much more sensitive than those in building automation systems. For OEMs, having an open standard like ZigBee means a dynamic sourcing market that drives costs down and boosts technological innovation. Eventually, these competitive pricing strategies will benefit end customers.

Beside cost issues, power consumption is also an important factor, as mentioned in section 3.2. Allowing customers to easily reposition switches and other low-end products while getting rid of power mains is a strong driver for adopting a technology like ZigBee. Although batteries are not foreseen to disappear soon, scavenging energy to make control devices fully autonomous brings great prospects to both new construction and renovation markets. Principal business drivers for residential control applications are summed up below:

Drivers for Adopting ZigBee in Home Control

- Competitive sourcing market and multi-vendor platforms
- Low power consumption / Possibility for future batteryless solutions
- Flexibility to reposition control and sensing devices across the home
- Increased energy efficiency and comfort
- Vendor interoperability ensured by an open international standard
- Lower installation costs (first in retrofit installations, then in new constructions)

4.2.1 Next Steps

A successful deployment roadmap for home control applications has to deal mainly with the following provisions:

- ZigBee standardization is more advanced on the residential side than on the commercial building side. Although the HA application profile has been completed, there are still some uncertainties as to whether to go for ZigBee or ZigBee Pro as the underlying stack profile. Many OEMs are willing to make use of the forthcoming ZigBee Pro stack profile to benefit from its advanced features.
- Wireless coexistence is also an important concern for home control product manufacturers. In contrast with building automation environments, enforcing network installation rules and recommendations might be more problematic among end customers and installers.

  While carefully choosing WiFi channels or constraining WiFi throughput during heavy operations like streaming is feasible in a commercial environment, this is clearly not achievable at home. Addressing interference problems will need to rely mainly on built-in mechanisms, such as frequency agility.

4.3 Automated Meter Reading

4.3.1 Business Drivers

Automated meter reading (AMR) was first deployed 40 years ago by AT&T and a group of utilities. At the time, phone-based services were provided at about four times more than the cost of a person to read the meter. After these successful but expensive experiments, AMR enjoyed constantly decreasing operating costs and became a widely accepted technology. Today key drivers for installing an AMR system are lowering costs, avoiding reading errors, and collecting relevant data as quickly as possible.

The use of wireless technologies is mainly propelled by the prospect of further cost reductions. According to research house ON World, wireless-based AMR can cut water, gas, and electricity costs by up to 25%. In a world where utility costs are rising by as much as 10% per year, such figures are of great appeal to this industry. ON World believes wireless networks will grow the fastest of all AMR networking solutions over the next years. New products are appearing at a fast pace for all geographical needs (urban, suburban, rural) and across various application requirements (industrial metering, utility metering, and sub-metering in large commercial facilities).

Positioned on the low-cost edge of wireless technologies, ZigBee can be considered as a strong contender in the race for advanced AMR solutions. Aggressive pricing trends are likely to be their best advocates. Although latency is not a critical

factor as it is for lighting, for instance, guaranteeing secure data delivery is of prime importance. Against other low-rate wireless technologies, ZigBee may also play the relatively high-performance advantage. Principal business drivers for metering applications are listed below:

Drivers for Adopting ZigBee in Automated Meter Reading

- Cost reductions in deploying and monitoring utility networks
- High performance and built-in security
- Competitive sourcing market and multi-vendor platforms

4.3.1 Next Steps

In contrast to building and residential markets, interoperability is not among the main drivers for AMR applications. Deployment of AMR solutions is still highly fragmented across regional or local markets, and utilities and OEMs enjoy the possibility of providing legacy systems, if cost requirements are fulfilled. From that perspective, having a non-ZigBee but IEEE 802.15.4-based solution may be a viable option for some AMR players.

Nevertheless, the availability of a public ZigBee profile addressing AMR needs may facilitate and accelerate product development. Resolving current uncertainties around this public profile will be key in adopting ZigBee in this industry.

4.4 Industrial Automation

4.4.1 Business Drivers

Industrial automation is a rather ambiguous market segment for ZigBee. Although initially presented as one of its primary target industries, industrial control now tends to be perceived as a likely late adopter, given the specific constraints of manufacturing and process environments. Robustness and real-time communication requirements are often cited as potential obstacles to using wireless technologies for industrial control. An even bigger impediment is the conservative mindset that characterizes the adoption of new technologies among industrial automation customers.

Given the large and diverse range of communication requirements in industrial environments, the deployment roadmap needs to be split into two major application categories:

- *Open-loop applications* refer to typical monitoring and data collection tasks that are not part of the control loop itself. Wireless communications are used in these cases for non-critical operations such as commissioning, diagnostics, condition monitoring, etc. ZigBee is here well-positioned in terms of cost efficiency and flexibility to relocate industrial devices across the plant.
- *Closed-loop applications* are in contrast related to critical tasks since they are part of the control loop. Examples include real-time process control or safety-related operations like commanding valves. Such scenarios will likely not be part of wireless technology roadmaps for 3 to 10 more years.

In summary, one can pinpoint the following main business drivers for adopting ZigBee in the industrial market segment:

Drivers for Adopting ZigBee in Industrial Automation

- Flexibility to adapt to changing plant configuration or functionality
- Lower installation and maintenance costs
- Competitive sourcing market and availability of multi-vendor platforms
- Vendor interoperability ensured by an open international standard

4.4.1 Next Steps

The first set of applications that will be ready for deployment in manufacturing and process environments are open-loop systems. Combined wired/wireless networks might be the first instantiation, in which control data are transported through wires, and a wireless link is used for commissioning and configuration purposes. Other promising areas include the monitoring of process variables, production equipment, or clamp-on types of instruments for permanent or temporary installations. While acquiring sensing data is fairly easy, conveying these data

to a central control system remains a significant challenge since it requires tight integration with existing automation infrastructures.

If a large number of devices is to be deployed, power consumption may become a critical issue. Changing batteries frequently is either not feasible or not wanted by plant operators. This means that advanced energy optimization techniques need to be devised to maintain an acceptable service level while significantly decreasing duty cycles.

Further down the road are closed-loop applications involving wireless communications. Several technical hurdles, such as guaranteeing high quality of service and implementing appropriate time synchronization mechanisms, need to be addressed before these scenarios become reality. It is still unclear at this stage whether ZigBee will be the most appropriate technology to take up these challenges.

5. CONCLUSIONS

This paper has highlighted some important challenges on the road toward deploying ZigBee in residential, commercial, and industrial applications. Although much work has been carried out since the completion of the first specification in December 2004, there still are some issues to be addressed before we will see mass market adoption. In particular, this article has pinpointed three important matters:

- Finalizing public application profiles and elaborating appropriate compliance testing and certification programs. The most urgent actions are clearly required on the commercial building side since interoperability is a less attractive feature for metering, and to some extent industrial applications. The question of choosing which stack profile for which application profile also needs to be sorted out. Having several stack profiles could also be questioned.
- Addressing end customer concerns, in particular coexistence with WiFi networks. This article has presented experimental results showing that, although interference may have to be dealt with, mitigation techniques can be implemented. ZigBee traffic is never completely interrupted, and installation rules can greatly help reduce the impact of WiFi perturbations.
- Improving power usage of current and future energy sources. Evolution of ZigBee deployment scenarios will have to align with progress made on the power technology side. One can expect an increasing demand for better power consumption schemes involving either improved batteries or batteryless systems relying on alternative scavenging methods. Such enhancements will gain increasing momentum in residential and industrial market segments.

The journey toward mass market deployment of ZigBee technology is not over yet but is heading the right way. The recent involvement of large OEMs and the increasing number of ZigBee radio solutions indicate that the market is now ready for end-product offerings. Adopting an open, international standard that provides multi-source development platforms is today the best strategy for OEMs to be successful in their markets. From that perspective, ZigBee today appears as the only viable alternative in the low-cost, low-power, wireless control segment. The capability of the ZigBee Alliance to work out the remaining technical and application profile issues will be instrumental in accelerating mass market adoption.

REFERENCES

[1] ZigBee Alliance website: www.zigbee.org

[2] IEEE Std 802.15.4-2003. Part 15.4: "WIRELESS MEDIUM ACCESS CONTROL (MAC) AND PHYSICAL LAYER (PHY) SPECIFICATIONS FOR LOW-RATE WIRELESS PERSONAL AREA NETWORKS (LR-WPANS)," 1 October 2003.

[3] "ZIGBEE SPECIFICATION," Version 1.0. 14 December 2004.

[4] A. Gutiérrez, E. H. Callaway, R. L. Barrett, "LOW-RATE WIRELESS PERSONAL AREA NETWORKS – ENABLING WIRELESS SENSORS WITH IEEE 802.15.4," IEEE Press. 2004.

[5] A. Sikora, "COMPATIBILITY OF IEEE 802.15.4 (ZIGBEE) WITH IEEE 802.11 (WLAN), BLUETOOTH, AND MICROWAVE OVENS IN 2.4 GHZ ISM-BAND – TEST REPORT," Steinbeis-Transfer Center, University of Cooperative Education, Lörrach. 12 September 2004.

[6] MC1319x Coexistence. Freescale Semiconductor Application Note AN2935. July 2005.

[7] **Wireless Sensors in Building Automation – Technical Insights**, Frost & Sullivan. February 2005.

■ CONTACT: ST.JOURNAL@ST.COM ■

LOCALIZATION IN SENSOR NETWORKS

K. Yao[1], F. Lorenzelli[2]

(1) University of California, Los Angeles
(2) STMicroelectronics

In this paper, we review the concepts of source and node localization with reference to a sensor network. We introduce several algorithms that have appeared in the literature as well as some that have been developed and used by the authors over the years. The methods presented are based on different measured signal quantities, such as received signal strength, angle of arrival, time or time-difference of arrival, all of which are defined and described in the paper.

1. INTRODUCTION

The progress made in communication, computer, networks, and microelectronics in the last fifty years made possible the emergence of sensor networks (SNs). In recent years, SNs constitute one of the most active academic research topics in the system area. It is believed that sensor networking in the 21st century will be enormously significant by providing measurement of the spatial and temporal physical phenomena around us, leading to a better understanding and utilization of this information in a wide range of applications. Sensor networking will be able to bring a finer-grained and fuller measurement (using acoustic, seismic, magnetic, infrared, imaging, etc., data) and characterization of the world around us to be processed and communicated, so the decision makers can use the information to take actions in near-real-time.

Sensor Networks combine the technology of modern microelectronic sensors, embedded computational processing systems, and modern communication and computer networking methodology. However, due to the limited power and energy from the battery, hostile radio propagation environment, constrained radio and processor capabilities, and random/unknown sensor placements, the analysis, design, and realization of sensor networks to perform meaningful tasks remain challenging.

1.1 Localization

Localization, the determination of the location of an object, whether that of a source emitting some energy (e.g., RF, acoustic, ultrasonic, optical, IR, seismic, thermal, etc.) of interest, or a sensor node in the SN, is a major issue of interest. Two families of algorithms can be identified under the heading of "localization": those used to locate the individual sensors in a SN with respect to each other and possibly to a point of reference will be referred to as *node localization* algorithms. This scenario arises when the nodes are not (or cannot be) placed with desired accuracy, due perhaps to a hostile environment or other reasons. Once the sensors' locations are known, either through a preliminary phase of node localization, or by accurate placement and calibration, then it becomes possible to locate external sources of energy. When this is the case, we will be talking of *source localization* algorithms.

As an example of node localization familiar to many, in the U.S., the Federal Communications Commission (FCC) has mandated the E911 regulation, (Europe has recommended the E112 regulation), requiring the cellular telephone provider to determine the location of a cell phone user (in an emergency mode) to tens of meters. Source localization may arise in various military scenarios, such as the localization of a vehicle, or in automated manufacturing, for instance the localization of a robotic device.

Localization of an object with an appropriate receiver can be achieved by many means. Satellite-based localization based on global positioning systems (GPS), and variations including the Global Navigation Satellite System (GLONASS), and the European Geostationary Overlay Service (EGNOS), have been used for many years for military, commercial, and consumer applications. In open-field scenarios, these localization systems may be adequate. However, localization based on these systems may not be appropriate due to their limiting precisions caused by severe propagation degradations inside buildings and in non-line-of-sight (NLOS) situations, or due to the exceeding complexity of the overall system, especially in low-cost SN deployment situations.

In this tutorial paper, we introduce the basic operations, including some equations, and their applications, to various localization schemes. In Section 2, we deal with several methodologies for sensor node localization. We explore the issue of statistical characterization of the measurements and how it affects the performance of the location estimators. We also describe a Gauss-Newton iterative node localization method in its centralized and distributed versions. In Section 3, we deal with several methodologies for source localization. They include, the trilateration/multilateration method, time-difference of arrival (TDOA) method, the received-signal-strength (RSS) method, and the direction-of-arrival (DOA) method (also called the angle-of-arrival (AOA) method).

2. NODE LOCALIZATION

In sensor networks it is assumed that small, inexpensive, collaborative, and relatively autonomous nodes are placed with varying degrees of accuracy, with the purpose of monitoring the environment. Depending on the applications, sensors are required to respond to changes in the environment by sending out the collected information to neighboring nodes or to a central station. For these reasons, the nodes are equipped with sensing, computation and wireless communication devices.

Applications may range from earthquake to agricultural (water and fertilizer) monitoring, from automobile traffic control to security alarm systems, etc. In many situations, it is paramount that the sensors be aware of their relative, and sometimes absolute, location. This is known as a node localization problem in the sensor network literature. Only by knowing the position of the sensor can the sensed information be placed in the context and made relevant for the application at hand. It is assumed that the individual devices are provided with small batteries that are replaced relatively infrequently [18]. Because of this, localization techniques based on global positioning systems may be considered too power-hungry, and alternative techniques are then called for.

The different techniques for node localization can be generally classified into two categories: localization based on anchors (or

beacons), and anchor-free localization [21]. Anchor nodes are special nodes equipped with special positioning devices. Other nodes try to determine their position relative to the anchors. Once their positions have been estimated, regular nodes may in turn become anchors, etc. Clearly, anchor-based localization algorithms need another positioning system in place to provide initialization information. This may be attained outdoors with special nodes having global positioning system information, or indoors by nodes whose positions are known by design. In anchor-free localization algorithms, all nodes collaborate with each other (usually with their neighbors) in order to determine a relative map. A post-processing is subsequently performed to convert the relative map into an absolute position information.

Another categorization that can be made of localization algorithms is among centralized, localized and distributed algorithms [11]. In centralized algorithms, all the information collected by the sensors is transferred to a central node that performs the localization algorithm and computes each node's location. The problem to solve typically is to search for a global maximum of a likelihood function. This problem is in general very hard, because of the multi-modal structure of likelihood functions in most realistic situations. In addition, the function to maximize is usually model-dependent, and any deviations from the assumed model may significantly degrade the estimates. The problem may be formulated as a convex program, and semidefinite programming formulations have been proposed [1]. The localization algorithm may be sometimes distributed, so that each node runs a subset of the operations required to obtain the localization estimates, and passes information to its neighboring nodes. In this way, the communication bottleneck due to the need to forward all the information to a central processor is obviated. More energy will probably be spent to exchange information locally, as opposed to what is required to send all the information to the central node, and it is conceivable that a trade-off can be achieved. After a few iterations and a few information passes between adjacent nodes, the localization information should converge for all nodes. Ideally, the distributed algorithm will find the same global optimum as the centralized algorithm. In making the algorithm distributed, convergence issues need to be addressed to insure that this is indeed the case. In particular circumstances, the position can be computed locally by the individual nodes, for instance when each node receives a signal from a set of beacon nodes, and estimates its own location based on the received signals. Localized algorithms use the estimated distance between the node and the anchor nodes in much the same way as trilateration and multilateration in source localization algorithms. It is generally assumed that wireless sensor networks, in order to minimize installation expenses, are relatively sparse, and anchor nodes may be separated by large distances. Moreover, low-power sensor nodes are usually not in the range of a sufficient number of anchors. In these cases, localized algorithms are probably not the best choice.

Many localization algorithms rely on measurements of distance between nodes [13]. These algorithms are known as range-based. The types of distance estimation that is usually performed are based on measurements of received signal strength (RSS), or estimates of time of arrival (TOA) or time difference of arrival (TDOA). As opposed to range-based algorithms, range-free techniques do not rely on point-to-point distance estimation, but try to estimate the nodes' locations by computing quantities such as the angle of arrival (AOA).

Not all localization algorithms make use of the characteristics of the signals exchanged among sensors (such as amplitude, phase or delay). Some algorithms are based on the network connectivity information. A typical case in point is the family of algorithms that estimate the distance between nodes by the number of hops required for the packets to travel from source to destination nodes. These types of algorithms will not be further discussed in this paper.

2.1. Measurement Characteristics

The parameters to estimate are the coordinates of the sensor nodes. The assumption is that in order to compute these estimates,

physical quantities of the physical signals are measured, such as RSS, AOA, TOA or TDOA. The signals involved can be seismic, acoustic, RF, etc., or combinations thereof, depending on the applications. Naturally, the measurements as well as the anchors' positions are affected by a degree of uncertainty. Range and angle measurements are degraded by both time-varying errors (such as noise or interference) and environment-dependent errors. In the assumption that sensors are arranged in a mostly static network, environment-dependent errors are usually static in nature and due to the arrangements of objects (trees, buildings, furniture) in the area of operation of the sensor network. Time-varying errors can be averaged out by taking multiple measurements over time. The statistical characterization of these measurements is important to assess the estimation quality of the various algorithms. Usually a number of simplifying assumptions are made. For instance, measurements are typically assumed independent. In reality, correlations among measurements exist due to obstructions, or channel characteristics etc. Another assumption is the choice of statistical distributions used to characterize the different quantities. This choice is usually backed up by experiment, even though often there simply is not enough data to give a complete characterization.

In the following we consider the four quantities, RSS, AOA, TOA and TDOA individually.

2.1.1 RSS

RSS is the strength (voltage) of the signal measured by the receiver's received signal strength indicator (RSSI) in arbitrary units. Sometimes, it is reported as the squared magnitude of the received signal strength. In the 802.15.4 PHY standard, RSS can be measured by the link quality indication (LQI), used to report the signal strength of the received packet to the higher layers. RSS measurements are relatively inexpensive and simple to implement, and for this reason they are widely used in localization algorithms. Unfortunately they are also notoriously unpredictable.

The attenuation increases with distance (for instance, in free space signal power decreases with the second power of the distance), and is averaged over fast fading. Environment-dependent errors are due to obstructions that the signal must go through or diffract around. These effects are modeled as random. The ensemble mean power at distance d is usually modeled as

$$P(d) = P_0 - 10n_p \log \frac{d}{d_0}, \quad (1)$$

where P_0 is the received power in dBm measured at a reference distance d_0. The parameter n_p is the path-loss exponent and varies between two and four. The difference between the measured power and its ensemble average is assumed to be log-normal distributed (i.e., Gaussian, when expressed in decibels). This model is justified by measurements and analysis [14,16]. The standard deviation, expressed in dBm, σ_{dB}, is relatively constant with distance and varies between four and twelve. Because the standard deviation is constant with range when measured in decibels, the RSS-based range estimates have variance proportional to the actual range. RSS errors are for this reason multiplicative in nature, as opposed to the additive TOA errors. RSS measurements depend on the manufacturing and calibration of both the transmitter and the receiver, as well as transmitter's battery level, and this causes much of the unpredictability of these measurements.

2.1.2 AOA

Information regarding the angle of arrival may be obtained by using array processing techniques. Each sensor node has to be equipped with at least two sensing devices (e.g., microphones or antennas). The AOA is computed by calculating the difference in times of arrival or, in case of a narrowband signal, in phase difference. An alternative method makes use of difference in RSS at two directional antennas placed on the sensor, for points where the beam-patterns overlap. In either case, the need for multiple sensing devices at each node may cause exceeding cost and size. In other situations, such as environmental monitoring, acoustic sensor arrays may be a very reasonable option [4]. AOA is affected by noise and multipath. In particular, all the SNR of the line-of-sight (LOS) path may be low due to multipath, to

the point that the LOS path may not be the strongest arriving signal. In some cases, there may not even be a LOS path, due to obstructions. In this case the AOA estimate will be incorrect. Non-LOS (NLOS) situations may be detected by computing the variance of the received signal strength, which will be much higher in a NLOS situation. Even when a LOS path is present, its peak may be smeared by early-arriving multipath signals. AOA measurements are usually modeled as Gaussian, with ensemble mean equal to the true value and standard deviation, σ_α, on the order of two to six degrees. Because the estimate depends on the orientation of the sensors, it is sometimes required to add the unknown orientation to the parameters to estimate.

2.1.3 TOA

TOA is the time of arrival due to the propagation of the signal from transmitter to receiver and depends on the nature of the signal being sent. For instance, RF signals travel at 10^6 times the speed of sound. The travel time can be computed in a synchronized network, or from the round-trip time from transmitter to receiver and back. Time delays in the transmitter and receiver may also affect the estimate. As for AOA, measurements are affected by noise and multipath. In NLOS situations, the estimate will have a positive bias that needs to be corrected. Denser networks allow for more accurate TOA measurements between nearby nodes because of the higher power of the LOS component. Accuracy also increases with the bandwidth of the signals, even though this requires in turn a higher speed signal processing and potentially higher costs. The statistical model of choice for short-range TOA measurements is Gaussian, with mean equal to $d_{ij}/\upsilon_p+\mu_T$ and standard deviation equal to σ_T, where υ_p is the propagation velocity and d_{ij} is the inter-node distance. For instance, measurements of RF wideband signals have reported values for μ_T in the order of units to tens of nanoseconds and σ_T in the order of units of nanoseconds. In order to account for large observed errors, which point to a fatter-tailed distribution than Gaussian, mixture distributions have been proposed. A two-mode Gaussian mixture can be used to model the case of LOS and NLOS situations [13].

TOA estimates rely on network synchronization. In case the network cannot provide a level of synchronization accuracy suited to the signal (good accuracy for acoustic signals may not be sufficient for RF signals) [12], or in case of asynchronous networks, two-way TOA measurements are made, assuming that the internal delay between signal reception and re-transmission is known or can be estimated. Alternatively, the estimation problem can be augmented by including in the parameters to be estimated the sensor's clock biases.

2.1.4 TDOA

Time difference of arrival of the signal at two different antennas can be used for localization purposes, with the advantage that network synchronization information does not need to be communicated or estimated.

2.2. Performance Bounds

The "goodness" of an estimator to provide a good estimate of the desired parameters, in our case the node positions, can be computed by evaluating the lower bound to the estimator mean square error. Such lower bound takes the name of Cramér-Rao bound (CRB) [15]. This bound can be a very useful tool to the designer as well, not only because it sets a reference point to the estimator accuracy, but also because it is function of the design parameters, such as the network geometry, channel parameters, number of sensors and number of anchors, etc.

For instance, in the case of TOA, the CRB takes the form

$$\mathrm{CRB(TOA)} = \frac{1}{8\pi^2 B T_s f_c^2 \,\mathrm{SNR}}, \qquad (2)$$

where B is the signal bandwidth, f_c its central frequency and T_s its duration. The CRB shows that better accuracy is expected when the bandwidth or the transmitted power is increased. The CRB for the TOA-based location estimates does not change with a scaling of the network geometry, while the CRB of RSS or AOA-based estimates is actually proportional to the size of the system [11]. The CRB is proportional to σ_T for TOA and σ_α for AOA,

while it is proportional to d_{dB}/n_p for RSS. This result shows that a higher path-loss exponent makes the estimates more accurate, at the expense of higher transmitted power. From the CRB results (which of course are a first-order approximation because do not take into account secondary effect, such as channel parameter variation with path length), it appears that TOA-based estimates are less sensitive than AOA or RSS to inter-node distances, and are therefore preferable in sparser networks. AOA, TOA and TDOA can achieve higher accuracy than RSS. On the other hand, RSS-based estimates can be made more accurate in denser networks, with shorter inter-node distances, and are considerably cheaper than the alternatives. Whenever possible, it is advisable to use the different estimation algorithms in combination, especially when they seem to have complementary performance.

2.3 Estimators

Most estimators compute the sensors' positions based on an optimization criterion, such as the maximization of a likelihood function, or the minimization of a least-squares problem. The least-squares problem may in turn be weighted or regularized.

As examples of estimators, we present below an algorithm based on a Gauss-Newton search, both in its centralized and distributed formulation [8].

2.3.1 A Distributed Gauss-Newton Method for Node Localization

Suppose the K anchor sensors at coordinates $\mathbf{a}_k$, $k = 1,\ldots, K$, are placed in known locations, while the $\mathbf{x}_i$, $i = 1,\ldots, N$, sensors at coordinates d_{ij}, have unknown locations. Let d_{ij} be the measured distance between nodes i and j and assume this distance has been obtained through RSS or TOA measurements. The problem is then to solve the following nonlinear least-squares problem with respect to the node locations, $\mathbf{x}_i$, $i = 1,\ldots, N$,

$$\min_{\{\mathbf{x}_i\}} F(\{\mathbf{x}_i\}) = \min_{\{\mathbf{x}_i\}} \left\{ \sum_{i,j} |r_{ij}|^2 + \sum_{i,k} |s_{ik}|^2 \right\}, \quad (3)$$

where $F(\{x_i\})$ is the global cost function, $r_{ij,} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 - d_{ij}^2$ represents the residual error in the placement of pairs of sensors i and j and $s_{ik,} = \|\mathbf{x}_i - \mathbf{a}_k\| - d_{ik}^2$ is the error in the placement of sensor i with respect to anchor k. The calculation is restricted to only the pairs of nodes within a given radius from each other (the so-called radio range). This (centralized) problem can be solved by means of a Gauss-Newton method. This algorithm can be turned into a distributed algorithm by defining the cost function at node i as

$$F_i = \sum_{j \to i} |r_{ij}|^2 + \sum_{k \to i} |s_{ik}|^2, \quad (4)$$

where $j \to i$ and $k \to i$ represent the indices of the nodes and anchors that are neighbors (in the radio range) of node i. It is assumed that the nodes are ordered numerically, as indicated by their indices. Each node will try to minimize the function F_i, assuming that all nodes adjacent to node i and index $j < i$ have been updated in the current iteration, and all the nodes adjacent to node i and index $j > i$ have been updated in the previous iteration. In this way, the distributed, but sequential, algorithm is performing the same operations as the centralized algorithm above. Each local minimization can be performed by approximating the problem with a linear least-squares problem and then using the Gauss-Newton method.

The algorithm can once more be changed in such a way as to become distributed and parallel. The problem being solved is now a first-order approximation of (4), and can be written as a linear least-squares problem as follows

$$\min_{\{\mathbf{p}_i\}} \left\{ \sum_{j \to i} |r_{ij} - \nabla r_{ij}^T \mathbf{p}_i|^2 + \sum_{k \to i} |s_{ij} - \nabla s_{ij}^T \mathbf{p}_i|^2 \right\}. \quad (5)$$

If $\mathbf{p}_i$ is the solution to the problem above, then the update of the coordinates of sensor i is given by

$$\mathbf{x}_i := \mathbf{x}_i + \alpha_i \hat{\mathbf{p}}_i, \quad (6)$$

for a properly chosen step size α_i. The calculations can be performed in parallel in all nodes. In general, the parallel algorithm requires a slightly higher number of iterations than

the sequential algorithm for the same accuracy, and thus the computation complexity is slightly increased. Nonetheless the parallel algorithm has much faster processing time because all computations occur simultaneously. The step sizes can be chosen so as to make the functions F_i non increasing with the iteration index. Although it is possible that some nodes will require at times $\alpha_i = 0$, in all the simulations performed to date the step size has always been strictly positive and the functions F_i strictly decreasing from iteration to iteration.

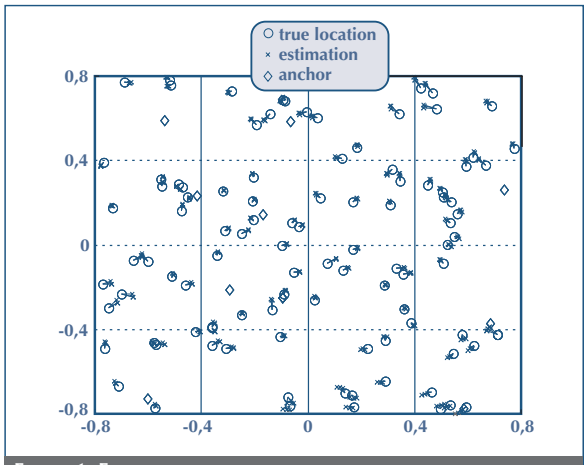

FIGURE 1: ESTIMATION RESULTS OF THE CENTRALIZED, SEQUENTIAL AND PARALLEL ALGORITHMS.

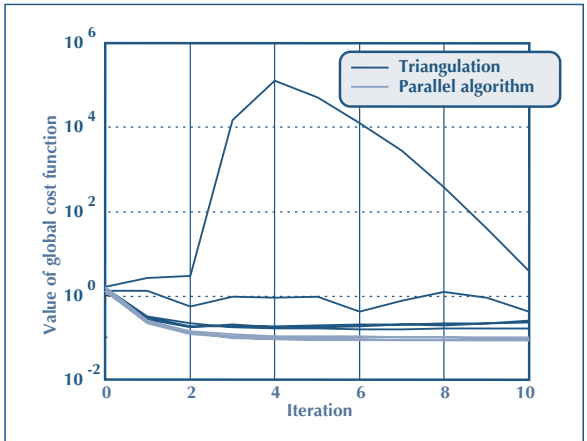

FIGURE 2: ESTIMATION ERROR VERSUS ITERATION FOR DIFFERENT ALGORITHMS. THE DIFFERENT CURVES REPRESENT 5 DIFFERENT RANDOM REALIZATIONS.

Fig. 1 shows an example of $N = 100$ sensors and $K = 10$ anchors placed randomly in a 1.6 X 1.6 region where the radio range is 0.35 and the noise factor is 0.1. The starting points are chosen as random perturbations from the true sensor locations, where the perturbation is given by a uniformly distributed random variable in [–0.1, 0.1]. The estimation results of the centralized, sequential and parallel algorithms are given in Fig. 1, represented by "x". The true sensor locations are given by the "o". Anchors are denoted by "◊". It can be seen from the figure that distributed algorithms converge to almost the same results as the centralized Gauss-Newton algorithm.

The parallel algorithm has been compared to the distributed triangulation (or one-hop multilateration) method in [19] and [20], where the node locations are calculated in two steps: firstly an approximate distance is calculated by computing the shortest hop distance and using knowledge of an average hop size and secondly a more accurate distance estimate is obtained by iteratively solving a local least squares problem. Fig. 2 shows comparisons of the convergence behavior of the parallel algorithm and the distributed triangulation method over 5 random realizations.

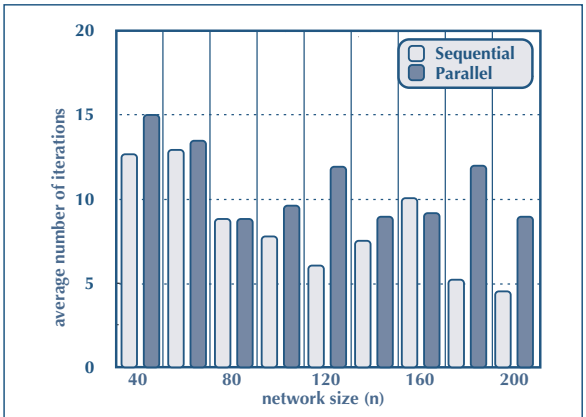

FIGURE 3: AVERAGE NUMBER OF ITERATIONS.

We subsequently increase the size of the network, N, while keeping the density of both sensors and anchors unchanged, that is, N sensors and $N/10$ anchors located randomly in a $1.6\sqrt{N/100}$

x1.6$\sqrt{N/100}$ region. Fig. 3 shows the average numbers of iteration over 20 realizations for both sequential and parallel algorithms to reach a given accuracy (within ±0.01 from the true location). The radio range is 0.35 and the noise factor is set to zero in all cases. It can be seen that the sequential algorithm requires a few more iterations.

On the other hand, the total processing time in Fig. 4 is much larger for the sequential algorithm. This is because the steps $\mathbf{p}_i$ in the sequential algorithm must be computed sequentially, while in the parallel algorithm they can be computed simultaneously (for simplicity, we ignore the processing time for finding the step length).

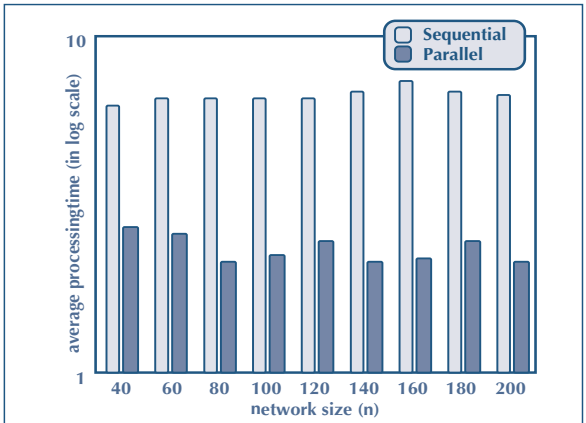

FIGURE 4: AVERAGE PROCESSING TIME (IN LOG SCALE).

Finally, we present some empirical results on the communication energy cost. We use the model given by [17] that the propagation loss is proportional to the fourth power of distance. For centralized algorithms, each sensor only needs to transmit to the central processor once.

However, as the size of the network increases or, equivalently, the region increases, each sensor will need more energy to transmit the information to a central processor (we assume the central processor lies at location (0,0)). In distributed algorithms, each sensor only communicates to the neighboring sensors. Thus the total energy cost depends linearly on the total number of sensors and number of iterations required.

Fig. 5 shows the average energy consumption over 20 realizations of different algorithms. It can be seen that when the size of the network is large, distributed algorithms have a reduced energy consumption.

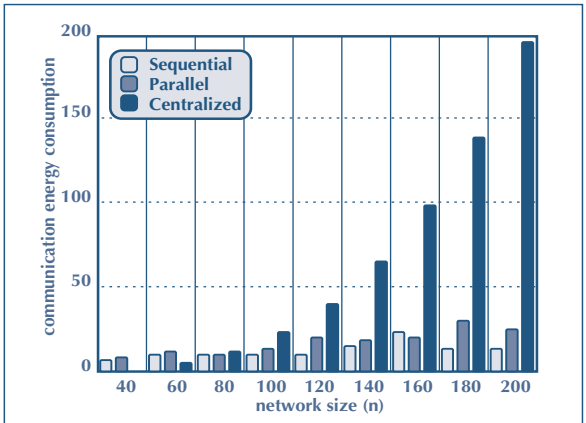

FIGURE 5: AVERAGE ENERGY CONSUMPTION DUE TO COMMUNICATION.

3. SOURCE LOCALIZATION

Distributed sensor networks (DSNs) have been proposed for a wide range of applications. The network may be used to monitor an area for military, environmental, and manufacturing purposes. In these and other applications, an important task is to perform source localization by estimating the location of one or more stationary or moving objects. The object(s) may be radiating one more more form(s) of energy(ies). Sensor nodes in the DSN, using one or more modality of reception can collect the signature(s) of the object(s) and perform source localization.

3.1 Trilateration/Multilateration

Trilateration is conceptually the simplest method to perform source localization of a single source based on range estimation. Here we assume each sensor node can estimate the range from the source to itself. There are various methods on how this range estimate can be made in practical scenarios. In order to find the location of the source, we need to reference it with respect to some coordinate system. Furthermore, one must assume that the

locations of the sensor nodes be known with respect to the same coordinate system. Again, the issue of determining the locations of the sensor nodes have been discussed in some details in Sec. 2.

In trilateration, consider all the range-sensing nodes and the source are situated on the x–y plane. Range-sensing nodes A, B, and C are located at (x_A, y_A), (x_B, y_B), (x_C, y_C) as shown in Fig. 6. If node A estimates a source at range d_A and B node estimates that source at range d_B, then the intersection of these two circles yields two possible ambiguous source locations marked by S and S'. Similarly, if node C estimates a range of d_C, then the intersection between node B and node C yields two possible ambiguous source locations S and S'', while the intersection of node C and node A yields two possible ambiguous source locations of S and S'''. However, the true location of the source is at the intersection of all three circles at S. Of course, in the presence of noisy estimations on d_A, d_B, d_C, the three circles will not intersect at a single point to yield a unique location S. In practice, denote the source coordinates (x, y), then d_A, d_B, d_C satisfy the equations of

$$(x-x_A)^2+(y-y_A)^2=d_A^2,\ (x-x_B)^2+(y-y_B)^2=d_B^2,$$
$$(x-x_C)^2+(y-y_C)^2=d_C^2. \quad (7)$$

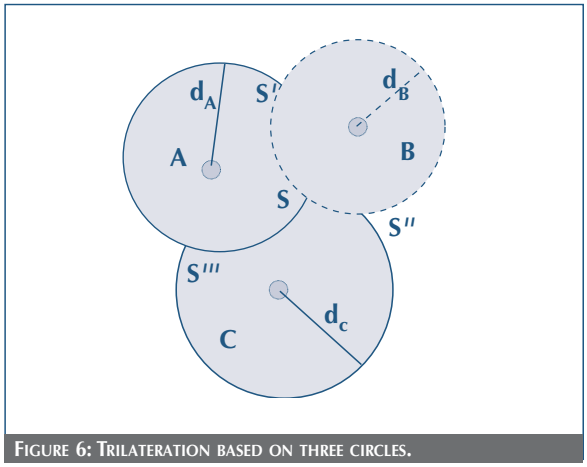

FIGURE 6: TRILATERATION BASED ON THREE CIRCLES.

The solution $[x, y]^T$ of (7) can be written as the least-squares (LS) solution of

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2(x_A-x_C) & 2(y_A-y_C) \\ 2(x_B-x_C) & 2(y_B-y_C) \end{bmatrix}^{-1} \begin{bmatrix} x_A^2-x_C^2-y_A^2-y_C^2+d_C^2-d_A^2 \\ x_B^2-x_C^2-y_B^2-y_C^2+d_C^2-d_B^2 \end{bmatrix}. \quad (8)$$

In multilateration, N range-sensing nodes with known locations (x_1, y_1), (x_2, y_2), …, (x_N, y_N) are used instead of three nodes in trilateration. Their distances from the unknown source are denoted by $d_1, d_2, \ldots, d_N$. Then corresponding to (7), we have

$$\begin{aligned} &(x-x_1)^2+(y-y_1)^2=d_1^2, \\ &(x-x_2)^2+(y-y_2)^2=d_2^2, \\ &\vdots \\ &(x-x_N)^2+(y-y_N)^2=d_N^2. \end{aligned} \quad (9)$$

The LS solution of $\mathbf{A}\mathbf{r}_s = \mathbf{b}$ corresponding to (8) is now given by

$$\mathbf{r}_s = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}, \quad (10)$$

where

$$\mathbf{A} = \begin{bmatrix} 2(x_1-x_N) & 2(y_1-y_N) \\ \vdots & \vdots \\ 2(x_{N-1}-x_N) & 2(y_{N-1}-y_N) \end{bmatrix}, \quad \mathbf{r}_s = \begin{bmatrix} x \\ y \end{bmatrix},$$

$$\mathbf{b} = \begin{bmatrix} x_1^2-x_N^2+y_1^2-y_N^2+d_N^2-d_1^2 \\ \vdots \\ x_{N-1}^2-x_N^2+y_{N-1}^2-y_N^2+d_N^2-d_{N-1}^2 \end{bmatrix}. \quad (11)$$

2. LINEAR INTERSECTION

Consider now the situation where two range-sensing nodes, A and B compute their distances from the source S as d_A and d_B, and assume it is known that the points A, B and S are in a clockwise sequence on the x–y plane. Also assume that the distance between A and B is known and equal to d. By simple geometric considerations, one can define the two quantities

$$\alpha = \frac{d_A^2 - d_B^2 + d^2}{2d^2}, \beta = \sqrt{\frac{d_A^2}{d^2} - \alpha^2}, \quad (12)$$

and calculate the source location as $x = x_A + \alpha(x_B - x_A) + \beta(y_B - y_A)$, $y = y_A + \alpha(y_B - y_A) + \beta(x_B - x_A)$. Of course this formulation assumes that the ranges are computed with no error. In case more node pairs are available, all the computed source locations can be combined in a weighted average, where the weights can be chosen according to an assumed statistical distribution of the errors [2].

3. MAXIMUM-LIKELIHOOD SOURCE LOCALIZATION

The general situation where M sources radiate a broadband signal and N synchronized sensors are used as in a sensor array to compute the source location estimates can be described by the following model

$$x_r(n) = \sum_{m=1}^{M} a_r^{(m)} s_0^{(m)}(n - t_r^{(m)}) + w_r(n),\ n = 1,\ldots,L,\ r = 1,\ldots,N, \quad (13)$$

where $a_r^{(m)}$ is the signal level of the mth source at the rth sensor, $s_o^{(m)}$ is the source signal, $t_r^{(m)}$ is the time delay and $w_r(n)$ is zero-mean i.i.d. Gaussian noise with variance σ^2. In the frequency domain the array signal model is given by

$$\mathbf{X}(k) = \mathbf{D}(k)\mathbf{S}_0(k) + \mathbf{Z}(k), \quad (14)$$

where $\mathbf{D}(k)$ is the steering matrix at frequency bin k and $\mathbf{Z}(k)$ is the noise spectrum vector. Given certain assumptions on the noise distribution, the source estimation problem can be cast into a the problem of maximizing the likelihood function, which in turn is equivalent to the following maximization

$$\max_{\mathbf{r}_s} J(\mathbf{r}_s) = \max_{\mathbf{r}_s} \sum_k \left\| \mathbf{P}(k, \mathbf{r}_s)\mathbf{X}(k) \right\|^2, \quad (15)$$

where $\mathbf{r}_s$ is the source location vector, $\mathbf{P}(k,\mathbf{r}_s) = \mathbf{D}(k)\ \mathbf{D}^+(k)$ is a projection operator, and $^+$ denotes pseudo-inverse. The summation does not have to encompass all energy bins, but may be limited to the frequencies that exhibit higher received signal energy. In practice, due to the edge effect in the finite-length FFT operation used to obtain the frequency-domain data, the above described method is denoted as the Approximate Maximum-Likelihood (AML) method. Fig. 7 shows the simulation result of the near-field normalized $J(\mathbf{r}_s)$ metric of a vehicle source having high values about its location inside the convex hull of an array of five sensors. Thus, for near-field scenarios, this metric is capable of estimating the location of the source. Fig. 8 shows the simulation result of the far-field normalized $J(\mathbf{r}_s)$ metric of a vehicle source having high values in the angular sector in the direction of the source. Thus, for far-field scenarios, this metric is capable of estimating direction-of-arrival (DOA) of the source. The details of this algorithm and these and other simulation and field-measured estimated results can be found in [3]. Note that the set of parameters to estimate can be augmented with the propagation velocity and the noise variance, if these are unknown. This algorithm requires a fairly high computational power, mostly due to the maximization of a function, $J(\mathbf{r}_s)$, that is usually non-convex and displays a large number of local maxima. Moreover all the sensors that participate in the estimate need to be synchronized and be fully collaborative.

The algorithm is centralized, i.e., the relevant information has to be sent to a central node provided with sufficient computational power. These are issues in networks of low-power sensors, but may not be a problem in environment monitoring networks where nodes are more complex and have higher available energy supplies.

3.4 TDOA

Consider the waveform emitted by a source and then received by a group of synchronized passive sensors with different time delays due to different propagation paths. By using these sensor data, time-difference of arrivals (TDOAs) can be used to perform source localization.

Denote the source location in Cartesian coordinates by $\mathbf{r}_s = [x_s, y_s, z_s]^T$ and the rth sensor location by $\mathbf{r}_r = [x_r, y_r, z_r]^T$. Without

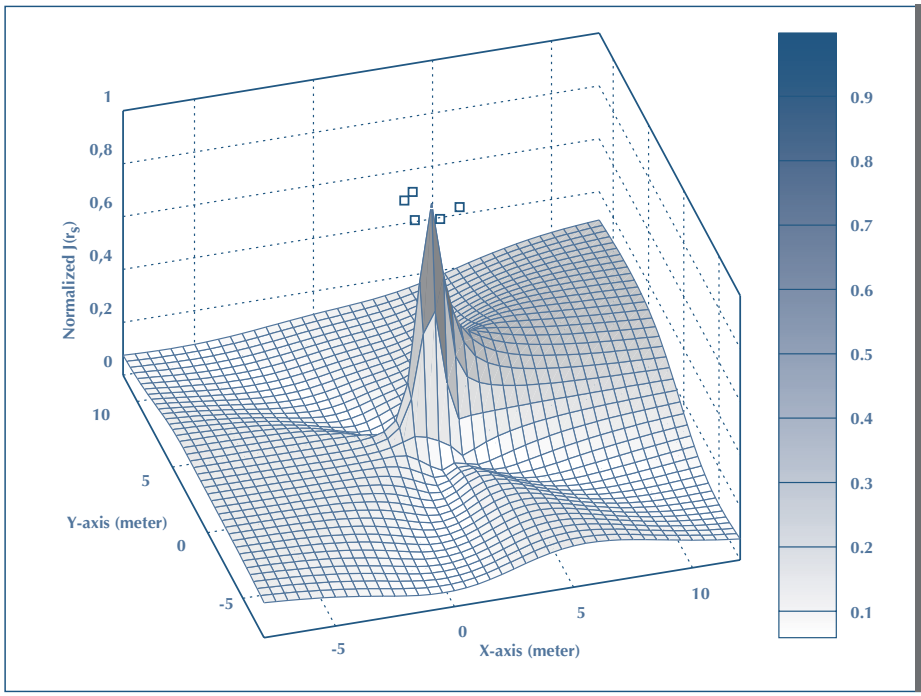

FIGURE 7:
PLOT OF $\mathbf{J}(\mathbf{r}_s)$ VALUES OF A NEAR-FIELD SOURCE CAPABLE OF SOURCE LOCALIZATION.

loss of generality, we choose $r = 1$ as the reference sensor for differential time-delays. Let the reference sensor be the origin of the coordinate system for simplicity. The speed of propagation v in this formulation can also be estimated from the data. In some problems, v may be considered to be partially known (e.g., acoustic applications) while in others it may be considered to be unknown (e.g., seismic applications). The differential time-delays for N sensors satisfy

$$t_{r1} = t_r - t_1 = \frac{\| \mathbf{r}_s - \mathbf{r}_r \| - \| \mathbf{r}_s - \mathbf{r}_1 \|}{v},$$

for $r = 2, \ldots, N$. This is a set of $N - 1$ nonlinear equations which makes finding its solution $\mathbf{r}_s$ non-trivial.

However, the above equations can be reformulated as

$$\| \mathbf{r}_s - \mathbf{r}_r \|^2 - \| \mathbf{r}_s \|^2 = \| \mathbf{r}_r \|^2 - 2 (x_s x_r + y_s y_r + z_s z_r).$$

The left hand side of (17) is equivalent to

$$(\| \mathbf{r}_s - \mathbf{r}_r \| - \| \mathbf{r}_s \|) (\| \mathbf{r}_s - \mathbf{r}_r \| + \| \mathbf{r}_s \|) = v t_{r1} (2 \| \mathbf{r}_s \| + v t_{r1}).$$

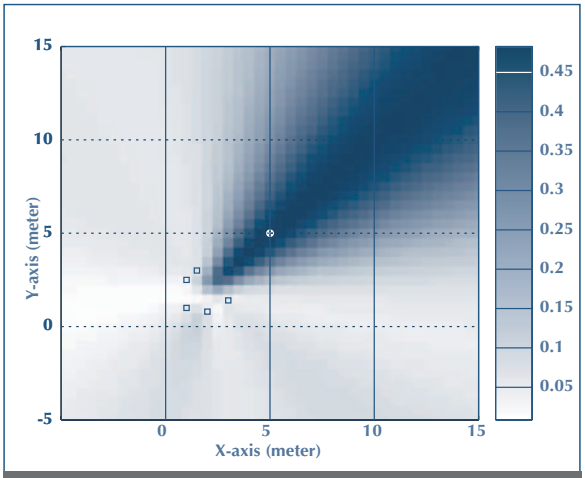

FIGURE 8: PLOT OF $\mathbf{J}(\mathbf{r}_s)$ VALUES OF A FAR-FIELD SOURCE CAPABLE OF DOA SOURCE ESTIMATION.

in the case of $\mathbf{r}_1 = 0$. Upon combining both expressions, we have the following linear relation

$$\mathbf{r}_r^T \mathbf{r}_s + v t_{r1} \| \mathbf{r}_s \| + v^2 t_{r1}^2 / 2 = \| \mathbf{r}_r \|^2 / 2,$$

for the rth sensor. With N sensors, we formulate the least-squares solution by putting $N - 1$ linear equations into the following matrix form

$$\mathbf{A}\mathbf{y} = \mathrm{b}, \tag{20}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{r}_2^T & t_{21} & t_{21}^2/2 \\ \mathbf{r}_3^T & t_{31} & t_{31}^2/2 \\ \vdots & \vdots & \vdots \\ \mathbf{r}_N^T & t_{N1} & t_{N1}^2/2 \end{bmatrix},$$

$$\mathbf{y} = \begin{bmatrix} \mathbf{r}_s \\ v\left\|\mathbf{r}_s\right\| \\ v^2 \end{bmatrix}, \quad \mathbf{b} = \frac{1}{2}\begin{bmatrix} \left\|\mathbf{r}_2\right\|^2 \\ \left\|\mathbf{r}_3\right\|^2 \\ \vdots \\ \left\|\mathbf{r}_N\right\|^2 \end{bmatrix}. \tag{21}$$

For 3-D uncertainty, the dimension of $\mathbf{A}$ is $(N - 1)$ X 5. In the case of six or more sensors, the pseudo-inverse of matrix $\mathbf{A}$ is given by

$$\mathbf{A}^+ = (\mathbf{A}^T\mathbf{A})^{-1}\,\mathbf{A}^T. \tag{22}$$

The LS solution for the unknown vector can be given by $\mathbf{y} = \mathbf{A}^+\mathbf{b}$. The source location estimate is given by the first three elements of $\mathbf{y}$ and the speed of propagation estimate is given by the square-root of the last element of $\mathbf{y}$.

In the three dimensional case, there are five unknowns in $\mathbf{y}$. To obtain an overdetermined solution, we need at least five independent equations, which can be derived from the data of six sensors. However, placing sensors randomly does not provide much assurance against ill-conditioned solutions. The preferred approach would be to use seven or more sensors, yielding six or more relative delays, and to perform a least squares fitting of the data. In the 2-D problem, the minimum number of sensors can be reduced by one. If the propagation speed is known, then the minimum number of sensors can be further reduced by one.

Notice in the unknown vector $\mathbf{y}$ of (29) that the speed of propagation estimate can also be given by

$$\hat{v} = \frac{\widehat{v\left\|\mathbf{r}_s\right\|}}{\left\|\hat{\mathbf{r}}_s\right\|}, \tag{23}$$

using the fourth and the first three elements of $\mathbf{y}$. To exploit this relationship, we can add another nonlinear constraint to ensure equivalence between the speed of propagation estimates from the fourth and the fifth elements. By moving the fifth element of $\mathbf{y}$ to the other side of the equation, we can re-write (20) to the following

$$\mathbf{A}\mathbf{y} = \mathbf{b} + v^2\,\mathbf{d} \tag{24}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{r}_2^T & t_{21} \\ \mathbf{r}_3^T & t_{31} \\ \vdots & \vdots \\ \mathbf{r}_N^T & t_{N1} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{r}_s \\ v\left\|\mathbf{r}_s\right\| \end{bmatrix}, \tag{25}$$

$$\mathbf{b} = \frac{1}{2}\begin{bmatrix} \left\|\mathbf{r}_2\right\|^2 \\ \left\|\mathbf{r}_3\right\|^2 \\ \vdots \\ \left\|\mathbf{r}_N\right\|^2 \end{bmatrix}, \quad \mathbf{d} = -\frac{1}{2}\begin{bmatrix} t_{21}^2 \\ t_{31}^2 \\ \vdots \\ t_{N1}^2 \end{bmatrix}. \tag{26}$$

In this case, the dimension of $\mathbf{A}$ is $(N - 1)$ X 4 for 3-D uncertainty. The constrained least-squares (CLS) solution for the unknown vector can be given by $\mathbf{y} = \mathbf{A}^+\mathbf{b} + v^2\,\mathbf{A}^+\mathbf{d}$. Define $\mathbf{p} = \mathbf{A}^+\mathbf{b}$ and $\mathbf{q} = \mathbf{A}^+\mathbf{d}$. The source location and speed of propagation estimates can be given by

$$\begin{aligned} x_s &= p_1 + v^2 q_1, \\ y_s &= p_2 + v^2 q_2, \\ z_s &= p_3 + v^2 q_3, \\ v\,\|\,\mathbf{r}_s\,\| &= p_4 + v^2 q_4, \end{aligned} \tag{27}$$

where p_i and q_i are the ith entry of $\mathbf{p}$ and $\mathbf{q}$, respectively. The number of unknowns appears to be five, but the five unknowns only contribute four degrees of freedom due to the following nonlinear relationship

$$\| \mathbf{r}_s \|^2 = x_s^2 + y_s^2 + z_s^2. \quad (28)$$

By substituting (45) into (46), the following third order constraint equation results

$$\alpha(v^2)^3 + \beta(v^2)^2, \gamma(v^2) + \delta = 0, \quad (29)$$

where

$$\begin{aligned} \alpha &= q_1^2 + q_2^2 + q_3^2, \\ \beta &= (p_1q_1 + p_2q_2 + p_3q_3) - q_4^2, \\ \gamma &= q_1^2 + q_2^2 + q_3^2 - 2p_4q_4, \\ \delta &= -p_4^2. \end{aligned} \quad (30)$$

In practice, there are various methods to estimate the TDOA among the sensors. A blind beamforming method was proposed by [5]. This method is based on the use of the maximum-power collection criterion to obtain array weights from the dominant singular vector or eigenvector associated with the largest singular value or eigenvalue of the space-time sample correlation matrix. This approach not only collects the maximum power of the dominant source, but provides some rejection of other interferences and noise. Theoretical justification of this approach uses a generalization of Szegő's theory of the asymptotic distribution of eigenvalues of the Toeplitz form.

The relative phase information among the weights yields the relative propagation time delays from the dominant source to the array sensors. These TDOA estimates can be used to perform source localization as well as detection, signal enhancement, DOA estimation, and delay-steered beamforming. More details on the TDOA estimation are given in [5] and [6].

A circular array with ten microphones located on a circle of four-feet radius (at each hour of the clock except for the ones at seven and ten o'clocks) and one microphone at the center was used to estimate the TDOA values of a moving vehicle from measured data supplied by ARL (the Army Research Lab). The data sampling rate was 1 kHz and approximately two minutes of data of the vehicle at the far field were used. The TDOAs of the ten microphones on the circle relative to the center microphones were estimated and shown in Fig. 9. Fig. 10a shows the estimated DOA and Fig. 10b shows the estimated speed of propagation using the TDOA-CLS method.

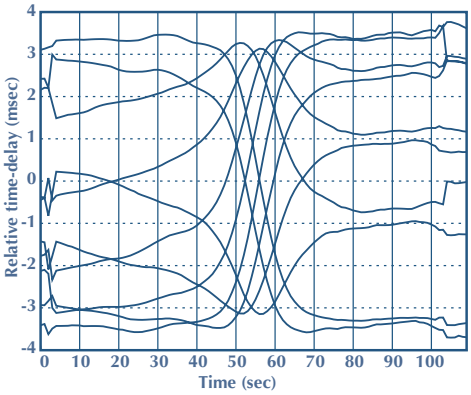

FIGURE 9: RELATIVE TIME-DELAY ESTIMATIONS OF 10 MICROPHONES RELATIVE TO THE CENTER MICROPHONE.

3.5. RSS

The Received Signal Strength (RSS) method is a conceptually simple energy-based source localization procedure. An acoustic or a RF source in free space radiating omni-directionally will attenuate at a rate inversely proportional to the square of the distance. In practice, we can assume an isotropic exponential attenuation model of $e_i(t) = s(t)/\|\mathbf{r}(t) - \mathbf{r}_i\|^{\alpha}$. Here, $e_i(t)$ is the energy value at the th sensor located at $\mathbf{r}_i$ whose location is assumed to be known, $\mathbf{r}(t)$ is the unknown coordinate of the source, $s(t)$ is the unknown source waveform, and α is the decay exponent which is assumed to be known or can be estimated for a given scenario. In ideal free space, $\alpha = 2$. In using this approach, the ratio of $e_i(t)/e_j(t)$ is computed for all pairs of sensors.

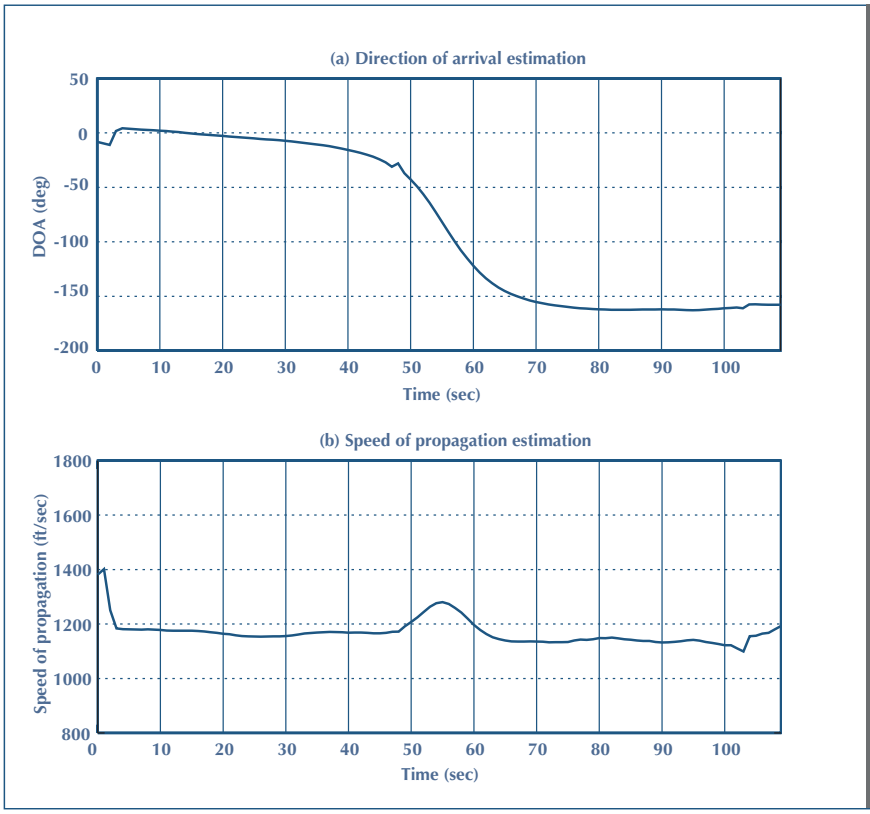

FIGURE 10: (A) DOA AND (B) SPEED OF PROPAGATION ESTIMATION USING THE TDOA-CLS METHOD.

In this ratio, the unknown $s(t)$ waveform is eliminated and will not be needed in further processing. Each of these ratios defines a circle on which $\mathbf{r}(t)$ may reside. In the absence of noise (or at high SNR), with N sensors, only $N-1$ out of the total $N(N-1)/2$ of these ratios are independent. For noisy situations, more than $N-1$ of these ratios need to be used to determine $\mathbf{r}(t)$. Additional details for acoustic sensors are given in [9] and [10] and for EM sensors are given in [11].

3.6 DOA

In Sec. 3.1 and Sec. 3.2, two different methods can be used to estimate the DOA in an array system. In this section, we describe some experimental results on using the DOAs to perform source localization in the far-field. In [7], we built a wireless acoustical testbed using the Compaq iPAQ 3760 Pocket PCs as the testbed nodes. Each iPAQ has a microphone, an ADC, a codec, and a 802.11b wireless card. The combination of the COTS hardware and open-source operating system was used to implement both the TDOA-CLS and the AML DOA estimation algorithms. The choice of these devices, which admittedly may be quite different from the nodes of a sensor network, was motivated by the desire to set up a working testbed that would allow the generation of significant data as well as the development and testing of localization tracking algorithms. Fig. 11 shows a configuration with three linear subarrays with each subarray having three iPAQ microphones to obtain three DOA estimations. Fig. 12 shows the results of six source localizations by using cross-bearings of the three DOA values using both the TDOA-CLS and AML methods for a music source. Next, consider a configuration with four square subarrays, with each subarray having four iPAQs. Fig. 13 shows

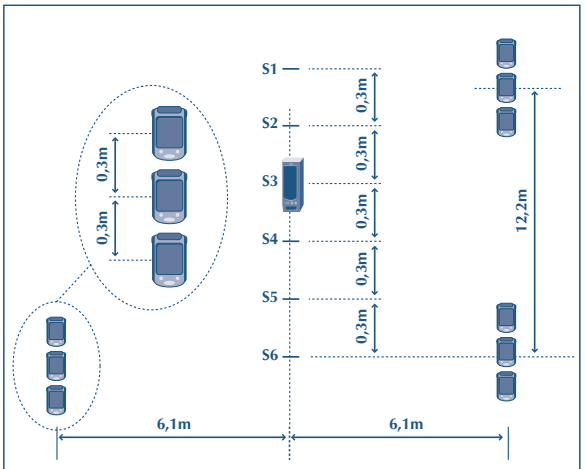

FIGURE 11: CONFIGURATION OF THREE LINEAR SUBARRAYS.

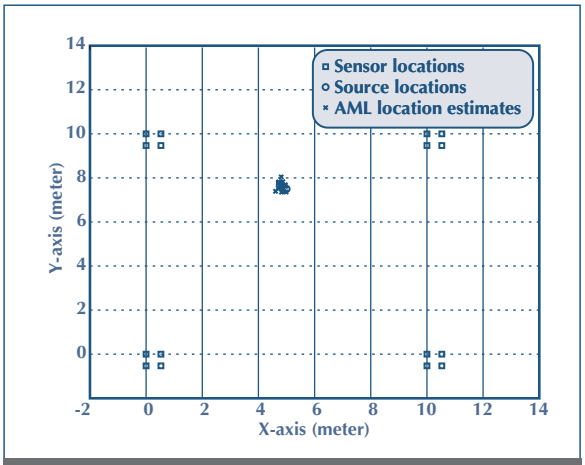

FIGURE 13: LOCALIZATION BY CROSS-BEARING OF DOA ESTIMATES OF A VEHICLE SOURCE.

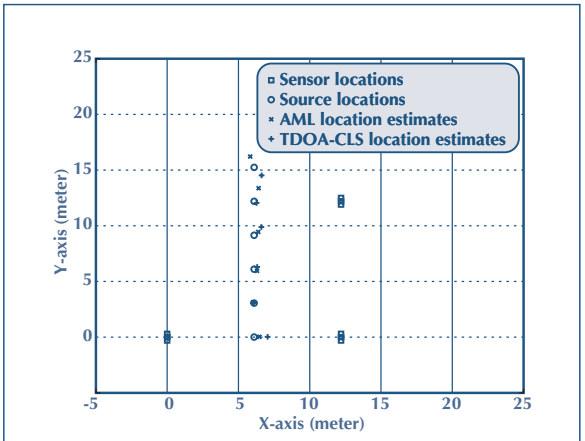

FIGURE 12: LOCALIZATION BY CROSS-BEARING OF DOA ESTIMATES OF A MUSIC SOURCE AT DIFFERENT LOCATIONS.

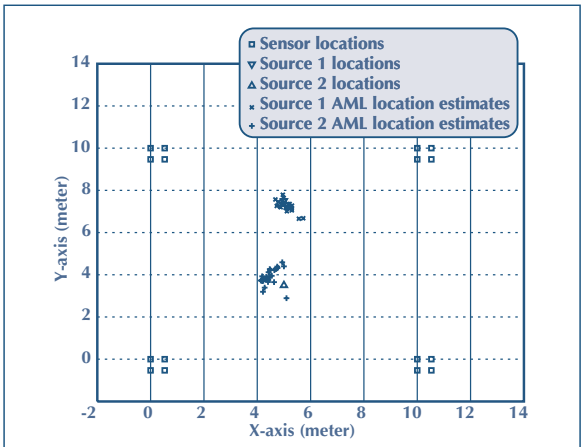

FIGURE 14: LOCALIZATION BY CROSS-BEARING OF DOA ESTIMATES OF TWO SOURCES.

the cross-bearings of the four DOAs to estimate the location of a speaker playing a music sound. Then two speakers, one playing the vehicle sound and another playing the music sound simultaneously were used. Fig. 14 shows that the AML method is capable of performing source localization of both speakers.

4. CONCLUSIONS

Sensor networks are being deployed in a wide range of applications. In order for the sensed data to be relevant, it is crucially important that the sensors' geographical positions, i.e., their locations, be known or estimated. If the nodes cannot be placed with certainty, or in case the nodes might be displaced or moved during the course of their operation, then node localization techniques must be employed. Location estimation may be performed by one centralized processor, or by the individual nodes in a distributed fashion. The function of sensors is often to detect the occurrence of an event, which in most cases is the presence of a source within their sensing range. The sensor

network performs this task by applying techniques of source localization.

In this paper we have described the problems of source and node localization in sensor networks. The literature on both these topics is very broad and by no means have we explored all possible facets of the localization problem. We have covered in some detail a few selected algorithms in which we have been personally involved. We have also explored some issues related to the statistical characterization of measurements and their impact on the quality of the location estimates.

REFERENCES

[1] P. Biswas and Y. Ye, "Semidefinite Programming for Ad-hoc Wireless Sensor Network Localization," Info. Proc. in Sensor Networks, pp. 552–553, 2004.

[2] M.S. Brandstein, J.E. Adcock, and H.F. Silverman, "A Closed-Form Location Estimator for Use with Room Environment Microphone Arrays," IEEE Tr. Speech and Audio Proc., vol. 15, no. 1, pp. 45–50, January 1997.

[3] J. Chen, R.E. Hudson, and K. Yao, "Maximum-Likelihood Source Localization and Unknown Sensor Location Estimation for Wideband Signals in the Near-Field," IEEE Trans. Signal Processing, vol. 50, pp. 1843–1854, Aug. 2002.

[4] J.C. Chen, K. Yao, and R.E. Hudson, "Source Localization and Beamforming," IEEE S.P. Mag., vol. 19, no. 2, pp. 30–39, March 2002.

[5] K. Yao, R.E. Hudson, C.W. Reed, D. Chen, and F. Lorenzelli, "Blind Beamforming on a Randomly Distributed Sensor Array System," IEEE Jour. on Sel. Areas in Communications, vol. 16, pp. 1555-1567, Oct. 1998.

[6] J.C. Chen, K. Yao, T.L. Tung, C.W. Reed, and D. Chen, "Source Localization and Tracking of a Wideband Source Using a Randomly Distributed Beamforming Sensor Array," Inter. Jour. of High Performance Computing Applications, vol. 16, pp. 259-272, Fall 2002.

[7] J.C. Chen, L. Yip, J. Elson, H. Wang, D. Maniezzo, R.E. Hudson, K. Yao, and D. Estrin, "Coherent Acoustic Array Processing and Localization on Wireless Sensor Networks," Proc. of the IEEE, vol. 91, pp. 1154-1185, Aug. 2003.

[8] B.H. Cheng, R.E. Hudson, F. Lorenzelli, L. Vandenberghe, and K. Yao, "Distributed Gauss-Newton Method for Node Localization in Wireless Sensor Networks," Signal Proc. Advances in Wireless Comm., pp. 915–919, 2005.

[9] D. Li, K.D. Wong, Y.H. Hu, and A.M. Sayeed, "Detection, Classification, and Tracking of Targets," IEEE Signal Processing Magazine, pp. 30-29, March 2002.

[10] D. Li and Y.H. Hu, "Energy-Based Collaborative Source Localization Using Acoustic Microsensor Array," Eurasip Jour. on Applied Signal Processing, pp. 321-337, March 2003.

[11] N. Patwari, J.N. Ash, S. Kyperountas, A.O. Hero III, R.L. Moses, and N.S. Correal, "Locating the Nodes," IEEE Signal Processing Magazine, pp. 54-69, July 2005.

[12] L. Girod, V. Bychkovskiy, J. Elson, and D. Estrin, "Locating Tiny Sensors in Time and Space: A Case Study," IEEE Intl. Conf. Comp. Design, pp. 214–219, 2002.

[13] F. Gustafsson and F. Gunnarson, "Mobile Positioning Using Wireless Networks," IEEE S.P. Mag., vol. 22, no. 4, pp. 41–53, July 2005.

[14] H. Hashemi, "THE INDOOR RADIO PROPAGATION CHANNEL," Proc. IEEE, vol. 81, no. 7, pp. 943–968, July 1993.

[15] S.M. Kay, **Fundamentals of Signal Processing — Estimation Theory**, Englewood Cliffs, NJ, Prentice Hall 1993.

[16] Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda, "FIELD STRENGTH AND ITS VARIABILITY IN VHF AND UHF LAND-MOBILE RADIO SERVICE," Rev. Elec. Commun. Lab., vol. 16, pp. 9–10, 1968.

[17] G. Pottie and W. Kaiser, "WIRELESS INTEGRATED NETWORK SENSORS," Comm. of the ACM, vol. 43, no. 5, pp. 51–58, May 2000.

[18] V. Raghunathan, C. Schurgers, S. Park, and M.B. Srivastava, "ENERGY-AWARE WIRELESS MICROSENSOR NETWORKS," IEEE S.P. Mag., vol. 19, no. 2, pp. 40–50, March 2002.

[19] C. Savarese, J. Rabaey, and K. Langendoen, "ROBUST POSITIONING ALGORITHMS FOR DISTRIBUTED AD-HOC NETWORKS OF SENSORS," USENIX Technical Annual Conf., June 2002.

[20] A. Savvides, H. Park, and M.B. Srivastava, "THE BITS AND FLOPS OF THE -HOP MULTILATERATION PRIMITIVE FOR NODE LOCALIZATION PROBLEMS," ACM Mobile Net. and App., vol. 8, no. 4, pp. 443–451, 2003.

[21] G. Sun, J. Chen, W. Guo, and K.J.R. Liu, "SIGNAL PROCESSING TECHNIQUES IN NETWORK-AIDED POSITIONING," IEEE S.P. Mag., vol. 22, no. 4, pp. 12–23, July 2005.

■ CONTACT: ST.JOURNAL@ST.COM ■

Inertial Sensors for Wireless Body Area Networks: The WiMoCA Solution

Elisabetta Farella[1], Augusto Pieracci[1], Luca Benini[1], Andrea Acquaviva[2]

(1) DEIS - University of Bologna, Italy
(2) ISTI - Urbino University, Italy

Wireless Body Area Sensor Networks (WBASN) are an emerging technology enabling the design of natural Human Computer Interfaces (HCI). Automatic recognition of human motion, gestures, and activities is studied in several contexts. For example, mobile computing technology is being considered as a replacement of traditional input systems. Moreover, body posture and activity monitoring can be used for entertainment and health-care applications.

However, until now, little work has been done to develop flexible and efficient WBASN solutions suitable for a wide range of applications. Their requirements pose new challenges for sensor network designs, such as optimizing traditional solutions for use as environmental monitoring-like applications and developing on-the-field stress tests. In this paper, we demonstrate the flexibility of a custom-designed WBASN called WiMoCA with respect to a wide range of posture and activity recognition applications by means of practical implementation and on-the-field testing. pproach taken in the design of WiMoCA provides the necessary effectiveness. Software support, communication protocols and hardware architecture have been designed to match requirements of WBASN applications. Nodes of the network mounted on different parts of the human body exploit tri-axial accelerometers to detect its movements. The advanced digital Micro-electro-mechanical system (MEMS) based inertial sensor has been chosen for WiMoCA because it demonstrated high flexibility of use in many different situations, providing the chance to exploit both static and dynamic acceleration components for different purposes. Furhermore, the sensibility and accuracy of the sensing element is perfectly adequate for monitoring human movement, while keeping cost low and size compact, thus meeting our requirements.

We implemented three types of applications, stressing the WBASN in many aspects. In fact, they are characterized by different requirements in terms of accuracy, timeliness, and computation distributed on sensing nodes. For each application, we describe its implementation, and we discuss results about performance and power consumption.

1. INTRODUCTION

Technology advances in the design of low power digital architectures and component miniaturization pushed the development of wireless networks of sensors, which have been extensively studied in the last decade in the context of environmental monitoring, security of buildings and public spaces, traffic, and, more recently, health monitoring. Body area networks represent a recent evolution of this technology for the development of a new generation of human-computer interfaces (HCIs). Sensor technology enables the development of small form-factor devices that can be mounted on wearable nodes and communicate to each other and to a sensor fusion element, the latter of which can be either a mobile device such as a cellular phone or, a palmtop, or a personal computer.

Similar to traditional wireless sensor networks, body sensors collect information about the environment (the human body), that is subsequently correlated for monitoring and/or actuation purposes. Current design and technology trends (such as Smart Dust sensors [34] let us envision in the near future the availability of dust-like non-intrusive and extremely low power sensor nodes that can be easily worn by users.

In this scenario, body area sensor networks can be used to build general purpose natural interfaces, enabling the user to interact with the surrounding intelligent space by providing a new range of services. For example, personal devices (such as mobile phones or palmtop computers) can be enhanced by posture recognition technologies [24], [20], [32], replacing traditional input systems. Replacement of standard interfaces is also coming up for user identification. In fact, personalized and secure access will be ensured by body-based identification techniques [19], [40], [16].Going beyond the simple replacement of traditional interfaces, body area networks open the field to a new range of services and applications that can be built on top of them. For instance, health monitoring applications can exploit sensors for real-time tracking of body movements for rehabilitation purposes [7], [15]. Collaborative workplaces are another example of innovative application. Here, the body network is bridged to an external network that controls the surrounding context. Workers can interact and collaborate with each other through the company network. Finally, entertainment and cultural applications can profit from the enhanced immersivity of user interaction in virtual games, 3-D artificial worlds, and virtual heritage [18], [17], [48].

Even if the research in this field is very active and fruitful, many challenges have still not been faced. Wearability and autonomy of the nodes require power consumption optimization. While, at the same time, real-time sensing, processing, and communication capability impose tight performance constraints. Recent work has shown that wearability and low-power consumption requirements can be handledthrough an efficient design of sensor nodes [20], [32] in the context of a posture recognition application. However, these requirements must be satisfied in conjunction with ever increasing application demands for fast response time and flexibility and must be tuned depending on the application domain.

Moreover, compared to traditional sensor networks for environmental monitoring that are commonly designed to handle sporadic events, original requirements are imposed to achieve an efficient body monitoring system implementation, both from the node software design and network organization.

From a network perspective, by mounting the nodes on the human body, they are all in the same coverage area and directly connected to a base station (installed as an extension board on a palmtop PC). Thus, the topology of the body network does not require the implementation of routing and data aggregation algorithms. The base station also represents a gateway that connects the network to a personal computer (either a palmtop or a desktop computer) where the user application is located. On the other end, tight design constraints are imposed to match fast response time requirements of interactive applications, while providing configurability and flexibility, depending on

the application context. Moreover, to enable the development of a general purpose body-centric natural interface, flexibility of node architecture and network organization is a key issue.
In this paper, we present three applications exploiting natural human movement detection interface based on our flexible Wireless Body Area Network solution, called WiMoCA [20], where sensors are represented by tri-axial integrated MEMS accelerometers [27], [28].

Sensitive, compact, and inexpensive inertial sensors based on MEMS technology enable the design of a system that is easier to use, more functional, and more reliable [54]. Today, communication, consumer, and industrial markets are the main drivers for the development of commercial MEMS devices, and their huge volumes, the tight power, and form factor constraints strongly push toward lower price, size, and power consumption. Clearly, many applications reap significant benefits from these trends [10]. When a sensor has to be worn, accelerometers are effective for recognition purposes because of their small size and their immunity from electromagnetic interference and obstacles [6], [23], [4]. All of the presented applications exploit accelerometers to detect postures and gestures of the human body, but they are different in terms of processing requirements, real-time constraints, and distribution of the computation.

A posture detection system is presented as a first application that implements a natural human-computer interface. Body sensors collect inclination data from various parts of the human body and perform a preliminary processing, then forward them to the gateway node, which is connected to a palmtop PC. Software running on the palmtop correlates data coming from various sensors and compares the resulting position with a predefined set of postures.

Body area sensor networks are a promising enabling technology for health monitoring applications. In this paper, we show a rehabilitation system where accelerometers in sensor nodes are used to monitor user balance with respect to the center of mass concerning the body part under control. A real-time data processing algorithm implemented in a palmtop computer detects the wrong position and sends a feedback signal through a Bluetooth headset in order to stimulate the user to correct its posture. In this kind of application, accuracy and real-time constraints are critical for design guidelines.

WiMoCA has been exploited not only to detect static postures (directly extracted from data by the firmware running on the microcontroller), but also dynamic every-day life actions, like different kind of gait. We designed a gait recognition system based on a complex algorithm. Due to its complexity, the algorithm runs on a desktop computer equipped with the gateway node.

The applications described in this paper have been successfully implemented and tested in the field. They represent real-life test cases for the design of body-area wireless sensor networks for natural interfaces. Thanks to their heterogeneous characteristics, they allow the implementation of various processing models, where computation is either performed on the gateway node, or partially performed locally on the nodes. This allows us to stress the capabilities of the WiMoCA body area network solution.

Extensive experimental tests highlight that the power consumption characteristics of the body area network are suitable for human-centric applications, even in the worst case of continuous activity.

2. BACKGROUND ON WIRELESS SENSOR NETWORKS FOR HCI

To be considered wearable and unobtrusive interfaces, accelerometers must be mounted on lightweight and tiny boards without wires and cables for power supply and connectivity. To this end, some research laboratories recently targeted low-power, wireless, and wearable devices. Hand and head motion tracking systems have been designed [21] for Virtual Reality applications. This class of systems is dedicated and optimized for tracking specific body parts and movements. WiMoCA targets general

purpose gesture recognition and for this reason is designed to be installed on any body part. Researchers have done a consistent amount of work to build absolute orientation detection devices mainly targeted to robot navigation [4], [23]. In these kinds of devices, accelerometers are used to compute orientation referred to local vertical direction, while magnetometers are used to compute the direction of gravity with respect to the Earth. Recently, MARG modules equipped with triads of orthogonally mounted accelerometers and magnetometer and angular rate sensors have been designed [4] ; these modules target 3 DoF (Degrees of Freedom) orientation measurements of static or slowly moving rigid bodies. Since a module does not have wireless communication capabilities, data must be transmitted over a wiredlink for further processing. Accelerometers, magnetometers, and gyroscopes are exploited in [23] to build a wireless portable absolute orientation estimation device. The device uses a Kalman filter to filter out accelerations due to user movements. Designed to be part of a sensor network, the device uses a Bluetooth module for network communications.

A new sensor node solution that has recently been developed is the InterSense Wireless InertiaCube [26]. This solution is composed of a set of sensors without processing capabilities connected to a receiver, which acts as a gateway. Processing must take place at the host side to which the gateway is connected.

Compared to our solution, it features a higher accuracy, but it has a bigger form-factor and lacks processing capabilities on the nodes, thus resulting in less flexibility. WiMoCA, on the other hand, is a network of sensor nodes that enables the implementation of distributed computation, thanks to node's processing capability. For example, in the posture recognition application we describe in this paper, this capability is exploited to perform partial recognition of postures inside the nodes. Moreover, the cost of InertiaCube3 is one order of magnitude higher than that of the WiMoCA nodes. Moreover, because WiMoCA is fully wireless and equipped with standard interfaces, it makes the system suitable for a wide range of applications for wearable and mobile devices, like the ones presented in this paper.

Comparing this class of sensors to WiMoCA, the main difference is that they are designed to perform tracking of absolute orientation. In our design, we are interested in gesture recognition instead of tracking, and for this reason we do not need complex filtering algorithms to compensate for the error introduced by the integration steps. Instead, we use a geometric algorithm which uses accelerometers as inclinometers. Moreover, because they are not designed to detect absolute orientation, we do not use gyroscopes and magnetometers. For this reason, WiMoCA shows a smaller form-factor, making it more suitable to be mounted on various body parts without impacting normal user movements.

Accelerometers and gyroscopes have also been used to build human-computer interfaces [44], [42], [12]. This class of HCI interfaces differs from WiMoCA either because they target a specific movement or because they are not designed to build a distributed general purpose recognition system. In fact, WiMoCA provides advanced wireless networking hardware and software support to communicate with other WiMoCA modules to build a body area network.

Wearable inertial sensors are a low-cost, low-power solution to track gestures and, more generally, movements of a person. The implementation of a body-centric network mounting inertial sensors has been explored in many fields. Examples are in context-aware applications [35], [45] and monitoring of patient activity in the medical domain [2]. A combination of accelerometers, magnetometers, and temperature and light sensors to be worn by users has been applied to help indoor navigation [22] and to infer a user's location. Many research studies have focused on hybrid sensor networks or sensor fusion techniques. The techniques whose sensing elements are distributed along the body and data processed both off-line or in real-time [56], [33]. Even if these solutions are suitable to specific applications, they are not

tailored to applications requiring high wearability and very-low power consumption.

Commercial solutions like Mica Motes [30] are designed to handle sporadic or slowly changing events (such as temperature and pressure variations) and to interface with web applications for environmental monitoring. As such, they are equipped with embedded operating systems that support a complete network stack. Compared to these solutions, our system has been designed for real-time interactive applications with low-power requirements, and, for this reason, we focused on minimizing software overheadby implementing our own component drivers and communication layer. Moreover, Mica are equipped with analog bi-axial accelerometers, which require additional ADC conversion and are less tailored to gesture recognition applications compared to digital integrated tri-axial devices equipped by WiMoCA.

The software support we developed to drive the module exploits the low power consumption of the sensing unit to improve battery lifetime and provide a network stack, (physical and MAC layer) to drive the wireless transceiver, to interface with other modules, and to organize accelerometer data in packets to be sent to a remote machine for further processing.

3. WBASN ARCHITECTURE

WiMoCA [20] sensing node is designed to be wearable and low-power. It has a modular architecture to ease fast replacement and update of each component. It is composed by three sections (Fig. 1) namely MCU/sensors, RF, and power supply. The module sensibility obtained is 2mg, and the maximum throughput is 36Kbps. The core of the WiMoCA node is the low-cost, low-power ATmega8 8-bit microcontroller based on the AVR RISC architecture. The sensing is based on a MEMS tri-axial digital accelerometer by STMicroelectronics (LIS3LV02DQ [29]). This "one-chip three-axis" MEMS-based accelerometers with digital interface chipsets, combines small size, high resolution, and low power consumption. In fact, maximum power consumption is 5mW in active mode and 1 μW in sleep mode (at supply voltage of 3.3V). The RF section is based on the RFM TR1001 transceiver operating at 868 MHz and reaching a maximum bit rate of 100kbps.

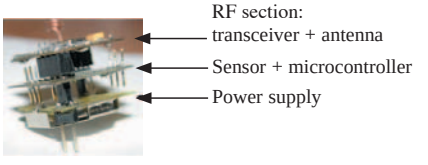

FIGURE 1: WiMoCA SYSTEM STRUCTURE.

Among other nodes, the gateway is a special one which mounts two layers dedicated to communication interface. Similarly to other nodes, it has an RF section for exchanging data with the body network, but it also interfaces to the external world, e.g., a PC or workstation, through an interchangeable link such as RS232, Bluetooth, or Ethernet.

The WBASN is organized with a star topology and can be easily worn. WiMoCA nodes are end-points of the star, and they are responsible for sensing and acquiring data from the environment before sending them to the gateway node (Fig. 2). The gateway performs bridging functions but does not have a sensing device. It is used to form the backbone of the system by connecting the WBASN to a host machine or an external network. Coordination among transmissions of various nodes to the gateway is performed at MAC-level.

WiMoCA software implements a collision free MAC protocol, inspired by real-time MAC presented in [9]. Compared to collision avoidance protocols developed for sensor networks such as TMAC [52], collision free protocols are suitable for handling real-time traffic, since they avoid the overhead imposed by collision detection packets by allocating a time frame to each message [50]. In fact, the typical usage of the body area sensor network imposes a periodic sampling of acceleration values from a subset of nodes. For instance, a typical application of WiMoCA

is to recognize if the user is walking or sitting down through sensors placed on arms and legs. This is obtained through real-time processing of acceleration values. Moreover, this protocol is suitable for networks with a constant number of nodes, such as the one we target in this work.

The CFRT (Collision Free Real-Time) protocol basically divides time into frames in which only one node is allowed to transmit. The scheduling order is derived by a Message Table stored in each node and is identical for all the nodes so that each of them knows when it has the right to transmit. The table contains an entry for each node allowed to transmit or receive in a frame. Fields in the entry specify source, destination nodes, message length, and message period. Compared with [9], where the table is built by applying earliest deadline first scheduling protocol (EDF) to a queue of waiting messages; here, the table is application dependent. The scheduling order is programmed depending on the job that must be performed by the body network. Moreover, we do not allocate frames for inter-cell messages since, in the body network, all the nodes belong to the same cell.

For this reason, a cross-layer approach has been followed where the application is responsible for performing message scheduling. The Message Table at each node can be dynamically re-programmed via wireless channel at the beginning of each frame. At this time, the gateway can notify nodes that a table-update is required. As a consequence, nodes prepare to receive the new table from the gateway. This solution allows applications to program the Message Table, depending on the sensing pattern, which in turn depends on the movement or gesture to be recognized. In general, the table keeps constant during an acquisition session, limiting time/energy spent for re-programming.

During table-updating phase, a time synchronization message is also sent by the gateway to each node, since all the nodes must be synchronized to share the bandwidth. The synchronization message is sent by the gateway at a predefined time interval, as explained in the next section. In general, in each round of the table, called *table period*, a time slot (*management slot*) is allocated for synchronization and table updates. When it is not used for synchronization or table updates, transceivers of all the nodes are switched off for its entire duration. The timeline of MAC protocol operations is shown in Fig. 3.

4. APPLICATIONS

Given the architecture described in the previous section, we show in the remainder of this paper a set of three applications

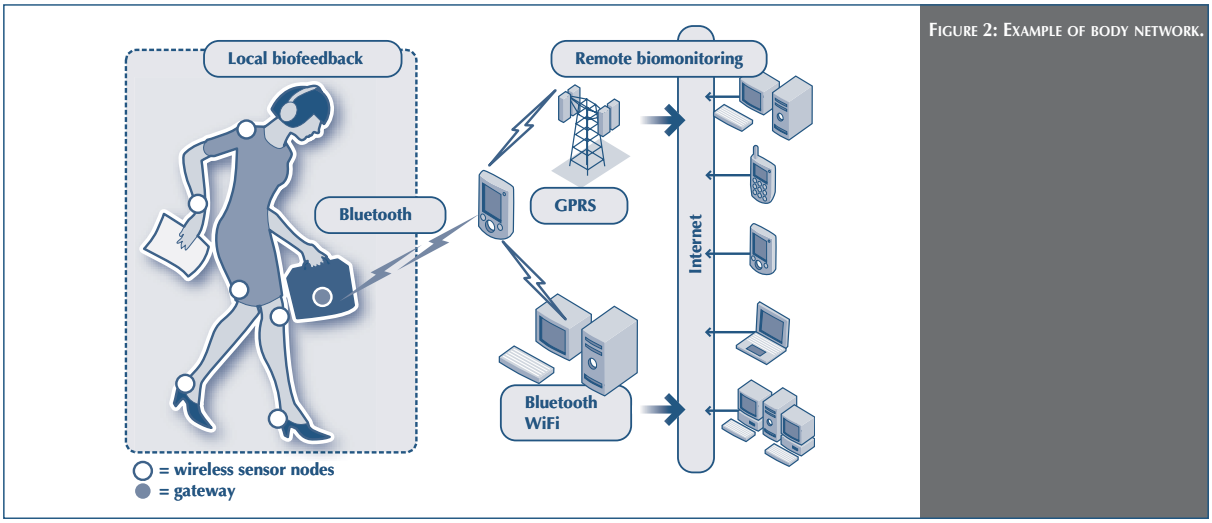

Figure 2: Example of body network.

enabled by the WiMoCA wireless body area sensor network. We implemented these applications to test the flexibility of the body area network software and hardware.

They can be distinguished based on i) the distribution of computation among nodes and gateway; ii) the type of data transmitted on the network; iii) the level of accuracy required in data sampling and processing and iv) the type of processing performed at the host side and real-time constraints. Results of characterization of each application in terms of performance and power consumption are also described.

4.1 Posture Recognition Application

Posture recognition is aimed at detecting user postures among a set of different possibilities. The size of this set depends on the application, which in turns affects the number of sensing nodes to be used.

Each sensing module monitors the inclination of a certain part of the body, acquiring acceleration samples along three axes. Acceleration values are averaged first, then the module tilt with respect to the gravity is computed and encoded. Thus, in this application, nodes are heavily involved in the overall computation. After processing, data are sent to the gateway according to the schedule imposed by the contention-free MAC protocol implemented on each node. Each module inclination is then collected at the gateway side frame by frame and subsequently combined to interpret body posture. Finally, the gateway communicates the result of detection to a Java application running on the host machine, where a Graphical User Interface (GUI) is updated to display the current user posture. Being only interested in a predetermined set of postures, accuracy is not critical in this kind of application. After averaging, inclination data are only transmitted when they overcome a threshold that determines the cross-point between two different postures.

In the present implementation, communication between the gateway and the host machine can be performed both through RS232 and Bluetooth protocol. The host machine can be a Desktop PC but also a palmtop computer. In this latter case, the Bluetooth solutionis preferred. The choice of a Java implementation for the application provides the necessary portability.

4.1.1 Setting the context

Even though the idea of exploiting user postures to implement alternative input systems for cell-phones and personal digital assistants (PDA) has been explored in the last few years [24], [3], many challenges have not yet been faced. From an implementation viewpoint, recent work showed that wearability and low-power consumption requirements can be faced through an efficient design of sensor nodes [20], [32]. However, these

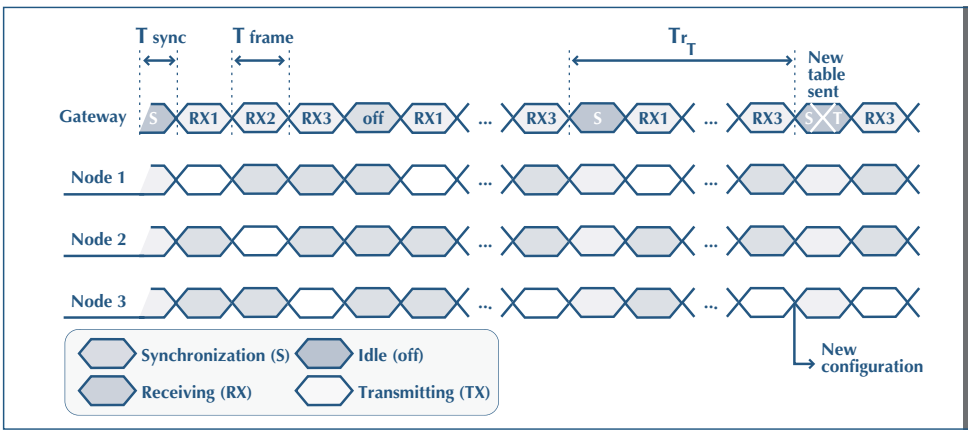

FIGURE 3: TIMELINE OF MAC PROTOCOL.

requirements must be satisfied in conjunctionwith ever increasing application demands for fast response time and flexibility. As a consequence, a complete implementation has not been presented up to now, to the best of our knowledge.

Posture recognition is often a primary goal of activity recognition system for elderly care [15]. The distinction between static postures (standing, sitting, lying supine, on a side, and prone) is investigated in [53], where the user is monitored with two accelerometers, one mounted tangentially on the thigh, and the second mounted radially on the sternum. The signals were acquired for subsequent off-line assessment. In a recent study [8], more than 20 (sub)postures and motions are distinguished, using four bi-axial accelerometers, worn by the user on different parts of the body and wired by cables among them and with a data recorder worn in a belt around the waist. After collection, data are processed off-line, and feature extraction and pattern recognition techniques are applied. This solution is suitable for motion and posture analysis, but the system is obtrusive and cannot be used for real-time interaction and posture recognition.

4.1.2. Posture recognition system

In the accelerometer-based posture monitoring application, the user is equipped with 3 sensing modules placed along the body (Fig. 4), precisely on the trunk (node A), the thigh-bone (node B), and the shinbone (node C). If we refer to the plane defined by the direction of gravity and the ideal line between the shoulders, the projection onto this plane of axes relative to each sensor module is shown in the right part of Fig. 4. The application implemented aims atdetecting user posture among seven different possibilities, as shown in Fig. 6 (sitting, standing, and lying in four different manners).

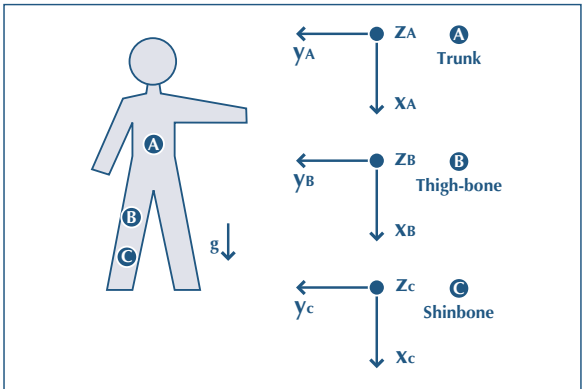

FIGURE 4: SCHEMA OF THE USER AND THE MODULE SETUP AND ORIENTATION ALONG THE BODY.

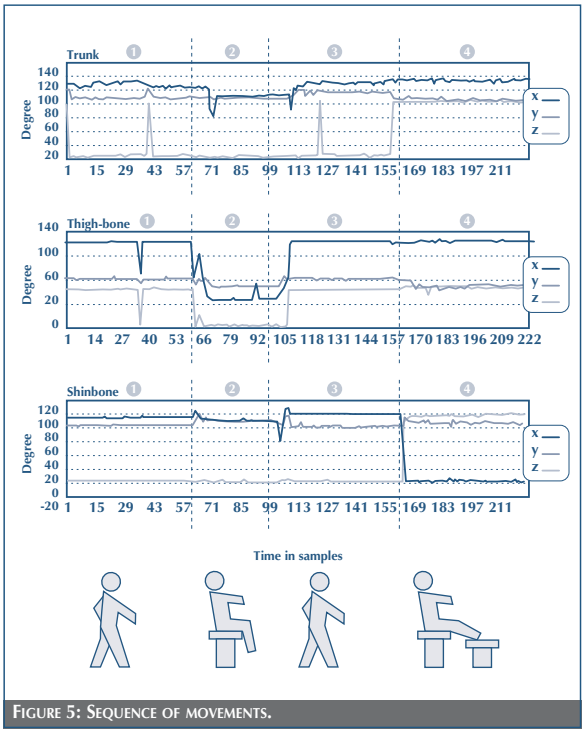

FIGURE 5: SEQUENCE OF MOVEMENTS.

Fig. 5 shows a sample of tilt data collected by the three end nodes. The stream corresponds to a sequence of postures: 1) standing; 2) seated; 3) standing and 4) seated with legs extended (as lying on a table). A different sequence can indicate, for example, that the user has fallen down and then used to generate an alarm or to contact assistance. The plot reports computed angle degrees versus time (expressed as number of samples).

Programming end nodes detect posture by acquiring sensing data and computing average acceleration in a predefined time

Figure 6: Seven static posture detected.

window to filter out temporary variations of user position or noise. The choice of a time window of 8 samples simplifies the averaging process, which is performed through a right shift of 3 bit. The acceleration value is 16 bit for each axis in two's complement encode (9 byte of memory are needed). In order to reduce the amount of information to be sent over the wireless channel, nodes identify one configuration among a known set. For each node, in fact, there is a set of valid configurations that is stored in the node's memory. These configurations are the result of tilt positions that are computed based on the sensor orientation w.r.t. the gravity component [6]. Accelerometers provide data in the range [-2g, 2g], though when performing tilt detection, the range is restricted to values within [-1g, 1g]. We are interested in detecting three different situations for each axis, directed as orthogonal to or opposed to the gravitational acceleration. The range of acceleration values has been quantized in classes, and the actual acceleration value is thus classified in one of three sub-ranges.

Combining the discretized value for each axis, the set of three accelerations is associated with a specific 4-bit configuration identifying the tilt of the module (see Table 1). The number of possible tilts recognized can be varied according to application need, introducing more than three sub-ranges for acceleration quantization.

The sub-ranges don't necessarily cover all the possible acceleration values as in our case. Thus, the process of module tilt identification also considers the possibility of providing a code corresponding to "unclassified" (last row in the table) as output. At the end of the tilt identification process, the module tilt code is sent to the gateway.

X	Y	Z	Module Tilt (4 bit)
p	o	o	0000
o	o	g	0001
o	p	o	0010
o	g	o	0100
o	o	p	1000
g	o	o	1010
none	none	none	0101

Table 1: End-node possible inclination: p = parallel but opposed to $\vec{g}$, g = directed as $\vec{g}$, o = orthogonal to $\vec{g}$.

Trunk	Thigh-bone	Shinbone	User posture (ASCII)
0000	0000	0000	1 (Standing)
0000	0001	0000	2 (Seated)
0000	0001	0001	3 (Seated with leg up)
1000	1000	1000	4 (Prone)
0001	0001	0001	5 (Supine)
0010	0010	0010	6 (Lying on right side)
0100	0100	0100	7 (Lying on left side)

Table 2: Posture Table in a practical case.

On the gateway side, the firmware acquires and accumulates tilt data from each end-module for a given frame and combines all module tilts to identify a known body configuration. Finally, it sends the identified position to the application. The gateway is also responsible for handling the set-up of nodes and network (e.g., Message Table transmission, association of each node with a body part).

According to the Message Table, which is identical on end-nodes and gateway, the latter waits for the reception of tilt data from each end-node. On the gateway, a table, called Body Location Table (BLT), containing correspondence among node ID and position on the body is stored. A default BLT is initialized each time the gateway is reset; nevertheless, the table can be programmed by the application running on the general purpose system at any time.

Supported by the BLT, data coming from different nodes are combined to determine user posture among a set of possible postures contained in another table, the Posture Table (see Table 2

and Fig. 6). For clearness, we refer to the actual application where the end-nodes are three, displaced on the trunk, on the thigh-bone, and on the shinbone and, all oriented as shown in Fig. 4.

After encoding, the user posture is sent through the serial (or Bluetooth) interface to the Java application.

The Java application on the host side accomplishes a set of simple tasks: (i) continuous data acquisition from the serial or Bluetooth port exploiting the Java COMM API; (ii) user posture encode extraction from data stream; (iii) GUI implementation for real-time visualization of an image corresponding to actual user posture; (iv) initialization control commands, set-up of all tables (BLT, Message Table, etc.) resident on the network.

Most of the time, the application listens to the COM port (Bluetooth or RS232) for acquiring data coming from the gateway. The extraction of the encoded posture is committed to a specific thread that first identifies the start byte of the packet sent by the gateway, then the time-frame through the counter byte, and finally the posture code. These steps are continuously repeated as in a state-machine. The code extracted is communicated to the Graphical User Interface (GUI) thread, which updates the picturedisplayed (Fig. 7). The GUI offers the user the chance to enter a new BLT through specific text fields and send buttons. This can be useful if a certain module is placed in a different location from the default one at run-time. When the button is clicked, a control character is sent to the gateway and subsequently the new BLT with correspondence among nodes and parts of the body.

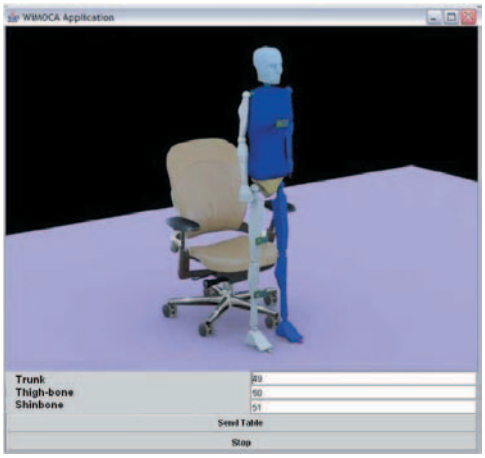

Figure 7: Graphical user interface snapshot.

4.1.3. Performance and Power Consumption

To assess power consumption of the node while running the posture recognition algorithm, we performed additional characterization tests, whose results are shown in Fig. 8. With 3 nodes, we can achieve a maximum sample rate of 651 positions per second. In practical cases, the maximum frequency of human movement is 30Hz so that 60 positions per second is a sufficient rate for detecting postures without losing information.
Battery lifetime depends on the frequency of samples. We note that the strong dependency of the transmission power on the type of bit also affects battery lifetime in a considerable way at full node activity.

Phase	Our case (Vcc= 3.3V)
Transmission of a data packet	$P_{TX\text{-}DP\ (M=11)}$ = 138,5 mW $P_{TX\text{-}DP\ (M=0)}$ = 21,60 mW
Total power consumption	$P_{NH\ (M=11)}$ = 46,00 mW $P_{NH\ (M=0)}$ = 16,85 mW
Battery life at latency = 1,536msec (651 positions per second)	$T_{BLH\ (M=11)}$ = 7h $T_{BLH\ (M=0)}$ = 19h
Battery life at latency = 16,7msec (60 positions per second)	$T_{BL\ (M=11)}$ = 18h $T_{BL\ (M=0)}$ = 21h

Figure 8: Power consumption and battery life-time of the node during posture recognition.

In addition, it must be noted that battery lifetime decreases significantly much lower (from 19h to 7h) as the position sample rate increases (from 60 pos/sec to 651 pos/sec). This is because of the higher power efficiency of the transceiver at the higher bit rate.

4.2 Bio-Feedback Application

In this application, the sensor network is distributed on the user's body for balance monitoring and correction. It represents a relevant feasibility study for the implementation of a mobile

recognition system using feedback actuation. Software architecture is characterized by nodes performing basic sensing functions with almost no on-board processing. All the computation is performed on the palmtop computer (we refer to it as PDA - Personal Digital Assistant) that interfaces with the wireless body network by means of the gateway node. The PDA is responsible for data processing and is in charge of activating the actuator that provides feedback to the user about the correct user posture. In the implemented set-up, we adopted audio feedback through a headset.

This application differs from the posture recognition system because of the software organization but also because of the use of accelerometers for fine position tracking. Accuracy requirements are much more stringent w.r.t. posture recognition. Data collected and transmitted are acceleration values that are then processed at the host side. The hardware and software architecture (communication protocols, power management policies, and application-level control) have been tuned to optimize cost, battery autonomy, and real-time performance required for this application.

4.2.1 Setting the context

Inertial sensors (accelerometers in particular) and biofeedback have proved to be essential elements in applications for balance control because of their small size, portability, and supply of useful kinematic information [46], [51]. Postural unbalance is a crucial aspect in many musculoskeletal, neurological, and age-related diseases [31], [47], [38] and in several sport disciplines [49]. Recent studies have shown that improvement of balance may be gained by use of biofeedback, based on the principles of physiological adaptations and brain plasticity [11]. Nevertheless, biofeedback systems for balance control are usually cumbersome and expensive; furthermore, they are mainly for ambulatory use and require specific expertise (e.g. [1]). Hence, a goal in the development of biofeedback systems for balance and motor rehabilitation is easy accessibility, specifically with respect to unobtrusiveness, portability, and cost.

4.2.2 Bio-WWS system

Bio-WWS is composed of four kinds of nodes: one or more sensor nodes; the audio actuator node; the PDA, which acts as a general purpose node; and the gateway node. At present, the system has been tested with three sensor nodes, one located on the trunk for sway monitoring and the other two on a leg (thigh and calf) for determining whether the user is walking or standing (Fig. 9).

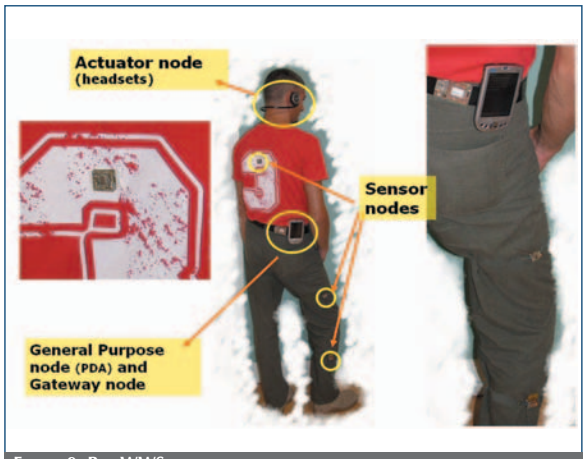

Figure 9: Bio-WWS set-up.

A star topology is used as in posture application. With respect to previous applications, here nodes do not perform preliminary processing. They instead send acceleration data to the gateway, which performs signal pre-processing and data normalization. It also organizes data retrieved from the sensor nodes in packets to forward them to the PDA over a Bluetooth link. Finally, the PDA processes data from the body sensors in order to resolve a human's posture condition and to give adequate feedback to the user.

The host PDA device we used for testing the biofeedback application is a HP iPAQ 5550 with Bluetooth and WiFi capabilities. To reach a fully wearable system, we used a wireless link through a PDA and a headphone representing the auditory feedback actuator.

The system considers trunk planar accelerations (in the forward-backward and left-right directions), both gravitation and

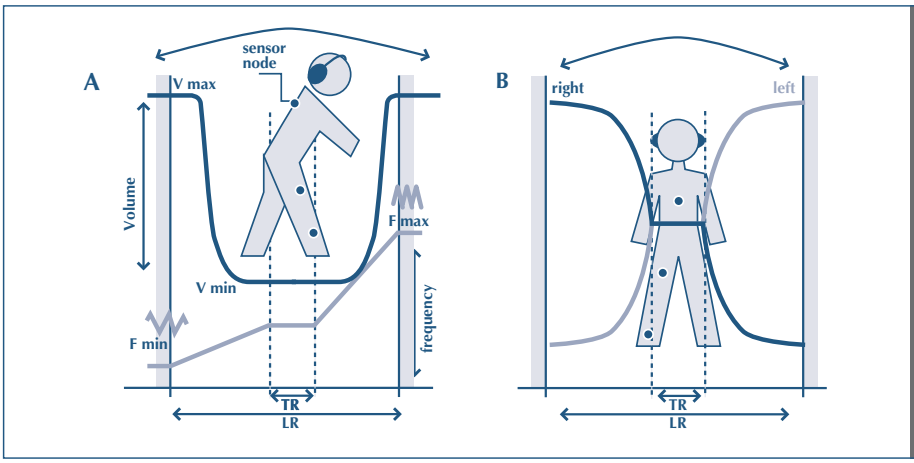

Figure 10: Audio biofeedback sound dynamics.
A. Forward-Backward direction
B. Left-Right direction

dynamic, and determines an estimate of sway of the user's center of mass. Accelerations collected by the gateway are coupled with a timestamp in order to have a screenshot of all sensors placed in the human body at the same time. Then, the integrated accelerations are sent to the PDA using a Bluetooth link.

To generate the feedback signal, accelerations are coded into a stereo sound using a suitable algorithm, aimed at providing high comfort level to the user while improving his/her stability while standing [16], [17], [19]. In particular, accelerations in the forward-backward direction are coded by frequency and amplitude (volume) modulation of the sound (Fig. 10.a). The frequency codes the value of the instantaneous acceleration while the volume increases with the distance from the *Target Region* (TR). The TR is a range of acceleration values which are considered to be safe for the user. When in this region, the user perceives a low but constant tone.

This has been shown to be more effective for the user than suddenly turning on the audio signal. Thanks to the feedback signal, the user notices when he/she goes outside the TR. When the user's accelerations are inside the TR range, the sound volume and frequency are fixed [13]. The user's accelerations in the left-right direction are coded by Left/Right balance modulation of the sound (Fig. 10.b). Besides TR, another parameter of the biofeedback algorithm is the *Limit Region* (LR). This region defines the operating range for the algorithm. If theuser crosses the LT, a constant tone is played, meaning that the user is in a dangerous position. A GUI was implemented on the PDA to support easy setting of parameters of the biofeedback algorithm.

4.2.3 Power and performance characterization

Measurement of power consumption was performed with an acceleration sampling frequency of 60Hz (to obtain a useful bandwidth of 30Hz, adequate for capturing human movements). Communication speed was set at 32kbps and at 230kbps from gateway to PDA, transmitting via Bluetooth at 2,4GHz. Power consumption when all components are active and the transceiver is continuously sending data corresponds to 200mW for the gateway and 45 mW for a single end-node. In an idle state (where only the reception mode is active), the power lowers respectively at 25mW and 20 mW. When active, the gateway battery (500mAh) and the single node battery (100mAh) have a lifetime of 8 hours. The sleep mode corresponds to having all devices in the minor consumption state (1.5mW for gateway and 10 W for end-nodes). For a 20% duty cycle, the lifetime reaches 40 hours for the gateway and 38 hours for the node.

In terms of power consumption, tests revealed that the bottleneck for a system lifetime is the palmtop computer (with a 3.7 V

battery of 1000mAh), which consumes around 350mW. As a consequence, the power budget of the WBASN is suitable for this kind of application.

Accelerometer accuracy and data transfer rate of the nodes abundantly matches the requirements of the specific application of Bio-WWS since human movements are below 50 Hz (even below 10 Hz in quiet standing), and trunk accelerations range from 5-10 mg [39] to 40-50 mg.

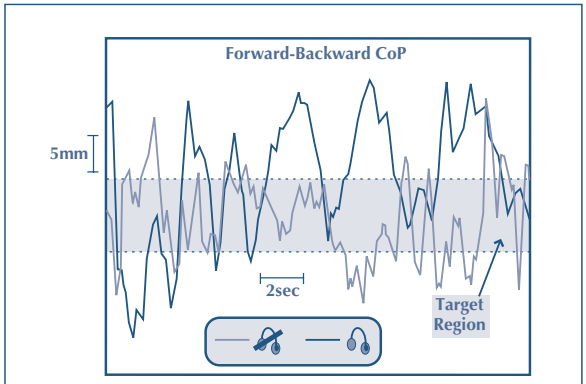

Figure 11: Results on a representative subject: BioWWS decreases the CoP excursion.

Tests in the field have been also performed to assess the effectiveness of the Bio-WWS application. Several experimental trials were performed on 5 young healthy users as preliminary validation procedure. Users were required to stand with eyes closed on afoam-rubber surface that simulated sensory alterations at foot level. Sway during quiet standing, which indicates the level of achieved balance, was measured by means of a force platform (Bertec 4060-08), which quantified movement of the center of pressure (CoP) [55], i.e., the application point of the ground reaction force. Subjects performed three trials first without Bio-WWS and then three trials with Bio-WWS. Preliminary results are exemplified in Fig. 11, where CoP displacement for a representative subject is shown, without BioWWS and with BioWWS (after the calibration time). Effectiveness of the system for this representative trial is proved by the minor CoP excursion with the use of BioWWS, then without it, in the same sensory-altered conditions. Future development of the system will include further investigation and experimental sessions to complete the system validation and to optimize the sound modulation parameters.

4.3 Gait

WBASN can be effective for implementing activity recognition systems. WiMoCA is used to implement a Gait Analyzer, as we called it, aimed at classifying human walking behavior, thus demonstrating the flexibility of our WBASN. In this application, we use asingle wireless end-node equipped with the triaxial MEMS accelerometer and placed on the ankle for the purpose of sampling accelerations to perform step recognition. Acceleration values are packed and transmitted. Hence, compared to posture detection system, accelerometers are not used as inclinometers. Moreover, after calibration (as explained later), only the accelerations on a single axis are transmitted. Similar to a bio-feedback system, computation is offloaded from the end-node. The gateway node isconnected to a personal computer that processes acceleration data coming from the node. Signal processing techniques are applied to data to isolate each step and classify the data among many possibilities. We considered four kinds of steps in the currentimplementation: i) step in place; ii) staircase step; iii) step forward slow and iv) step forward fast.

4.4 Setting the context

Detection and classification of walking behaviors through inertial measurements can be employed for activity recognition [14], gait monitoring for healthcare purposes [15], or for indoor navigation systems [36]. In [37], walking behavior detection is aimed at supporting an indoor localization system counting the number of steps and estimating a person's current location referenced on a known starting location. Many of the techniques applied for this purpose are based on a training phase on data collected during daily activity and afterward processed off-line. In [41], gait analysis targets clinical systems. Gait is continuously monitored though an instrumented shoe, which sends data coming from a

tri-axial accelerometer and a pressure sensor wirelessly and in real-time. The shoe has been more recently enhanced with other sensors obtaining a multimodal sensor interface used also to provide an interactive music device for dance performance [43]. Gait monitoring has also been studied collecting accelerometer data referred to the head and pelvis in young and older subjects as they are walking on a level and an irregular walking surface.

The aim of this investigation is to understand how ageing affects walking behavior and posture. In [14], a gyroscope attached to the shank is used to detect walking upstairs, toe-off, heel-strike, and foot-flat. Wavelet transformation in conjunction with a simple kinematics model have been exploited and the system tested in many different conditions. In [53], the dynamic contribution to acceleration of a monoaxial accelerometer fixed on the thigh has been used to distinguish among walking, stair ascent, stair descent, and cycling. Compared to previous work, our solution provides: (i) greater flexibility, since a single WBASN can be used for many applications; (ii) processing is performed run-time processing without the need of a pre-characterization phase and (iii) it is a completely wireless system.

4.5 Gait recognition application

To classify movements correctly, the Gait Analyzer [25] needs to know the exact orientation of the node as it has been placed on the leg. For this reason, a system calibration must be performed first. The user must place the sensor, which can be identified as a dice, with one of the faces orthogonal with respect to the gravity direction. The processing flow has four main tasks: (i) payload extraction; (ii) system calibration; (iii) step count; and (iv) gait identification.

The first task is similar to other applications discussed and does not need further elaboration. Instead, focusing on calibration is more critical. The calibration strategy we choose is one among several possible alternatives. In our case, it is performed with the user standing still. The maximum value of acceleration measurable for any components is close to one, since it is normalized with the gravity acceleration. During calibration, data are captured for a fixed time window to obtain an average value for each axis.

By comparing the computed average with thresholds shown in Table 3, the orientation of the node is identified. This helps the user set the node in the correct position. Note that the user is requested to set the node so that acceleration values are close to either zero or one (depending on which axis is orthogonal or parallel to the ground). However, the user is not asked to set a particular axis orthogonal to the ground plane; the node orientation need not be carefully set either. The averages computed during the calibration phase are then stored to be later used later as a reference during the recognition.

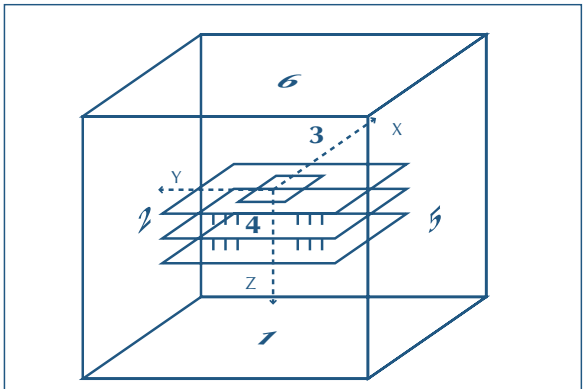

FIGURE 12: DICE FOR EXTRACTING ORIENTATION.

Next, absolute position needs to be then found out. In general, to identify the position of the tri-dimensional node in the space, it is sufficient to consider it as a dice (Fig. 12). To achieve this target, a number is associated with each surface of the dice. In Table 3, six different positions are identified depending on the acceleration values along the *X*, *Y* and *Z* axes. A position "1" means that the surface "1" is parallel to the ground plane. In Fig. 12, the surface "1" is identified by acceleration values along the *Z* axis as "-1".

The reference position in which the calibration takes place in our case is with the node having a surface in contact with the leg.

	Average Value ≥ 0,5	*-0,5 ≤ Average Value ≥ 0,5*	*Average Value ≤ -0,5*
Axis direction w.r.t ground	orthogonal, direction down	parallel to ground	orthogonal, direction up
Symbol	1	0	-1

Table 3: Position codes.

Position code	X	Y	Z
1	0	0	-1
2	0	-1	0
3	-1	0	0
4	1	0	0
5	0	1	0
6	0	0	1

Table 4: Position codes.

This surface lies on a plane vertical w.r.t. to the ground. If the surface of the dice lying on the leg is the one identified by the number "1", as shown in Fig. 13, the average acceleration values of the three axes are zero along *X* and *Z*, while they provide an average close to "-1" along *Y* axis because it is parallel to the ground plane. Absolute no deposition is then computed by combining the information obtained for each axis (Table 4). After calibration, by means of the knowledge of node position over time, it is possible to understand when the subject is walking.

The algorithm implemented to recognize single steps is based on the observation of acceleration values along the component orthogonal to the ground. This is the acceleration component which, during calibration phase, was parallel to the gravity acceleration. If the node is fixed on the ankle (see Fig. 13), the axis involved is *Y*.

The overall step recognition is composed of two distinct phases. First, the length of the step (i.e. the number of acceleration values) is determined in a first "row" recognition phase. Then, the algorithm performs a fine detection based on the step type.
Let us focus on the first phase. To perform *row* recognition, the human step is considered as a combination of a raising phase and a leaning phase. Based on this observation, when the foot is raised, the acceleration collected along the Y axis increases due to the dynamic component solicited by the upward movement. Alternatively, as the foot touches the ground, acceleration along Y decreases. At the end of the leaning phase, there is a bouncing effect that is used to recognize the end of the step.

In fact, to perform step recognition, two thresholds have been set (as shown in Fig. 14) that are used to compare positive and negative acceleration values respectively. More precisely, first the negative and the positive peaks are determined; then these values are compared against the thresholds. These threshold values have been obtained through extensive characterization in the field.

A step is identified by three consecutive events (Fig. 14):
i) the positive peak is larger than the positive threshold;
ii) the negative peak is lower than the negative threshold; and
iii) the acceleration values stay under the thresholds for at least three consecutive samples. All data referring to a single step are stored in a static array.

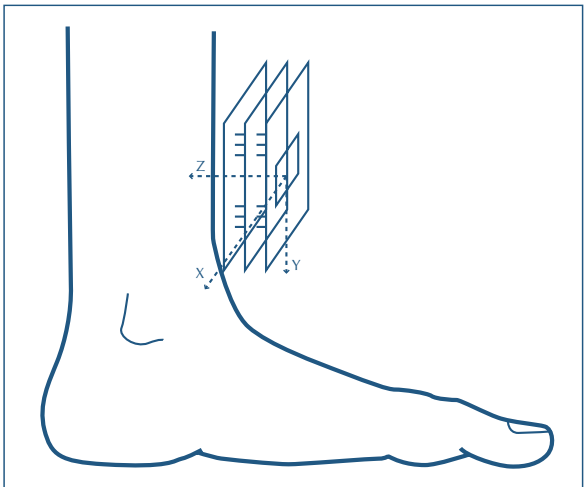

Figure 13: Setup on the user for the sensor node.

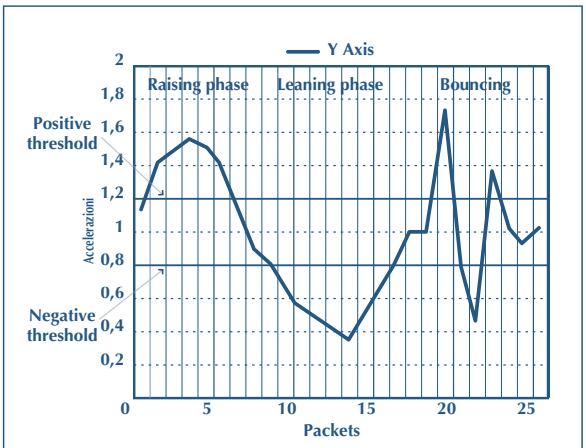

FIGURE 14: ACCELERATION PLOT CORRESPONDING TO A STEP FOR Y AXIS.

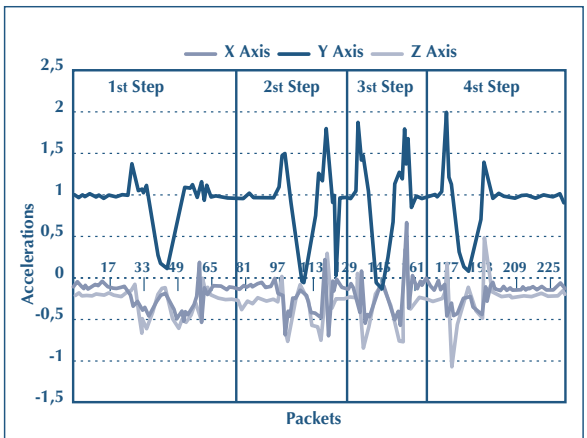

FIGURE 15: STEP ACCELERATION WAVEFORM.

Note that this procedure allows us to identify the length of the step in real-time. Other approaches in literature perform an off-line characterization phase that is then used to correlate sampled data with stored data [5].

Once a single step has been stored in all three vectors (one for each axis), data are processed to distinguish in real-time among four kinds of steps (more can be added):

1. going up on a staircase;
2. step in place
3. step forward fast
4. step forward slow

To identify the action of going up step by step on a staircase, an autocorrelation function is applied to the vertical axis vector. This vector contains acceleration values for increasing time instants.

Autocorrelation values computed are stored in an output vector, which is then processed to verify its monotonic behavior (see Fig. 16). If the function is not monotonic, the user has mounted a step of a staircase. If yes it is, the system verifies one of the other possibilities (i.e., step in place or step forward).

To recognize if the user is performing a step in place, the result coming from the auto-correlation function is processed to find its higher value, verifying whether is greater or not that a empirically defined threshold. If both results have a negative outcome, the user has probably performed a step forward. To identify the speed of the step, the standard deviation referred to the axis parallel to the movement direction is used. Standard deviation is compared with another empirically defined constant acting as a threshold. This threshold can also be set by the user depending on the speed that has to be identified.

5. EXPERIMENTS

In this subsection, we show experimental results about the characterization of human steps. First of all, we show in Fig. 15 acceleration values corresponding to a sequence of steps. It can be noted that each step has at least a positive acceleration peak, a negative peak, and an intermediate phase, where acceleration values stay under the thresholds.

Another set of experiments has been performed to tune the algorithm for recognition of various types of steps, as explained in the previous subsection. An example of such experiments is shown in Fig.16.

From Fig. 16, it can be noted that different types of steps are characterized by the different behaviors of the autocorrelation function. Steps performed in place lead to autocorrelation values

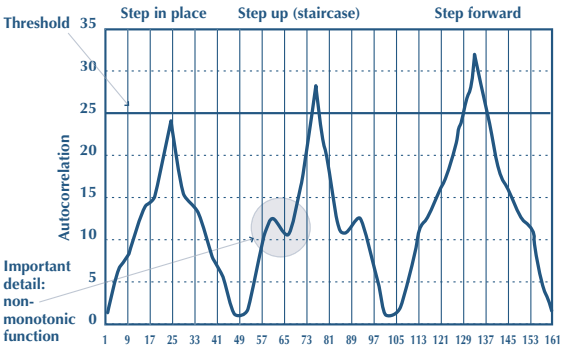

FIGURE 16: AUTOCORRELATION OF TWICE THE VERTICAL ACCELERATION (VECTOR CORRESPONDING TO Y AXIS).

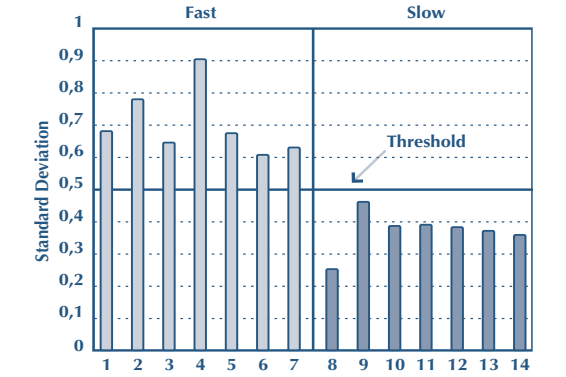

FIGURE 17: STANDARD DEVIATION FOR 14 DIFFERENT STEPS FORWARD, 7 STEPS PERFORMED FAST, 7 STEPS SLOW.

under a certain threshold. Staircase steps are identified by the non-monotonic behavior of the autocorrelation function. Steps forward are identified by the autocorrelation function overcoming a predefined threshold, as shown in Fig. As mentioned before, standard deviation can be used as an indication of the step speed. A threshold helps distinguish between steps considered slow and steps considered fast, as shown in Fig. 17.

5. CONCLUSIONS

In this paper, we have presented three applications exploiting a natural human computer interfacing system implemented by means of a WiMoCA wireless body area sensor network. We showed how WiMoCA is able to handle diverse application requirements, thanks to its flexibility. We performed tests in-the-field to stress the capability of the network. We proved that our custom-designed, vertically integrated solution for body area sensor networks allows for the implementation of a wide range of HCI applications. Future work addresses the integration of other kinds of sensors to collect more physiological data from the user, and we plan to add vibro-tactile actuators for bio-feedback provisioning. Moreover, we are already exploring energy scavenging techniques to apply on the basic node and evaluating the use of the Zigbee standard protocol.

REFERENCES

[1] www.onbalance.com. Ref Type: Internet Communication, 2005.

[2] K. Aminian, E. D. Andres, K. Rezakhanlou, C. Fritsch, Y. Schutz, M. Depairon, P.-F. Leyvraz, and P. Robert, "MOTION ANALYSIS IN CLINICAL PRACTICE USING AMBULATORY ACCELEROMETRY, " Lecture Notes i Computer Science, 1537(-):1–11, 1998.

[3] J. Angesleva, I. Oakley, S. Hughes, and O. S., "BODY MNEMONICS: PORTABLE DEVICE INTERACTION DESIGN CONCEPT," Proceedings of the 16th annual ACM symposium on User interface software and technology, 2003.

[4] E. Bachmann, X. Yun, D. McKinney, R. McGhee, and M. Zyda. "DESIGN AND IMPLEMENTATION OF MARG SENSORS FOR 3-DOF ORIENTATION MEASUREMENT OF RIGID BODIES," Proceedings. ICRA '03. IEEE International Conference on Robotics and Automation 2003, September 2003, pages 14–19

[5] L. Bao and S. S. Intille. "ACTIVITY RECOGNITION FROM USER ANNOTATED ACCELERATION DATA," In Proc. of Pervasive 2004, the Second International Conference on Pervasive Computing, Springer, pages 1–17, 2004.

[6] R. Barbieri, E. Farella, A. Acquaviva, L. Benini, and B. Riccò, "A low-power motion capture system with integrated accelerometer," First IEEE Consumer Communications and Networking Conference, 2004. CCNC 2004., pages 418–423, January 2004.

[7] D. Brunelli, E. Farella, L. Rocchi, M. Dozza, L. Chiari, and L. Benini, "Bio-feedback system for rehabilitation based on wireless body area network," Proceedings of Fourth IEEE international Conference on Pervasive Computing (Percom06) - Workshop Ubicare 2006, pages 527– 531.

[8] J. Bussmann, W. L. J. Martens, J. H. M. Tulen, F. C. Schasfoort, H. J. G. V. D. Berg-Emons, and H. J. Stam, "Measuring daily behaviour using ambulatory accelerometry: The activity monitor," Behavior Research Methods, Instruments and Computers, 2001, 33(3):349–356.

[9] M. Caccamo, L. Y. Zhang, L. Sha, and G. C. Buttazzo, "An implicit prioritized access protocol for wireless sensor networks," IEEE Real-Time Systems Symposium, December 2002, pages 39– 48

[10] M. Canina, A. Rovetta, F. Pasolini, M. Tronconi, and E. Chiesa, "Innovative system for the accumulation of energy of the step in a limb prosthesis," Proceedings of the 11th World Congress in Mechanism and Machine Science, April 2004.

[11] D. Cattaneo, M. Ferrarin, W. Frasson, and A. Casiraghi, "Head control: volitional aspects of rehabilitation training in patients with multiple sclerosis compared with healthy subjects," Arch. Phys. Med. Rehabil., 86:1381–1388, 2005.

[12] A. D. Cheok, K. G. Kumar, and S. Prince, in Proceedings of the Sixth International Symposium on InWearable Computers, 2002. (ISWC 2002).

[13] L. Chiari, M. Dozza, A. Cappello, F. B. Horak, V. Macellari, and D. Giansanti, "Audio-biofeedback for balance improvement: An accelerometry-based system," Biomedical Engineering, IEEE Transactions, 52:2108–2111, 2005.

[14] B. Coley, B. Najafi, A. Paraschiv-Ionescu, and K. Aminian, "Stair climbing detection during daily physical activity using a miniature gyroscope," Gait and Posture, 22(4):287–294, December 2005.

[15] K. M. Culhane, M. O'Connor, D. Lyons, and G. M. Lyons, "Accelerometers in rehabilitation medicine for older adults," Age Ageing, 34(6):556–560, 2005.

[16] K. Delac and M. Grgic, "A survey of biometric recognition Methods," In, Proceedings of the 46th International Symposium on Electronics in Marine, 2004 (Elmar 2004), pages 184 – 193, 16-18 June 2004.

[17] G. Faidley, J. Hero, K. Lee, B. Lwakabamba, R. Walstrom, F. Chen, J. Dickerson, D. Rover, R. Weber, and C. Cruz-Neira, "Developing an integrated wireless system for fully immersive virtual reality environments," In Proceedings of the Eighth International Symposium on Wearable Computers (ISWC 2004), 2004, volume 1, pages 178 – 179, 31 Oct.-3 Nov. 2004.

[18] E. Farella, D. Brunelli, L. Benini, B. Riccò, and M. E. Bonfigli, "Pervasive computing for interactive virtual heritage," IEEE Multimedia, 12(3):46– 58, July-Sept 2005.

[19] E. Farella, S. O ´ Modhrain, L. Benini, and B. Riccò, "Gesture signature for ambient intelligence applications: a feasibility study" In Proceedings of the, 4th International Conference on Pervasive Computing (PERVASIVE 2006), Dublin, Ireland, pages 288–304, May 7-10 2006.

[20] E. Farella, A. Pieracci, D. Brunelli, L. Benini, and B. Riccò, "Design and implementation of wimoca node for a body area wireless sensor network," Proc. of Systems Communications 2005 IEEE SENET, pages 342–347, August 2005.

[21] E. Foxlin and M. Harrington, "Weartrack: A self-referenced head and hand tracker for wearable computers and portable vr.," proceedings of the Fourth International Symposium on Wearable Computers, pages 155–162, October 2000.

[22] A. R. Golding and N. Lesh, "Indoor navigation using a diverse set of cheap, wearable sensors," Proceedings of the 3rd IEEE International Symposium on Wearable Computers (ISWC '99), 1999.

[23] T. Harada, H. Uchino, T. Mori, and T. Sato, "Portable orientation estimation device based on accelerometers, magnetometers and gyroscope sensors for sensor network," Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI2003), pages 191– 196, July 2003.

[24] R. Headon and G. Coulouris. "Supporting gestural input for users on the move," in Proceedings Eurowearable, pages 107–112, September 2003.

[25] www-micrel.deis.unibo.it/wsn/myvideo/walking0001.wmv.

[26] http://www.isense.com/products/prec/ic3/wirelessic3.htm.

[27] http://www.st.com/stonline/prodpres/dedicate/mems/mems.htm.

[28] http://www.st.com/stonline/prodpres/dedicate/mems/products/products.htm.

[29] http://www.st.com/stonline/products/literature/ds/11115/lis3lv02dq.pdf.

[30] http://www.xbow.com/Products/Wireless Sensor Networks.htm.

[31] G. B. Jarnlo and K. G. Thorngren, "Standing balance in hip fracture patients. 20 middle-aged patients compared with 20 healthy subjects," Acta Orthop. Scand., 62:427–434, 1991.

[32] E. Jovanov, A. Milenkovic, C. Otto, and P. C. de Groen, "A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation," Journal of Neuro-Engineering and Rehabilitation, 2:6–16, 2005.

[33] H. Junker, P. Lukowicz, and G. Troster, "Padnet: Wearable physical activity detection network," Proceedings of the 7th IEEE International Symposium on Wearable Computers, 2003 (ISWC '03).

[34] J. M. Kahn, R. H. Katz, and K. S. J. Pister, "Emerging challenges: Mobile networking for smart dust," Journal of Commun., 2(3):188, 2000.

[35] N. Kern, B. Schiele, and A. Schmidt, "Multi-sensor activity context detection for wearable computing," Proceedings of European Symposium on Ambient Intelligence, pages 220–232, November 2003.

[36] S.-W.Lee and K. Mase, "Recognition of walking behaviors for pedestrian navigation,"

Proceedings of the 2001 IEEE International Conference on Control Applications, 2001 (CCA '01), pages 1152 – 1155.

[37] S.-W. Lee and K. Mase, "Activity and location recognition using wearable sensors," IEEE Pervasive Computing, 1(3):24–32, Jul. Sept 2002.

[38] B. E. Maki and W. E. McIlroy, "Postural control in the older adult," Clin. Geriatr. Med., 12:635–658, 1996.

[39] M. J. Mathie, A. C. F. Coster, N. H. Lovell, and B. G. Celler, "Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement," Physiol Meas., 25:R1–20, 2004.

[40] J. Mäntyjärvi, M. Lindholm, E. Vildjiounaite, S.-M.Mäkelä, and H. Ailisto, "Identifying users of portable devices from gait pattern with accelerometers," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2005 (ICASSP '05), volume 2, pages ii/973 – ii/976, 2005.

[41] S. Morris and J. Paradiso, "Shoe-integrated sensor system for wireless gait analysis and real-time feedback," Proceedings of the Second Joint Conference on EMBS/BMES, 2002, 3:2468–2469, 23-26 Oct 2002.

[42] I. Oakley, J. Angesleva, S. Hughes, and S. O'Modhrain, "Tilt and fill: Scrolling with vibrotactile display," In Proc. of EuroHaptics, June 2004, pages 313–323.

[43] J. Paradiso, "Footnotes: Personal reflections on the development of instrumented dance shoes and their musical applications," Quinz, E., ed., Digital Performance, Anomalie, 2:34–49, 2002.

[44] J. K. Perng, B. Fisher, S. Hollar, and K. S. Pister, "Acceleration sensing glove (asg)," In IEEE Symposium on Wearable Computers, pages 178–180, October 1999.

[45] C. Randell and H. Muller, "Context awareness by analyzing accelerometer data," Proceedings of the 4th IEEE International Symposium on Wearable Computers (ISWC '00), 2000.

[46] R.Moe-Nilssen, "A new method for evaluating motor control in gait under real-life environmental conditions," part 1: The instrument. Clin Biomech (Bristol, Avon)., 13(4-5):320–327., Jun 1998.

[47] L. Rocchi, L. Chiari, and F. B. Horak, "Effects of deep brain stimulation and levodopa on postural sway in parkinson's disease," J. Neurol. Neurosurg. Psychiatry, 73:267–274, 2002.

[48] M. Sama, V. Pacella, E. Farella, L. Benini, and B. Riccò, "3did: a low-power, low-cost hand motion capture device," Proceedings of IEEE Design, Automation and Test in Europe Conference and Exhibition (DATE-06), 2006, 6-10 March 2006.

[49] J. M. Schmit, D. I. Regis, and M. A. Riley, "Dynamic patterns of postural sway in ballet dancers and track athletes," Exp. Brain Res., 163: 370–378, 2005.

[50] J. Stankovic, T. Abdelzaher, C. Lu, L. Sha, and J. Hou, "Realtime communication and coordination in embedded sensor networks," Proceedings of the IEEE, volume 91, pages 1002–1022, July 2003.

[51] G. Uswatte, W. H. R. Miltner, B. Foo, M. Varma, S. Moran, and E. Taub, "OBJECTIVE MEASUREMENT OF FUNCTIONAL UPPEREXTREMITY MOVEMENT USING ACCELEROMETER RECORDINGS TRANSFORMED WITH A THRESHOLD FILTER," Stroke, 31(3):662–7, Mar 2000.

[52] T. van Dam and K. Langendoen, "AN ADAPTIVE ENERGY EFFICIENT MAC PROTOCOL FOR WIRELESS SENSOR NETWORKS," Proceedings of the 1st international conference on Embedded networked sensor systems (SenSys '03), ACM Press, 2003, pages 171–180.

[53] P. H. Veltink, H. B. Bussmann, W. de Vries, W. L. Martens, and R. C. V. Lummel, "DETECTION OF STATIC AND DYNAMIC ACTIVITIES USING UNIAXIAL ACCELEROMETERS," IEEE Trans Rehabil. Eng., 4(4):375–85, Dec 1996.

[54] B. Vigna. "MORE THAN MOORE: MICRO-MACHINED PRODUCTS ENABLE NEW APPLICATIONS AND OPEN NEW MARKETS," Technical Digest of IEEE International Electron Devices Meeting 2005, (IEDM 2005), pages 1– 8, Dec. 5 2005.

[55] D. A.Winter, F. Prince, J. S. Frank, C. Powell, and K. F. Zabjek, "UNIFIED THEORY REGARDING A/P AND M/L BALANCE IN QUIET STANCE," J. Neurophysiol., 75:2334–2343, 1996.

[56] R. Zhu and Z. Zhou, "A REAL-TIME ARTICULATED HUMAN MOTION TRACKING USING TRI-AXIS INERTIAL/MAGNETIC SENSORS PACKAGE," IEEE Transactions on Neural Systems and Rehabilitation Engineering, 12(2):295–302, 2004.

MANAGING IMPULSIVE INTERFERENCE IN IMPULSE RADIO UWB NETWORKS

Manuel Flury, Ruben Merz,
Jean-Yves Le Boudec

EPFL, School of Computer
and Communication Sciences

Wireless sensor networks are ideally built on low-cost, low-complexity nodes that have a low power consumption to guarantee a long network lifetime. These are all properties that can potentially be achieved with impulse radio ultra-wide band (IR-UWB). In addition, IR-UWB has a fine timing resolution enabling accurate ranging and localization capabilities. For all these reasons, IR-UWB is an extremely interesting physical layer technology for wireless sensor networks. In this article, we consider the management of impulsive interference in IR-UWB networks. Impulsive interference is due to uncoordinated concurrent transmissions. It occurs, for instance, when several independent piconets operate in close vicinity and is also present in some MAC layer proposals that allow concurrent transmissions. If not properly addressed, impulsive interference can severely affect the throughput and energy consumption of an IR-UWB network; as such, it already needs to be taken into account in the design phase. First, we show that impulsive interference is a serious concern for IR-UWB networks. Second, we present techniques at the physical layer and at the link layer to cope with and combat such interference efficiently. Finally, we present DCC-MAC as an example of an interference-aware design.

1. INTRODUCTION

For the design of wireless networks, there are two choices with respect to interference: we can design a system that tries to control or even prevent interference, or we can intentionally allow interference. Systems that let interference happen use some form of adaptability to deal with the constantly changing environment. Systems to control or prevent interference use mechanisms such as tight power control, orthogonal communication channels, or mutual exclusion [1].

However, even in systems designed to control interference, there are always numerous external factors that are beyond the control

of the system designer. For instance, there might be coexisting, non-coordinated piconets that interfere with each other. This external interference is difficult to foresee, and adaptive mechanisms to cope with it are required.

We consider non-coordinated systems based on impulse radio UWB (IR-UWB) physical layers that allow concurrent transmissions without power control [2], [3]. Data transmission at the physical layer occurs in sequences of very short pulses[1] with a large pulse repetition time (PRT). The most frequently used physical layer model [4] is illustrated in Fig. 1 and briefly introduced in the following. Time is divided into frames of length T_f. Each user transmits one pulse of length T_p per frame. To provide some multi-access capability, a frame is further subdivided into non-overlapping chips of length T_c, where $T_c \geq T_p$. Each user chooses the chip in which to transmit its pulse randomly according to a (pseudo-random) *time-hopping sequence* (THS).

Such systems are subject to impulsive, non-Gaussian interference created by the system itself, or by other, similar systems. On Fig. 2, we can clearly observe the detrimental effect of impulsive interference on an IR-UWB physical layer. Further, like any other UWB system [5], they have to coexist with existing narrowband technologies like 802.11. Managing interference to and from such coexisting technologies has been extensively studied and is out of the scope of this article. In this paper, we concentrate on impulsive interference. The main source of impulsive interference in IR-UWB systems are pulse collisions between concurrently transmitting sources. Pulse collisions occur even though nodes from different piconets generally use different THSs. This is due to the fact that THSs in IR-UWB are usually not orthogonal

1. Or short bursts of short pulses

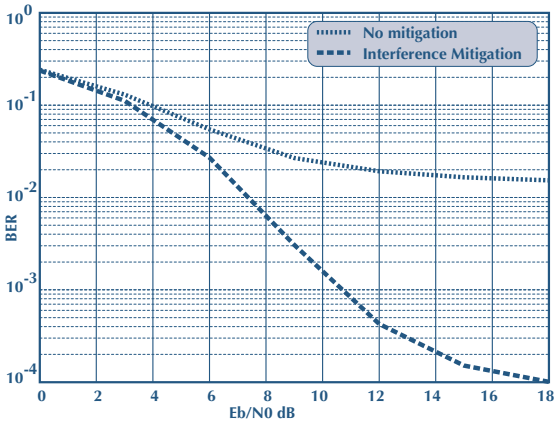

Figure 2: In uncoordinated IR-UWB networks, some form of interference mitigation at the physical layer is needed. We show the bit error rate (BER) versus signal-to-noise ratio at the receiver for a system with and without interference mitigation. The mitigation scheme used here is the one using interference modeling (further described in Section 2.3.2); the scenario for the simulation is the same as in Fig. 4. It can clearly be seen that the performance degradation is huge when not mitigating the effect of interference. Interference mitigation at the physical layer is discussed in Section 2.2, and a possible solution is given in Section 2.3.

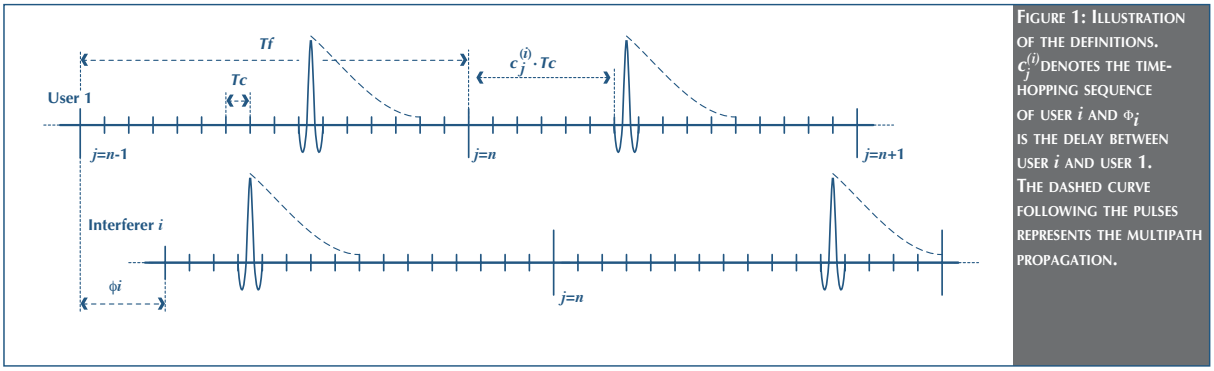

Figure 1: Illustration of the definitions. $c_j^{(i)}$ denotes the time-hopping sequence of user i and Φ_i is the delay between user i and user 1. The dashed curve following the pulses represents the multipath propagation.

and therefore do not completely prevent collisions. Furthermore, even if they were perfectly orthogonal, a tight synchronization between all the nodes in different piconets would be needed to prevent interference caused by misaligned THSs.

We focus on techniques and schemes that are used to react and adapt to interference. We do not discuss protocols or techniques that try to prevent or control interference (see [1] and the references therein). A multipath propagation channel at the physical layer further worsens the situation. The larger the delay spread of the channel, the more a pulse is spread in time. This increases the probability of pulse collisions. As IR-UWB systems are likely to be used in environments exhibiting severe multipath (indoor, factories, etc.), this is a serious issue. Another factor that increases the probability of pulse collisions is the number of users trying to transmit simultaneously. Even in systems with a generally low duty-cycle, it can happen that a lot of users access the channel at the same time. An example is a sensor network detecting a fire outbreak. In this case, a specific event triggers simultaneous transmissions from a large number of nodes.

Finally, one additional important factor concerning interference is the near-far effect. As the systems under consideration do not make use of power control, interferers close to the receiver might not have a signal of much higher strength than that of the user of interest. To ensure that small portions of these high power signals do not predominate the received signal, they have to be mitigated to prevent a huge performance loss.

Note that in a mobile ad hoc network, not only interference, but also the variable environment calls for adaptability of the system. Additionally, systems that try to prevent interference usually need tighter control than systems that let interference happen. This is often undesirable in an uncoordinated ad hoc network.

Impulsive interference in IR-UWB systems reduces the signal-to-interference-and-noise ratio (SINR) at the receiver. It affects the quality of the radio link, producing more packet losses, which result in an overall rate reduction and an increased energy consumption. Interference has a large impact on the system performance and needs to be taken into account as early as in the design phase. As we further show in this paper, interference management is a cross-layer issue. It has to be dealt with at the physical layer level as well as at the link layer level.

On the physical layer, some form of interference mitigation (Section 2.2) is needed to deal with the near-far effect. The benefits of an interference mitigation scheme are depicted in Fig. 2.

On the link layer, adaptive retransmission techniques must be used. Also, the overall rate of a source has to be variable in order to be adapted to the current level of interference at the receiver. Systems with a fixed rate must be designed in order to sustain the worst possible operating conditions, typically a poor channel between a source and its destination. This in turn imposes a low overall rate. Systems with an adaptive rate can take advantage of good channel conditions to transmit with a higher rate. In the case of degraded channel conditions, their adaptability prevents complete communication outages.

We do not discuss the effect of these schemes on energy consumption. There is, of course, a trade-off. A better system performance reduces the number of retransmissions and hence decreases the energy consumption. On the other hand, more complex transceiver designs increase the energy-consumption.

The organization of this paper is as follows. In Section 2, we present techniques to combat interference on the physical layer. In Section 3, we discuss link layer techniques to cope with interference. In Section 4, the DCC-MAC protocol [2] is presented as a concrete example of an interference aware design which is a rate-adaptive medium access control (MAC) protocol for IR-UWB networks. Finally, we conclude the paper in Section 5.

2. COMBATING INTERFERENCE AT THE PHYSICAL LAYER

Combating interference matters to all functions provided by the physical layer, be it decoding, channel estimation, or timing acquisition and detection. We will present some possible

solutions for all of these functions in Section 2.3. Currently used techniques to combat interference on the physical layer can be divided into two classes, both of which are shortly discussed in the following.

2.1. Techniques based on joint decoding

These are extensions or adaptations to UWB of classical, well-established techniques that are also used in other systems like CDMA [6]. They aim at cancelling or suppressing interference by jointly estimating and decoding the signals of a large number of users. For example, a near-far interferer would be jointly received instead of being treated as interference. This annihilates the near-far effect and makes joint decoding potentially attractive. However, an optimal joint processing of all users [7] is mostly not possible due to its very high complexity. Therefore, suboptimal methods like minimum mean-square error (MMSE) multi-user detectors (MUD) or receivers employing successive interference cancellation (SIC) are used [8]. All of these methods share the common factor that the receiver has to acquire and actively decode each of the users. This might be perfectly suited for a centrally coordinated and synchronized system, where a base station communicates with a large number of users at the same time. However, with a distributed IR-UWB system, synchronizing the receiver with all the users is extremely complex and impractical. In addition, the complexity of the decoding operation is excessively high.

2.2. Techniques based on interference mitigation

In contrast to joint decoding, signals from interfering users are treated as a common interference term. Techniques based on interference mitigation try to reduce this interference term and to mitigate and reduce its effect on the performance of the physical layer. We distinguish two possible options, interference modeling and thresholding:

2.2.1. Interference modeling

The interference term is assumed to follow an underlying statistical model. The background noise is often directly incorporated in the interference model. A receiver using interference modeling proceeds in two steps. It first tries to estimate the model parameters. In a second step, this model is exploited to mitigate the effect of the interference. Modeling interference is important as it has been shown that simply assuming it to be Gaussian is not accurate [9].

In order to estimate the parameters of the model, techniques based on interference modeling can either follow a data-aided [10], [11], [12] or a blind approach [10]. In the data-aided approach, a training sequence known to the receiver is used. The receiver estimates the statistics of the interference model exploiting the knowledge about the training sequence. In the blind (non-data-aided) approach, the receiver jointly estimates the model parameters as well as the unknown data sequence.

2.2.2. Thresholding

A simple thresholding mechanism can be applied. Samples of the received signal that have an amplitude exceeding a certain threshold are assumed to have a large interference contribution[2], [13], [14], [8]. Although thresholding is easy to implement, an issue common to all thresholding schemes is the determination of the optimal threshold. This is often left as an open issue, or it boils down to assuming an AWGN multi-user interference (MUI) model and setting the threshold based on the estimated average received noise power [13].

2.3 MUI-aware physical layer system design proposals

We now present some proposals for physical layer core functionalities that were specifically designed for asynchronous IR-UWB systems subject to impulsive interference and that use the above mentioned techniques.

2.3.1. Timing acquisition and detection

In conventional detection methods, the transmitter prepends each packet with an acquisition preamble known to the receiver. The receiver correlates the received signal with this acquisition preamble and performs a threshold check. If the output of the correlator exceeds a certain threshold, a good match between

the received signal and the acquisition preamble is assumed, and detection of the packet is declared. These methods have a severe drawback when MUI is present. If one of the pulses of the acquisition preamble at the receiver is aligned with a pulse of a near-by interferer, this interfering pulse can affect the correlation significantly. Consequently, a small number of aligned interfering pulses can dominate the output of the correlator and lead to a wrong detection. In [14], a power independent detection (PID) method that addresses this problem was proposed. As for the conventional methods, the PID uses thresholding. However, it splits the correlation with the whole acquisition preamble into a set of elementary correlations. Each elementary correlation corresponds to only one pulse of the acquisition preamble. A first threshold is applied at the output of each elementary correlation. If the energy captured by the elementary correlation exceeds the first threshold, detection of the corresponding pulse is declared. A second threshold is then applied to the number of detected pulses. If the number of detected pulses exceeds the second threshold, detection of the packet is declared. This procedure makes sure that all the pulses of the acquisition preamble contribute equally to the final decision. This procedure is therefore resistant to near-far interference.

2.3.2. Channel estimation and decoding

As already mentioned, a Gaussian model is not well suited to model MUI in an IR-UWB system. A popular non-Gaussian model is the Gaussian Mixture Model (GMM). The GMM assumes that the interference has an underlying probability distribution formed by a mixture of Gaussians with different variances. Each interference term is then assumed to be generated by one of these mixture components. The GMM seeks to classify each sample and typically attributes samples with high interference to mixture components with high variances. In [11], the GMM is proposed as a MUI model for IR-UWB. We will show how to perform channel and interference statistics estimation based on this model using a data-aided approach.

The GMM assumes that the mixture components are independently chosen. However, due to the multipath nature of the channel, this is not necessarily true, as samples with a high interference level are likely to occur in bursts. Therefore, we propose [12] to introduce correlation by modeling the sequence of mixture components with a homogeneous Markov chain. The resulting MUI model is a hidden Markov model (HMM), where each state is associated with a Gaussian output distribution. The GMM is just a special case of the more general HMM, where the choice of the next state is independent of the current state[2]. We find that the HMM is effectively better in modeling MUI than the GMM. However, the performance difference is not that huge and comes at the cost of increased complexity.

In [12], we also propose a coherent RAKE receiver that makes use of a combination of thresholding and interference modeling to mitigate interference in the decoding process and accounts for the multipath nature of UWB channels. Interference modeling is done using a data-aided approach. Let us assume for simplicity that the interference model (GMM or HMM) only has two states, s_1 and s_2, where s_1 corresponds to a low interference level and s_2 to a high interference level. The receiver proceeds in two steps. In a training step, the channel coefficients as well as the variances, $\sigma_{s_1}^2$ and $\sigma_{s_2}^2$, associated with each of the two states, are estimated based on the known training sequence. In the subsequent data reception step, the receiver estimates for each sample y_n the probability $\gamma_{s_1}(n)$ that it has an interference term generated by state s_1[3]. Before passing the received samples to the decoder, the receiver multiplies each sample with the following weighting vector:

$$w(n) = \frac{\gamma_{s_1}(n)}{\sigma_{s_1}^2} + \frac{\gamma_{s_2}(n)}{\sigma_{s_2}^2}$$

2. The reader is invited to read [12] for a more mathematically rigorous definition of the respective interference models.

3. Here we only consider a two state model, so the probability that the interference term of y_n was generated by s_2 is of course $\gamma_{s_2}(n)=1-\gamma_{s_1}(n)$.

Consequently, samples with an interference term that stems with high probability from s_2 get penalized through the factor $w(n)$. This ensures that samples with high interference do not contribute excessively to the decision made by the decoder. (Note the similarity to the power independent detection method described in Section 2.3.1). The effect of applying this weighting factor is shown in Fig. 3. Our coherent RAKE receiver also employs thresholding in addition to the interference modeling procedure described before. This is necessary in a data-aided approach since we are facing three different interference scenarios, only two of which are resolved by interference modeling. If interference occurs during packet reception, it must fall into one of the following three categories:

1. interference is present during both training and data reception (type 1)
2. interference is present during training only (type 2)
3. interference is present during data reception only (type 3)

Interference of type 1 is taken care of by interference modeling. Ideally, we estimate the interference during the training phase and then deal with it during data reception as explained above. Interference of type 2 should do even less harm; we have estimated it, but it is not present during data reception. Interference of type 3 is more difficult to tackle since it is not present during training. Therefore, the estimated variances of the interference term will be rather small (on the order of the background noise variance). Samples with a lot of interference will then still get a relatively high weight.

Hence, we propose the following thresholding mechanism. After the training phase, we determine the largest of the estimated variances. In the case of the two-state model discussed here, this is σ_{s_2}. We then determine a threshold ν, such that $P(X \geq \nu) \leq \varepsilon$, where $X \sim N(0, \sigma_{s_2})$, and ε is some predetermined small probability. We then erase the samples with an estimated interference and noise term exceeding the threshold ν by setting $\gamma_{s_1}(n)=\gamma_{s_2}(n)=0$ for these samples. This ensures that the samples that cannot be explained by the estimated interference model with high probability do not contribute to the decision made by the decoder. Interference of type 3 is thus mitigated by detecting a deviation from the estimated model. A similar thresholding approach, rejecting samples suffering from high interference, has

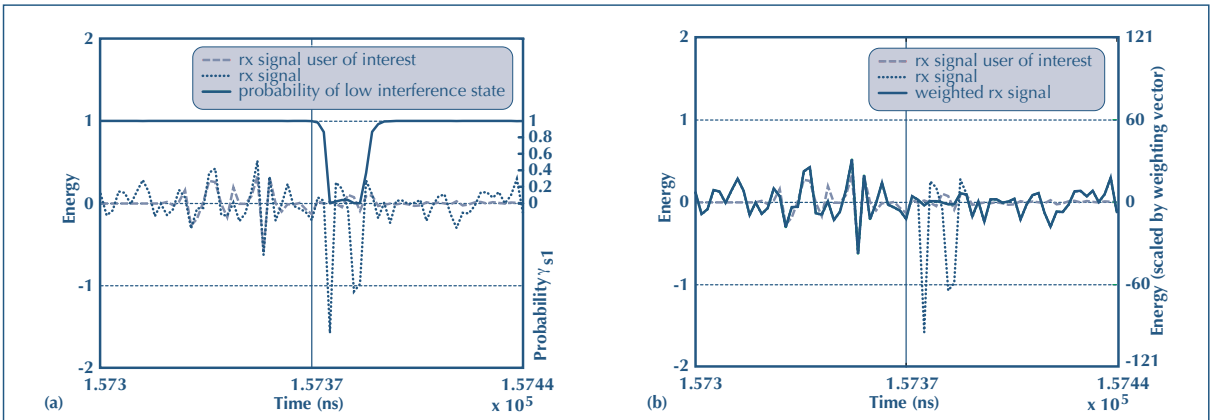

FIGURE 3: HERE WE SHOW HOW AN ALGORITHM BASED ON INTERFERENCE MODELING PERFORMS INTERFERENCE MITIGATION. A TWO-STATE HIDDEN MARKOV MODEL IS ASSUMED FOR THE MUI. IN (A), ONE PULSE OF THE RECEIVED SIGNAL AND ITS COMPONENT CORRESPONDING TO THE USER OF INTEREST IS SHOWN. FOR EACH SAMPLE, THE RECEIVER ESTIMATES THE PROBABILITY THAT IT HAS A LOW CONTRIBUTION FROM INTERFERING USERS (LOW INTERFERENCE STATE) OR THAT IT IS POLLUTED WITH A HIGH INTERFERENCE TERM (HIGH INTERFERENCE STATE). THE ESTIMATED PROBABILITY OF BEING IN THE LOW INTERFERENCE STATE IS ALSO SHOWN IN THE LEFT FIGURE. WE CAN SEE THAT THE ALGORITHM NICELY IDENTIFIES THE PART THAT SUFFERS FROM A HIGH INTERFERENCE TERM. BASED ON THIS ESTIMATION, THE RECEIVER DESIGNS A WEIGHT VECTOR THAT IS APPLIED TO THE RECEIVED SIGNAL. DIAGRAM (B) ADDITIONALLY SHOWS THE RECEIVED SIGNAL AFTER IT HAS BEEN MULTIPLIED WITH THE WEIGHTING VECTOR, AND WE CAN SEE THAT THE MUI HAS BEEN SUCCESSFULLY REMOVED.

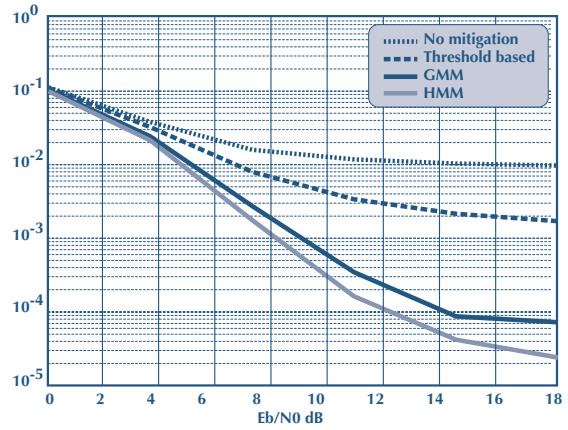

FIGURE 4: WE COMPARE OUR INTERFERENCE MITIGATION TECHNIQUE WITH A RECEIVER THAT NEGLECTS MULTI-USER INTERFERENCE (MUI) COMPLETELY AND WITH A RECEIVER PERFORMING ONLY SIMPLE THRESHOLDING. PHYSICAL LAYER PACKETS ARE GENERATED ACCORDING TO A POISSON PROCESS AT HALF THE PEAK DATA RATE. THE CHANNEL MODEL WE SIMULATE IS THE 802.15.4A INDOOR NLOS MODEL. FURTHER, WE HAVE FOUR NEAR INTERFERERS WITH POWER LEVELS OF 10DB, 13DB, 16DB AND 20DB WITH RESPECT TO THE USER OF INTEREST. IT CAN BE SEEN, THAT THE PERFORMANCE GAIN FROM MODELING THE INTERFERENCE IS SIGNIFICANT. USING THE MORE SOPHISTICATED HMM TO CHARACTERIZE MUI GIVES AN ADDITIONAL GAIN COMPARED TO THAT GIVEN THE GMM MODEL

been proposed in [8] for MUD in a synchronous UWB system, and in [13], [2] as a stand-alone method without interference modeling and without neglecting the multipath nature of the UWB channel. The performance gain of our proposal over these simple thresholding schemes can be seen in Fig. 4.

3. MANAGING INTERFERENCE AT THE LINK LAYER

In this section, we discuss several link-layer techniques that can be used to react and adapt to interference. Link layer techniques control transmission parameters and the retransmission behavior at the sender. Their goal is to adapt to the level of interference experienced at the receiver. When the interference at the receiver is low, adaptive transmission techniques allow for the increase of the throughput. On the other hand, when interference at the receiver is high, adaptive techniques avoid communication outages and ensure a minimum throughput.

In the case of IR-UWB communications, the transmission parameters to adapt can be the modulation order (number of bits per symbol), the power, the rate of the channel code, or the processing gain.

In order to adapt to these parameters, the transmitter must have an estimate of the level of interference at its intended receiver. In the context of uncoordinated networks, most techniques make use of feedback information from the receiver to the sender. Feedback information from the receiver can take various forms. It is often a function of the SINR. Several other examples can be found in [15]. However, with a UWB physical layer, measuring the SINR is difficult in practice due to the very low transmit power of UWB signals. For instance, the DCC-MAC protocol discussed in Section 4 relies on information produced by the channel decoder rather than on physical layer measurements.

3.1 Adaptation of the transmission parameters to the level of interference

Adaptive modulation [16] allows for the efficient adaption of the spectral efficiency to the level of interference. Adaptive modulation essentially varies the number of bits per symbols. In the case of IR-UWB, a simple and efficient technique is to use M-ary PPM [17]. A further technique is the use of joint modulation and coding, such as bit-interleaved coded modulation (BICM) [18], [19]. Even in the presence of multi-user interference (MUI), such a technique can considerably increase the throughput of an IR-UWB link [20].

The amount of redundancy of the channel code, and hence the rate, can be adapted to the level of interference. Practical schemes such as rate-compatible punctured convolutional (RCPC) codes [21] can be used. One of their main advantages is that only one decoder is necessary at the receiver for a given family of RCPC codes. Another interesting feature is that they can be used with incremental redundancy techniques (see Section 3.2). An issue that arises with IR-UWB physical layers and channel coding is the detrimental effect that impulsive interference can cause.

If soft-decision decoding is used, a large interference sample (for instance, in the case of near-far interference) can propagate through the trellis of the decoder and result in several decoding errors. However, if hard-decision decoding is used, this effect is prevented. But then, the performance when only regular Gaussian noise is present is impacted. Intuitively, the optimal decoding policy should consist of an adaptive combination of hard-decision when strong interferers are present and soft-decision otherwise [2]. Hence, an interference mitigation scheme (Section 2.2) should be used.

Adapting the processing gain for IR-UWB has been suggested in [17], [22]. It is possible to either change the average pulse repetition frequency or to change the number of pulses per symbol[4]. The issues of near-far interference also apply in this case. Changing the processing gain also has an impact on the average emitted power.

Note that adaptive modulation and adaptive channel coding are rate adaptation techniques. They are also procedures that are local to a single sender and receiver pair, that is, only communication between the source and the destination is required to perform them.

With power control [23], [24], a transmitter ensures that the received SINR at its destination remains higher than a given threshold. This threshold depends on the current level of interference at the receiver.

Contrary to adaptive modulation and adaptive channel coding, performing power control is a global procedure. Coordination is required not only between the source and its destination, but also with the neighbors of the transmitter. The transmitter should make sure that it does not destroy any ongoing transmission by reducing the SINR at receiving nodes in its vicinity. This requires the estimation of the channel gain between the transmitter and each node in the range of the transmitter.

4. Pulse repetitions are a special case of channel coding. Indeed, it is nothing but a repetition code.

The choice of rate adaptation and/or power control for IR-UWB networks is analyzed in [25]. When the objective is to maximize the overall throughput of the network, it turns out that the optimum is to use rate adaptation and no power control. If the primary objective is to minimize the energy consumption, this is still an open issue. Still, some results in [1] suggest that rate adaptation with no power control is not far from being optimal.

3.2. Adaptive retransmission techniques: Hybrid-ARQ with incremental redundancy

The techniques discussed in the previous section allows for the adaption of the parameters of the transmitted signal to the estimated level of interference at the receiver. However, there are two issues associated with these techniques. First, the feedback returned by the receiver is only an estimation of the level of interference at the receiver. Second, the level of interference can change significantly between the time the feedback is received and the time where the transmission occurs. In the first case, a solution is to include a safety margin. However, in the second case, an increase of the level of interference can arise and induce an error on the transmission between the source and the destination.

Hence, there is a need for an efficient retransmission procedure in case of a transmission failure. Such schemes have been extensively studied in the literature. They are denominated under the general term of Automatic Retransmission reQuest (ARQ). For an extensive overview of ARQ mechanism, the reader can consult [26], [15] and the references therein. In the remainder of this section, we will concentrate on adaptive mechanisms, the so-called hybrid-ARQ schemes.

In its simplest form, an ARQ scheme retransmits the same packet until successful reception occurs at the receiver. The feedback is binary and indicates whether or not the packet was properly received. However, this scheme will fail in the event of a strong and lasting interference; indeed, the data transmission will fail at each retransmission. Therefore, current ARQ techniques are adaptive. In most cases, the ARQ mechanism

is combined with a variable-rate channel code. If the initial data transmission fails, the retransmission occurs with the data encoded with a more powerful code, i.e., at a lower rate[5].

A further improvement of this scheme can be obtained by using incremental redundancy. Instead of retransmitting the whole data encoded at a lower rate, only the coded information necessary to obtain the lower rate is sent. For instance, channel codes such as the RCPC code in [21], [27] can provide incremental redundancy. For a specific family of RCPC codes, a code of a given rate is a "subset" of all the codes with a higher rate.

Note that these schemes do not specify how the rate should be adapted. Hence, there is a large amount of freedom left for how the overall retransmission mechanism can be designed.
The design of an adaptive ARQ scheme is largely dictated by the flexibility of the channel code and by the type of feedback available between a receiver and the sender.

4. DCC-MAC: AN UNCOORDINATED MAC PROTOCOL FOR UWB NETWORKS WITH RATE ADAPTATION AND INTERFERENCE MITIGATION

In this section, we present the case study of a system designed to be interference-aware. We consider the organization of non-coordinated and asynchronous medium-access (MAC) protocol for UWB networks. One proposal is the DCC-MAC protocol [2]. DCC-MAC is an interference-aware design that is conceived to operate in a flawless manner in the presence of strong impulsive interference.

In order to compare the performance of DCC-MAC against a non interference-aware protocol, we compare DCC-MAC with the $(UWB)^2$ protocol. $(UWB)^2$ is a more recent proposal that is not interference-aware and does not support any mechanisms to combat impulsive interference. The main characteristics of the two protocols are summarized in Table 1.

DCC-MAC	$(UWB)^2$
Interference-aware	Non interference-aware
Interference mitigation	n/a
Rate adaptation	n/a
No control channel required, no RTS-CTS	Common control channel with RTS-CTS

TABLE 1: MAIN ASPECTS AND DIFFERENCES OF THE DCC-MAC AND THE $(UWB)^2$ PROTOCOL FOR NON-COORDINATED IMPULSE-RADIO UWB NETWORKS

In the following, we first describe the two main components of DCC-MAC that permit to combat interference, namely rate-adaptation and interference mitigation. In addition, we briefly describe the main aspects of our protocol. Then, we present simulation results that compare DCC-MAC with the $(UWB)^2$ protocol [3].

4.1. Rate-adaptation and interference mitigation in the DCC-MAC protocol

The main ingredients of DCC-MAC to combat interference are a rate-adaptation mechanism and an interference-mitigation scheme. Rate-adaptation is obtained by using RCPC channel codes. The modulation and the processing gain is fixed. The family of RCPC codes is the one described in [27]. It offers a set of twenty-five channel code rates that can be extended to thirty. Only one pair of channel code encoder and decoder is necessary. The rate-adaptation scheme is based on an additive-increase, multiplicative-decrease (AIMD) policy similar to what is used by TCP. Whenever a packet is successfully received, the destination takes advantage of the decoding process to estimate the maximum rate at which the data transmission could have occurred [2]. The receiver subtracts a safety margin and returns this information back to sender in the acknowledgment packet. Hence, in the case of DCC-MAC, the feedback consists of the estimated rate at which the next data transmission should take place.

Interference mitigation is used, albeit in the simple form of thresholding. The mechanism is similar to what is explained in Section 2.2. The transmission of data to a destination is performed using a time-hopping sequence unique to the destination. This time-hopping sequence can be created by seeding a pseudo-random number generator with a unique

5. A different modulation could be used, but this is hardly done in practice.

identifier for the destination. Such an identifier can be, for instance, the hardware address.

A typical transmission consists of a data packet transmission from the source to the destination, an acknowledgment sent back by the destination and the transmission of an IDLE packet from the source. Hence, it has a simple design that does not require any common control channel nor the use of any RTS-CTS type of handshake.

Along with a subtle control of timers and a careful use of time-hopping sequences, the IDLE packet is necessary for the protocol to operate properly in the absence of carrier-sensing as well as in multi-hop environments [2].

4.2. Performance evaluation of the DCC-MAC and (UWB)^2 protocols

In order to emphasize the importance of an interference-aware design, we compare the DCC-MAC protocol with the (UWB)^2 protocol [3]. Contrary to DCC-MAC, (UWB)^2 needs a common control channel and uses an RTS-CTS handshake to arbitrate access to a destination. (UWB)^2 uses neither power-control nor rate-adaptation. Interference is not mitigated at the physical layer. Table 1 summarizes the main aspects and differences of the two protocols.

We use the ns-2 simulator [28] with an extension for UWB physical layers. The code for the UWB extension is available online at [29]. The parameters correspond to a typical 802.15.4a scenario. The maximum rate of the physical layer is 1 Mbit/s. For every scenario, the link distance is 10 meters. The transport protocol is UDP. The throughput is the saturation throughput.

In Fig. 5, the network throughput of the two protocols is compared in a multi-hop scenario. The topology is a line of n nodes where one extremity of the line sends to the other extremity. The throughput of (UWB)^2 drops dramatically as the number of hops increases. On the contrary, the throughput of DCC-MAC remains stable for more than three hops. In Fig. 7, the network throughput of the two protocols is compared in a near-far scenario. An example of the near-far topology for six nodes, i.e., three concurrent communication links, is represented in Fig. 6. Again, the throughput of the (UWB)^2 protocol decreases with the number of concurrent links. On the other hand, the DCC-MAC protocol can cope with the increasing number of concurrent transmissions. Additional simulation results for DCC-MAC can be found in [2] for various scenarios.

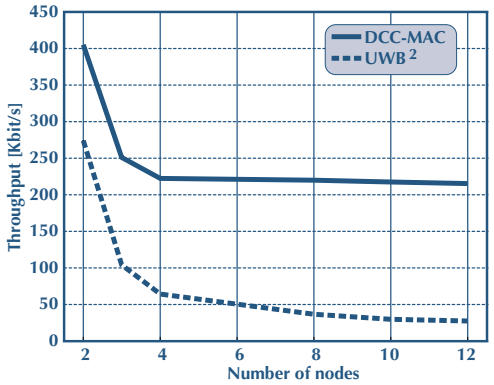

FIGURE 5: PERFORMANCE OF THE DCC-MAC PROTOCOL AND THE (UWB)^2 PROTOCOL IN A MULTI-HOP SCENARIO. THE TOPOLOGY IS A LINE OF NODES WITH A LINK DISTANCE OF 10 METERS. THE THROUGHPUT IS PLOTTED AGAINST THE NUMBER OF NODES. THE TRANSMITTER AND RECEIVER ARE LOCATED AT EACH EXTREMITY OF THE LINE. THE TRANSPORT PROTOCOL IS UDP. THE DCC-MAC PROTOCOL CLEARLY OUTPERFORMS THE (UWB)^2 PROTOCOL.

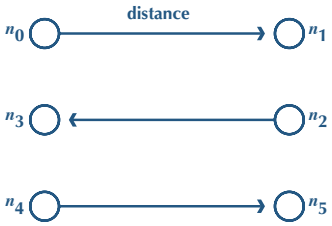

FIGURE 6: NEAR-FAR TOPOLOGY FOR SIX NODES, I.E., WITH THREE CONCURRENT COMMUNICATION LINKS. THE LINK DISTANCE IS d. THE DIRECTION OF THE COMMUNICATION IS ALTERNATED FROM ONE LINK TO ANOTHER.

5. CONCLUSION

We have discussed the management of impulsive interference in IR-UWB networks. We have shown that this kind of interference is an issue in IR-UWB and has therefore to be taken care of.

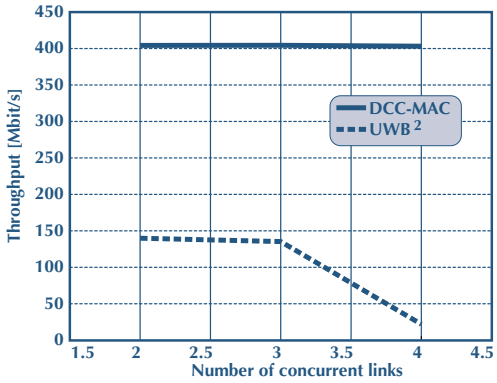

FIGURE 7: PERFORMANCE OF THE DCC-MAC PROTOCOL AND THE (UWB)^2 PROTOCOL IN A NEAR-FAR SCENARIO. THE LINK DISTANCE IS 10 METERS. THE THROUGHPUT IS PLOTTED AGAINST THE NUMBER OF CONCURRENT COMMUNICATION LINKS. THE TRANSPORT PROTOCOL IS UDP. THE DCC-MAC PROTOCOL CLEARLY OUTPERFORMS THE (UWB)^2 PROTOCOL. THE (UWB)^2 PROTOCOL SUFFERS FROM THE CONTENTION ON THE COMMON CHANNEL AS WELL AS THE RATE DROP DUE TO THE RTS-CTS EXCHANGE.

We have further presented several techniques and proposals that address impulsive interference on the physical as well as on the link layer. Additionally, interference could also be managed on the network layer. There is already some work on routing protocols that try to route packets such that interference is limited. There are also other aspects that we have left out. We have not discussed the effect of these schemes on energy consumption. There is of course a tradeoff. A better system performance reduces the number of retransmissions and hence decreases the energy consumption. On the other hand, more complex transceiver designs increase the energy-consumption. Another important aspect of IR-UWB is its ranging capability. As for detection or channel estimation, interference will most probably matter, and some ways to deal with it will have to be considered.

REFERENCES

[1] A. El Fawal, J.-Y. Le Boudec, R. Merz, B. Radunovic, J. Widmer, G. M. Maggio, "Tradeoff analysis of PHY-aware MAC in low-rate, low-power UWB networks," IEEE Communications Magazine, vol. 43, no. 12, pp. 147–155, December 2005.

[2] R. Merz, J. Widmer, J.-Y. Le Boudec, B. Radunovic, "A joint PHY/MAC architecture for low-radiated power TH-UWB wireless ad-hoc networks," Wireless Communications and Mobile Computing Journal, Special Issue on Ultrawideband (UWB) Communications, vol. 5, no. 5, pp. 567–580, August 2005.

[3] M.-G. Di Benedetto, L. Nardis, M. Junk, G. Giancola, "(UWB)^2: Uncoordinated, wireless, baseborn, medium access control for UWB communication networks," Mobile Networks and Applications, vol. 10, no. 5, October 2005.

[4] M. Z. Win, R. A. Scholtz, "Ultra-wide bandwidth time-hopping spread-spectrum impulse radio for wireless multipleaccess communications," IEEE Transactions on Communications, vol. 48, no. 4, pp. 679–691, April 2000.

[5] R. C. Qiu, H. Liu, X. Shen, "Ultra-wideband for multiple access communications," Communications Magazine, IEEE, vol. 43, no. 2, pp. 80–87, 2005.

[6] S. Verdù, **Multiuser Detection**. Cambridge University Press, 1998.

[7] Y. C. Yoon, R. Kohno, "Optimum multi-user detection in ultra-wideband (UWB) multiple-access communication systems," in ICC 2002 - IEEE International Conference on Communications, no. 1, Apr. 2002, pp. 812–816.

[8] E. Fishler, H. V. Poor, "Low-complexity multiuser detectors for time-hopping impulse-radio systems," IEEE Transactions on Signal Processing, vol. 52, no. 9, pp. 2561–2571, September 2004.

[9] B. Hu, N. Beaulieu, "Accurate evaluation of multiple-access performance in TH-PPM and TH-

BPSK UWB systems," IEEE Transactions on Communications, vol. 52, no. 10, pp. 1758–1766, Oct. 2004.

[10] V. Lottici, A. D'Andrea, U. Mengali, "Channel estimation for ultra-wideband communications," IEEE Journal on Selected Areas in Communications, vol. 20, no. 9, pp. 1638–1645, Dec. 2002.

[11] V. Cellini, G. Dona, "A novel joint channel and multi-user interference statistics estimator for UWB-IR based on Gaussian mixture model," in IEEE International Conference on Ultra-Wideband (ICU 2005), Sept. 2005, pp. 655–660.

[12] M. Flury, J.-Y. Le Boudec, "Interference Mitigation by Statistical Interference Modeling in an Impulse Radio UWB Receiver," in International Conference on Ultra-Wideband (ICUWB 2006), Waltham, MA, September 2006.

[13] W. M. Lovelace and J. K. Townsend, "Threshold discrimination and blanking for large near-far power ratios in UWB networks," IEEE Transactions on Communications, vol. 53, no. 9, pp. 1447–1450, Sept. 2005.

[14] A. El Fawal, J.-Y. Le Boudec, "A robust signal detection method for ultra wide band (UWB) networks with uncontrolled interference," IEEE Transactions on Microwave Theory and Techniques, 2006, to appear.

[15] S. Nanda, K. Balachandran, and S. Kumar, "Adaptation techniques in wireless packet data services," IEEE Communications Magazine, vol. 38, no. 1, pp. 54–64, January 2000.

[16] S. T. Chung, A. Goldsmith, "Degrees of freedom in adaptive modulation: a unified view," IEEE Transactions on Communications, vol. 49, no. 9, pp. 1561–1571, September 2001.

[17] N. August, R. Thirugnanam, and D. Ha, "An adaptive UWB modulation scheme for optimization of energy, BER, data rate," in UWBST, 2004, pp. 182–186.

[18] G. Caire, G. Taricco, E. Biglieri, "Bit-interleaved coded modulation," IEEE Transactions on Information Theory, vol. 44, no. 3, pp. 927–946, May 1998.

[19] Y. Souilmi, R. Knopp, "Code constructions for non-coherent on-off ultra-wideband systems," in IEEE international conference on Ultra Wide Band, September 2005, pp. 28–32.

[20] R. Merz, J.-Y. Le Boudec, "Effect of interfering users on the modulation order and code rate for UWB impulse-radio bit-interleaved coded M-ary PPM," in IEEE UWBnets workshop, October 2005.

[21] J. Hagenauer, "Rate-compatible punctured convolutional codes (RCPC codes) and their applications," IEEE Transactions on Communications, vol. 36, no. 4, pp. 389–400, April 1988.

[22] I. Guvenc, H. Arslan, S. Gezici, H. Kobayashi, "Adaptation of multiple access parameters in time hopping UWB cluster based wireless sensor networks," in IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS), October 2004, pp. 235–244.

[23] M. Krunz, A. Muqattash, S.-J. Lee, "Transmission power control in wireless ad hoc networks: challenges, solutions and open issues," IEEE Network, vol. 18, no. 5, pp. 8–14, Sept.-Oct 2004.

[24] F. Cuomo, C. Martello, A. Baiocchi, C. Fabrizio, "RADIO RESOURCE SHARING FOR AD HOC NETWORKING WITH UWB," IEEE Journal on Selected Areas in Communications, vol. 20, no. 9, pp. 1722–1732, December 2002.

[25] B. Radunovic, J. Y. Le Boudec, "OPTIMAL POWER CONTROL, SCHEDULING AND ROUTING IN UWB NETWORKS," IEEE Journal on Selected Areas in Communications, vol. 22, no. 7, pp. 1252–1270, September 2004.

[26] D. Costello, J. Hagenauer, H. Imai, S. Wicker, "APPLICATIONS OF ERROR-CONTROL CODING," IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2531–2560, October 1998.

[27] P. Frenger, P. Orten, T. Ottosson, A. Svensson, "RATE-COMPATIBLE CONVOLUTIONAL CODES FOR MULTIRATE DS-CDMA SYSTEMS," IEEE Transactions on Communications, vol. 47, no. 6, pp. 828–836, June 1999.

[28] "THE NETWORK SIMULATOR NS-2," http://www.isi.edu/nsnam/ns/."

[29] "UWB RESEARCH AT EPFL-IC," http://icawww1.epfl.ch/uwb/, 2006.

■ CONTACT: ST.JOURNAL@ST.COM ■

THE ALOHA ACCESS (UWB)^2 PROTOCOL REVISITED FOR IEEE 802.15.4A

Maria-Gabriella Di Benedetto, Luca De Nardis,
Guerino Giancola, Daniele Domenicali

School of Engineering
University of Rome La Sapienza

The IEEE 802.15.4a Task Group recently proposed Impulse Radio Ultra Wide Band (IR-UWB) for a physical layer that can provide combined communication and ranging in low data rate indoor/outdoor networks. At present, it is therefore particularly relevant to design IEEE 802.15.4a MAC strategies that are appropriately tailored on the physical layer. Previously, we proposed the Uncoordinated Baseborn Wireless medium access control for UWB networks (UWB)^2, a UWB-tailored MAC based on the low probability of pulse collision. The (UWB)^2 adopted the Aloha principle and enabled location-based network optimization by providing and storing estimates of distance between nodes. This paper first revisits the (UWB)^2 MAC protocol in view of its application to IEEE 802.15.4a. The structure of both control and data MAC protocol data units is defined based on the legacy 802.15.4 MAC in order to allow a seamless support, for both centralized and distributed network topologies, as defined in the parent standard. Secondly, this work extends and completes the analysis of (UWB)^2 since it takes into account multipath-prone channels. Channel parameters, for both indoor and outdoor propagation scenarios in Line Of Sight (LOS) and Non-Line Of Sight (NLOS) conditions, were derived from the channel model defined within the 802.15.4a channel sub-committee.
Results highlight that the (UWB)^2 protocol is robust to multipath, and provides high throughput and low delay, with performance scaling gracefully as a function of the number of users and the user bit rate. Results confirm and support the adoption of (UWB)^2 principles for low data rate UWB communications.

Index Terms - Ultra Wide Band, MAC, Low Data Rate

1. INTRODUCTION

Low data rate and low cost networks for mixed indoor/outdoor communications are of great interest in sensor and ad-hoc networking. The interest towards low data rate networks led in 2003 to the definition of the IEEE 802.15.4 standard for low rate, low complexity, and low power wireless networks [1]. The 802.15.4 standard also forms the basis of the ZigBee technology, which provides a comprehensive solution for low data rate networking from physical layer to applications [2].

Both IEEE 802.15.4 and ZigBee lack, however, an important requirement of forthcoming low data rate systems, namely, the capability of locating, with sufficiently high precision, objects and individuals by means of distributed, infrastructure-independent positioning algorithms.

The introduction of positioning in low data rate networks is a top priority of the recently formed IEEE 802.15.4a Task Group [3], which recently proposed Impulse Radio Ultra Wide Band (IR-UWB) as an attractive transmission technique for indoor and outdoor low data rate wireless networks [4], [5]. Thanks to its ultra wide bandwidth that spans over several GHz, IR-UWB has some interesting properties, in particular:

- an inherently high temporal resolution that provides good robustness in the presence of multipath, thereby allowing communication despite obstacles and Non-Line-Of-Sight (NLOS) propagation conditions.
- the capability of providing accurate ranging, thanks to its high temporal resolution. Distance information can then be used for deriving the physical position of terminals in the network.

The definition of the Uncoordinated Baseborn Wireless medium access control for UWB networks $(UWB)^2$ protocol was based on the above specific features of IR-UWB [6]. $(UWB)^2$ also evaluates and stores distances, which are then used by positioning and routing algorithms. The $(UWB)^2$ approach for propagation over AWGN channels was validated in [7].

In the present work, we revisit $(UWB)^2$ by redefining the structure of both control and DATA MAC Protocol Data Units (MACPDUs) based on the PDU structure of the original IEEE 802.15.4 MAC standard in order to guarantee compatibility of the new MAC protocol with both distributed and centralized network topologies defined in the 802.15.4 standard.

Next, we extend the analysis of the $(UWB)^2$ protocol to the case of multipath-affected channels, for both indoor and outdoor channel scenarios in Line Of Sight (LOS) and Non-Line Of Sight (NLOS) conditions. Channel parameters were obtained from the channel model proposed within the 802.15.4a Task Group, and a set of channel realizations were considered for each selected scenario.

Finally, Multi User Interference (MUI) was also included in the performance analysis based on an enhanced version of the Pulse Collision model specific for IR-UWB [7], which takes into account multipath. This MUI model is used to analyze performance of $(UWB)^2$ by simulation, as a function of channel, network size, and user bit rates.

The paper is organized as follows: Section 2 summarizes $(UWB)^2$ and the ranging scheme, and Section 3 defines the format of the $(UWB)^2$ MAC Protocol Data Unit, based on the structure defined within the preexisting 802.15.4 MAC. The Pulse Collision MUI model is introduced in Section 4. Performance evaluation of $(UWB)^2$ in presence of multipath and MUI for different number of users and offered traffic is carried out in Section 5. Section 6 draws conclusions.

2. THE $(UWB)^2$ MAC PROTOCOL

The high temporal resolution of IR-UWB signals has the beneficial side effect of reinforcing robustness to MUI, in particular for low data rate applications [4]. As a consequence, access to the medium in low data rate UWB networks can be based on a most straightforward solution, Aloha ([8], [6]), by which devices transmit in an uncoordinated fashion. Thanks to the resilience to MUI offered by impulse radio, correct reception

for multiple simultaneous links can be obtained. An Aloha-like approach may also favor lowering costs, since it does not rely on specific physical layer (PHY) functions, such as carrier sensing, and may thus be adapted with little effort to different PHYs.

As for the duty cycle of emitted signals, low data rate scenarios usually lead to an average Pulse Repetition Period (PRP), the average time between two consecutive pulses emitted by a device, on the order of $10^{-4} \div 10^{-5}$ s, with an average duration of emitted pulses typically on the order of 10^{-10} s. Theoretically, the duty cycle can thus be as low as 10^{-6}. However, a detailed analysis of this issue requires introducing the channel model in order to take into account propagation effects on pulse duration. When Time Hopping (TH) is the selected coding technique, TH – Code Division Multiple Access (TH-CDMA) is a natural choice for multiple access. The adoption of TH-CDMA can introduce an additional degree of freedom, since the effect of pulse collisions is further reduced by the adoption of different codes on different links. Two factors cooperate in determining the robustness to MUI: low duty cycle of emitted signals, and association of different TH-Codes with different links.

$(UWB)^2$ is a multi-channel MAC protocol that is based on the combination of Aloha with TH-CDMA [6]. $(UWB)^2$ adopts the combination of a common code for signaling, where terminals share the same code, and code collisions are avoided thanks to phase shifts between different links, and Transmitter codes for data transfers, where each terminal has a unique code for transmitting, and the receiver switches to the code of the transmitter for receiving a packet.

The packet exchange between transmitter TX and receiver RX that takes place during connection set-up may also serve for enabling a simple ranging procedure, based on a three-way exchange. During set-up, TX and RX prepare a DATA packet transmission by exchanging a Link Establishment (LE) packet transmitted on the Common Code, followed by a Link Confirm (LC) packet transmitted on the Transmitter Code of RX, and finally by the DATA packet on the Transmitter Code of TX. This handshake allows storing distance between TX and RX at both TX and RX. The protocol also foresees the presence of a procedure by which each terminal *i* maintains a ranging database for all neighboring terminals. Each entry of the database contains the ID *j* of the neighbor, the estimated distance to *j*, and a timestamp indicating the time at which the estimation was performed.

3. THE $(UWB)^2$ MACPDU FORMAT

The format of the MACPDU originally proposed in [6] was revisited and modified in order to take into account the characteristics of the future IEEE 802.15.4a PHY.

The MACPDU is composed of a header, a payload, and a trailer. The standard header, shared by all PDUs and long up to 23 bytes, is derived from the 802.15.4 header and is organized as follows:

- frame control (2 bytes)
- equence number (1 byte)
- destination PAN identifier (2 bytes)
- destination address (2/8 bytes)
- source PAN identifier (2 bytes)
- source address (2/8 bytes)

In the case of LE control packets (link set-up phase of $(UWB)^2$), the header includes the following additional fields:

- Time Hopping flag (1 bit), used to inform destination whether the standard Time Hopping code or a different one is going to be adopted in the DATA transmission;
- Time Hopping code (0/2 bytes), used for communicating the TH code to the destination (e.g., by including the code identifier, assuming that all nodes share a common codebook).

In the case of DATA PDU, the header contains the 1 byte additional field N_{PDU} that indicates to the destination the number of additional DATA PDUs that will be sent from the source. If N_{PDU} is different from 0, the destination will keep on

listening on the DATA TH code and wait for additional DATA PDUs. The length of the payload is set to 0 for LE and LC PDUs, while the ACK PDU has a 2 byte payload containing the status of the corresponding DATA PDU. Finally, DATA PDUs have a payload length of up to 103 bytes.

All PDUs include a 2 byte trailer consisting of a CRC code evaluated on the entire PDU.

The above PDU structure leads to a maximum PDU length of 129 bytes, corresponding to the case of a DATA PDU with full header (24 bytes), full payload (103 bytes), and the 2 byte trailer.

4. BER EVALUATION UNDER THE PULSE COLLISION MODEL

4.1. System model

We assume IR-UWB transmissions with Pulse Position Modulation (PPM) and TH coding. Signals generated at TX are described as follows:

$$s_{TX}(t) = \sqrt{E_{TX}} \sum_j p_0\left(t - jT_S - \theta_j - \varepsilon b_{\lfloor j/N_S \rfloor}\right), \quad (1)$$

where $p_0(t)$ is the energy-normalized waveform of the transmitted pulses, E_{TX} is the transmitted energy per pulse, T_S is the average pulse repetition period, $0 \le \theta_j < T_S$ is the TH time shift of the j-th pulse, ε is the PPM shift, b_x is the x-th bit of a binary source sequence $\mathbf{b}$, N_S is the number of pulses transmitted for each bit, and $\lfloor x \rfloor$ is the inferior integer part of x.

Propagation for link m occurs over a multipath-affected channel with impulse response given by:

$$h^{(m)}(t) = X^{(m)} \sum_{l=0}^{L^{(m)}} \sum_{k=0}^{K} \alpha_{k,l}^{(m)} \delta\left(t - \Delta t^{(m)} - T_l^{(m)} - \tau_{k,l}^{(m)}\right), \quad (2)$$

where $X^{(m)}$ is the amplitude gain, $L^{(m)}$ is the number of clusters, K is the number of paths that are considered within each cluster, $\delta(t)$ is the Dirac function, $\Delta t^{(m)}$ is the propagation delay, $T_l^{(m)}$ is the delay of the l-th cluster with respect to $\Delta t^{(m)}$, $\tau_{k,l}^{(m)}$ is the delay of the k-th path relative to the l-th cluster arrival time, and $\alpha_{k,l}^{(m)}$ is the real-valued tap weight of the k-th path within the l-th cluster. Tap weights are energy-normalized and thus verify:

$$\sum_{l=0}^{L^{(m)}} \sum_{k=0}^{K} \left(\alpha_{k,l}^{(m)}\right)^2 = 1, \quad (3)$$

For all channel parameters in (2), we adopt the statistical characterization that is suggested in [10] for 9 different propagation environments, i.e., *i*) residential LOS, *ii*) residential NLOS, *iii*) office LOS, *iv*) office NLOS, *v*) outdoor LOS, *vi*) outdoor NLOS, *vii*) industrial LOS, *viii*) industrial NLOS, and *ix*) open outdoor environment NLOS (farm, snow-covered open area).

For link m, both channel gain $X^{(m)}$ and propagation delay $\Delta t^{(m)}$ depend on distance of propagation $D^{(m)}$ between TX and RX. For $X^{(m)}$, in particular, one has:

$$X^{(m)} = 1 \Big/ \sqrt{10^{\left(PL^{(m)}/10\right)}}, \quad (4)$$

where $PL^{(m)}$ is the path loss in dB, which can be modelled as indicated in [10].

Reference TX and RX are assumed to be perfectly synchronized. The channel output is corrupted by thermal noise and MUI generated by N_i interfering and asynchronous IR-UWB devices. The received signal at RX input writes:

$$s_{RX}(t) = r_u(t) + r_{mui}(t) + n(t), \quad (5)$$

where $r_u(t)$, $r_{mui}(t)$, and $n(t)$ are the useful signal, MUI, and thermal Gaussian noise with double-sided power spectral density $N_0/2$, respectively. By denoting as 0 the reference link between TX and RX, the useful signal $r_u(t)$ writes as follows:

$$r_u(t) = \sqrt{E_0} \sum_j \sum_{l=0}^{L^{(0)}} \sum_{k=0}^{K} \alpha_{k,l}^{(0)} \cdot \cdot p_0\left(t - jT_S - \theta_j^{(0)} - \varepsilon b_{\lfloor j/N_S \rfloor} - \Delta t^{(0)} - T_l^{(0)} - \tau_{k,l}^{(0)}\right), \quad (6)$$

where $E_0 = (X^{(0)})^2 E_{TX}$ is the total received energy per pulse.

As for $r_{mui}(t)$, we assume all interfering signals to be characterized by the same T_S; thus:

$$r_{mui}(t) = \sum_{n=1}^{N_i} \sqrt{E_n} \sum_j \sum_{l=0}^{L^{(n)}} \sum_{k=0}^{K} \alpha_{k,l}^{(n)} \cdot \cdot p_0\left(t - jT_S - \theta_j^{(n)} - \varepsilon b_{\lfloor j/N_S^{(n)} \rfloor}^{(n)} - \Delta t^{(n)} - T_l^{(n)} - \tau_{k,l}^{(n)}\right), \quad (7)$$

where index n represents the wireless link between the nth interfering device and RX. In (7), $E_n = (X^{(n)})^2 E_{TX}$, and $\Delta\tau^{(n)}$ are the received energy per pulse and the delay for link n. The terms $\theta_j^{(n)}$, $b_x^{(n)}$ and $N_S^{(n)}$ in (7) are the time shift of the j-th pulse for user n, the x-th bit, and the number of pulses per bit, respectively for user n.

Both TH codes and data bit sequences are assumed to be randomly generated and correspond to pseudonoise sequences, that is, $\theta_j^{(n)}$ terms are assumed to be independent random variables uniformly distributed in the range $[0,T_S)$, and $b_x^{(n)}$ values are assumed to be independent random variables with equal probability to be "0" or "1." Based on the above assumptions, the N_i relative delays $\Delta\tau^{(0)}-\Delta\tau^{(n)}$, with $n = 1,...,N_i$ may be reasonably modelled as independent random variables uniformly distributed between 0 and T_S.

As well known, the optimum receiver structure for (6) consists of a RAKE receiver composed of a parallel bank of correlators, followed by a combiner that determines the variable to be used for the decision on the transmitted symbol. Each correlator of the RAKE is locked on one of the different replicas of the transmitted waveform $p_0(t)$. The complexity of such a receiver increases with the number of multipath components that are analyzed and combined before decision, and can be reduced by processing only a sub-set of the components that are available at the receiver input [4].

Such a reduction, however, entails a decrease in the available useful energy in the decision process, together with a consequent decrease in receiver performance. As a result, system designers have the possibility to trade the cost of the devices with the performance of the physical layer. For some application scenarios, for example, it might be better to have very cheap devices with modest performance with respect to high-priced terminals with better performance. In the examined scenario, we adopt a basic IR receiver that analyzes a single component of the received signal. This basic receiver is composed by a coherent correlator followed by an ML detector [4]. In every bit period, the correlator converts the received signal in (5) into a decision variable Z that forms the input of the detector. Soft decision detection is performed. For each pulse, we assume that the correlator locks onto the multipath component with maximum energy. By indicating with l_M and k_M the cluster and the path of the maximum energy multipath component for the reference user, the input of the detector Z for a generic bit b_x is as follows:

$$Z = \int_{xN_ST_S+\Delta T^{(0)}}^{(x+1)N_ST_S+\Delta T^{(0)}} s_{RX}(t)\, m_x\left(t - \Delta T^{(0)}\right) dt \quad (8)$$

where

$$\Delta T^{(0)} = \Delta t^{(0)} + T_{l_M}^{(0)} + \tau_{k_M,l_M}^{(0)} \quad (9)$$

and where

$$m_x(t) = \sum_{j=xN_S}^{(x+1)N_S} \left(p_0\left(t - jT_S - \theta_j\right) - p_0\left(t - jT_S - \theta_j - \varepsilon\right)\right) \quad (10)$$

By introducing (5) into (8), we obtain: $Z=Z_u+Z_{mui}+Z_n$, where Z_u is the signal term, Z_{mui} is the MUI contribution, and Z_n is the noise contribution, which is Gaussian with zero mean and variance $\sigma_n^2 = N_S N_0 \xi(\varepsilon)$, where $\xi(\varepsilon)=1-R_0(\varepsilon)$, and where $R_0(\varepsilon)$ is the autocorrelation function of the pulse waveform $p_0(t)$ [4]. Bit b_x is estimated by comparing the Z term in (8) with a zero-valued threshold according to the following rule: when Z is positive decision is "0", when Z is negative decision is "1."

4.2. BER estimation under the Pulse Collision approach

According to Section 4.1, the average probability of error on the bit at the output of the detector for independent and equiprobable

transmitted bits is: BER = $Prob\{Z<0|b_X=0\} = Prob\{Z_{mui}<-y\}$, where $y = Z_u+Z_n$ is a Gaussian random variable with mean:

$$\mu_y = N_S\xi(\varepsilon)\sqrt{\left(\alpha^{(0)}_{l_M,k_M}\right)^2 E_0} = N_S\xi(\varepsilon)\sqrt{E_u} \qquad (11)$$

and variance $\sigma_y^2 = N_S N_0 \xi(\varepsilon)$. The quantity E_u in (11) indicates the amount of useful energy conveyed by the maximum multipath contribution. The average BER at the receiver output can be evaluated by applying the Pulse Collision (PC) approach [11]. First, we compute the conditional BER for a generic y value, i.e., $Prob\{Z_{mui} < -y \mid y\}$, and we then average over all possible y values, that is:

$$\text{BER} = \int_{-\infty}^{+\infty} Prob\{Z_{mui} < -y \mid y\}\, p_Y(y)\, dy \qquad (12)$$

Next, we expand the conditional BER in order to take into account collisions between pulses of different transmissions. In a bit period, the number of possible collisions at the input of the reference receiver, denoted with c, is confined between 0 and $N_S N_i$, with N_S pulses per bit and N_i interfering users. Thus:

$$\text{BER} = \sum_{c=0}^{N_S N_i} P_C(c) \int_{-\infty}^{+\infty} Prob\{ Z_{mui} < -y \mid y, c \}\, p_Y(y)\, dy \qquad (13)$$

where $P_C(c)$ is the probability of having c collisions at the receiver input. For independent interferers, $P_C(c)$ can be expressed through the binomial distribution:

$$P_C(c) = \binom{N_S N_i}{c} (P_0)^c (1-P_0)^{N_S N_i - c}, \qquad (14)$$

where P_0 is the basic collision probability, which is defined as the probability that an interfering device produces a non-zero contribution within a single T_S. Given the receiver structure in (8), we approximate P_0 as follows:

$$P_0 = \frac{T_m + \varepsilon + \tau_{MAX}}{T_S}. \qquad (15)$$

where T_m is the time duration of the pulse waveform $p_0(t)$, and τ_{MAX} is the maximum among the values of the root mean square delay spread for the N_i channels between the interfering devices and RX. Note that (15) provides acceptable P_0 values if $T_S > T_m + \varepsilon + \tau_{MAX}$, which is reasonable for LDR systems with long pulse repetition periods. This condition guarantees that no Inter Frame Interference (ISI) is present at the receiver, even in the presence of multipath propagation.

As regards $Prob(Z_{mui} < -y \mid y,c)$, we adopt the linear model introduced in [11], that is:

$$Prob\left(Z_{mui} < -y \mid y, c\right) = \begin{cases} 1 & \text{for} \quad y \le -\zeta(n) \\ 1 - \dfrac{P_C(c)}{2}\left(1 + \dfrac{y}{\zeta(c)}\right) & \text{for} \quad \zeta(n) < y \le 0 \\ \dfrac{P_C(c)}{2}\left(1 - \dfrac{y}{\zeta(c)}\right) & \text{for} \quad 0 < y \le \zeta(n) \\ 0 & \text{for} \quad y > \zeta(n), \end{cases} \qquad (16)$$

where $\zeta(c)$ indicates the maximum interference contribution that can be measured at the output of the correlator. Based on [11], we propose here the following approximation for $\zeta(c)$:

$$\zeta(c) = \sum_{j=1}^{N_i} \left(\left\lceil \frac{c-j+1}{N_i} \right\rceil \sqrt{E_{\text{int}}^{(j)} \frac{T_m + \varepsilon}{\tau_{rms}^{(j)}}} \right) \qquad (17)$$

where $\{E_{int}^{(1)}, E_{int}^{(2)}, \ldots, E_{int}^{(Ni)}\}$ are the interfering energies $\{E_1, E_2, \ldots, E_{Ni}\}$ of (7), sorted in descending order so that $E_{int}^{(j)} \ge E_{int}^{(j+1)}$ for $j = 1,\ldots,N_i\text{-}1$. The expression in (17) indicates that the value of the maximum interference contribution at the receiver output is computed privileging dominating interferers, that is, those users with the highest interfering energies.

Note that in (17) we multiply the value of jth interfering energy $E_{int}^{(j)}$ by $(T_m+\varepsilon)/\tau_{rms}^{(j)}$.

This operation indicates that only part of the energy associated with a colliding pulse contributes to Z in (8), corresponding to the ratio between the correlator window $T_m+\varepsilon$ and the length of the pulse at the receiver, approximated by $\tau_{rms}^{(j)}$. By combining

(16) into (13), one has:

$$\mathrm{BER} \approx \frac{1}{2} erfc\left(\sqrt{\frac{1}{2}\frac{N_S E_u}{\mathrm{N}_0}\xi(\varepsilon)}\right) + \sum_{c=0}^{N_t N_S} \frac{P_C(c)^2}{2}\Omega\left(\frac{N_S E_u}{\mathrm{N}_0}\xi(\varepsilon), \frac{\varsigma(c)^2}{N_S \mathrm{N}_0 \xi(\varepsilon)}\right), \quad (18)$$

where

$$\Omega(A,B) = \frac{1}{2} erfc\left(\sqrt{\frac{A}{2}} - \sqrt{\frac{B}{2}}\right) + \frac{1}{2} erfc\left(\sqrt{\frac{A}{2}} + \sqrt{\frac{B}{2}}\right) - erfc\left(\sqrt{\frac{A}{2}}\right), \quad (19)$$

The first term in (18) only depends on the signal to thermal noise ratio at the receiver input, while the second one accounts for MUI. The proposed approach was demonstrated to guarantee high accuracy in estimating receiver performance for impulse-based transmissions, even in the presence of scarcely populated systems, systems with dominating interferers, or low-rate systems [11].

5. PERFORMANCE ANALYSIS

The (UWB)^2 protocol described in Section 2 was tested by simulation. Simulation results were averaged over L different simulation runs. In each simulation run, N nodes were randomly located inside a square region with area A. Next, a realization of the channel impulse response, path loss, and delay spread was generated for each pair of nodes, with characteristics depending on the considered propagation scenario. These quantities were used by the interference module for introducing errors on the received packets, according to the MUI model described in Section 4.2.

We considered the scenarios CM1, CM2, CM5, and CM6 defined within IEEE 802.15.4a, corresponding to indoor propagation in residential environments in LOS and NLOS conditions, and outdoor propagation in LOS and NLOS [10]. CM1, CM2, CM5, and CM6 channels will be indicated as Scenarios 1, 2, 3, and 4, respectively. The settings for the path loss at a reference distance and path loss exponent in the four channel scenarios are presented in Table I.

Scenario	Path Loss @ d=1m	Path Loss exponent
1	43.9 dB	1.79
2	48.7 dB	4.58
3	43.29 dB	1.76
4	43.29 dB	2.5

TABLE 1: CHANNEL SCENARIOS CHARACTERISTICS.

During all simulations, the maximum size of 1288 bits was adopted for the PHYPDU. This value was obtained by considering as PHY payload a full size MACPDU of 129 bytes coded with a Reed Solomon (43,51) Forward Error Correction code in compliance with the specifications for the UWB PHY of the future 802.15.4a standard [12]. The 1224 coded bits were then combined with a PHY synchronization trailer of length 64 bits, leading to a size of 1288 bits for each PHYPDU.
Table II presents the main simulation settings.

Parameter	Setting
L	10
Number of nodes	From 10 to 20
Area	10 m × 10 m (Indoor) 50 m × 50 m (Outdoor)
Network topology	Random node positions
Channel model	See eq. (2) and [10]
User bit rate R	From 10 kb/s to 30 kb/s
Transmission rate	966 kb/s
Power	36.5 μW (FCC limit for Bandwidth ≅ 0.5 GHz)
Packet traffic model	Poisson generation process, uniform distribution for destination node
DATA packet length	1224 bits (+ 64 bits for Sync trailer)
Interference Model	Pulse Collision (see section 4)
Physical layer settings	N_S = 4, T_s = 258.8 ns T_m = 2 ns, Reed Solomon (43,51) FEC

TABLE 2: SIMULATION SETTING.

Performance of (UWB)^2 was analyzed as a function of:

- channel characteristics (indoor vs outdoor)
- number of terminals
- user bit rate
- access strategy (pure vs. slotted).

The comparison between pure and slotted Aloha was motivated by the fact that, as well known, in narrowband networks, slotted Aloha guarantees a higher (up to two times) throughput with respect to pure Aloha, thanks to a lower probability of packet collision. Our goal was to verify if this large performance gap is also present in low bit rate UWB networks, where the negative impact of packet collisions is mitigated by the high processing gain.

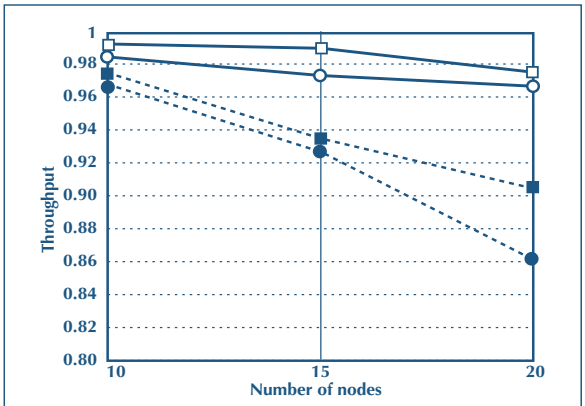

Figure 1: Throughput as a function of number of nodes for a user data rate R = 10 kb/s and Pure Aloha access (Filled line with white circles: Scenario 1 (Indoor LOS); Dashed line with full circles: Scenario 2 (Indoor NLOS); Filled line with white squares: Scenario 3 (Outdoor LOS); Dashed line with full squares: Scenario 4 (Outdoor NLOS).)

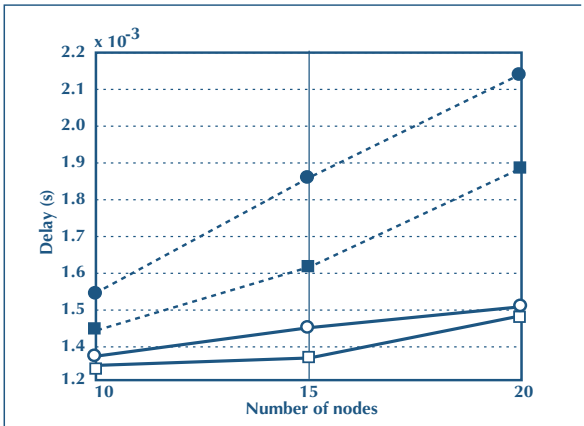

Figure 2: Delay as a function of number of nodes for a user data rate R = 10 kb/s and Pure Aloha access (Filled line with white circles: Scenario 1 (Indoor LOS); Dashed line with full circles: Scenario 2 (Indoor NLOS); Filled line with white squares: Scenario 3 (Outdoor LOS); Dashed line with full squares: Scenario 4 (Outdoor NLOS).)

Fig. 1 presents the throughput for the Pure Aloha strategy as a function of the number of nodes, for a user bit rate R = 10 kb/s.

Fig. 1 shows that in all cases throughput is greater than around 98% for both indoor and outdoor LOS scenarios and stays above 85%, even in NLOS conditions, where the higher path loss and the larger channel delay spread have a stronger impact on network performance.

Fig. 2 shows the delay measured in the same simulations, taking into account both the DATA PDU transmission time, equal to 1.33 ms, and the additional delay introduced by retransmissions following PDU collisions. Fig. 2 shows that for LOS scenarios the delay experienced by DATA PDUs is close to a minimum possible value, given by the DATA PDU transmission time. NLOS scenarios lead to a larger delay that is, however, below 2.2 ms in all cases.

Figs.3 and 4 show throughput and delay in the case of Slotted Aloha.

Fig.3 shows that Slotted Aloha leads to throughputs comparable to Pure Aloha, with values above 96% in LOS conditions and above 85% in the NLOS case.

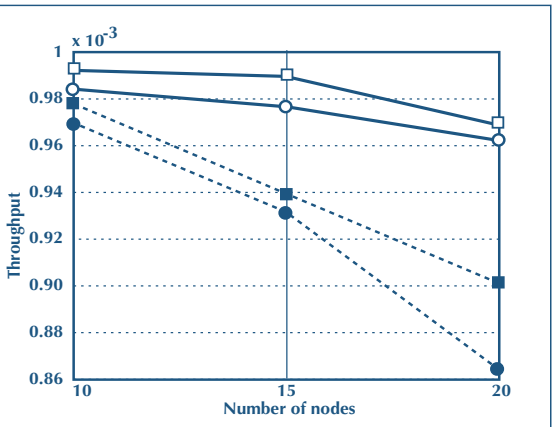

Figure 3: Throughput as a function of number of nodes for a user data rate R = 10 kb/s and Slotted Aloha access (Filled line with white circles: Scenario 1 (Indoor LOS); Dashed line with full circles: Scenario 2 (Indoor NLOS); Filled line with white squares: Scenario 3 (Outdoor LOS); Dashed line with full squares: Scenario 4 (Outdoor NLOS).)

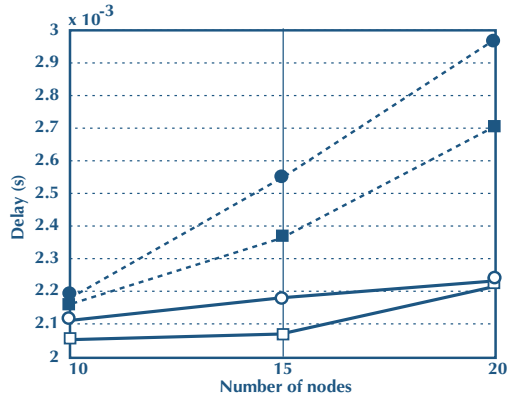

FIGURE 4: DELAY AS A FUNCTION OF NUMBER OF NODES FOR A USER DATA RATE R = 10 KB/S AND SLOTTED ALOHA ACCESS (FILLED LINE WITH WHITE CIRCLES: SCENARIO 1 (INDOOR LOS); DASHED LINE WITH FULL CIRCLES: SCENARIO 2 (INDOOR NLOS); FILLED LINE WITH WHITE SQUARES: SCENARIO 3 (OUTDOOR LOS); DASHED LINE WITH FULL SQUARES: SCENARIO 4 (OUTDOOR NLOS).)

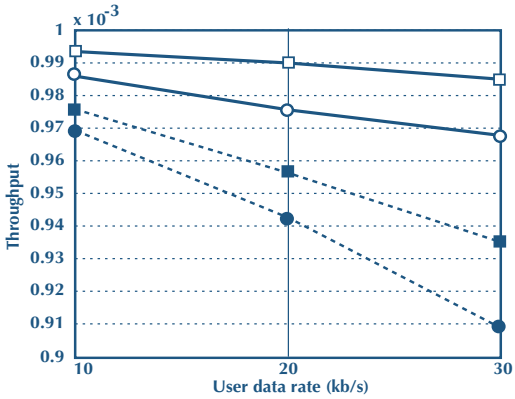

FIGURE 5: THROUGHPUT AS A FUNCTION OF USER DATA RATE FOR A NETWORK OF 10 NODES AND PURE ALOHA ACCESS (FILLED LINE WITH WHITE CIRCLES: SCENARIO 1 (INDOOR LOS); DASHED LINE WITH FULL CIRCLES: SCENARIO 2 (INDOOR NLOS); FILLED LINE WITH WHITE SQUARES: SCENARIO 3 (OUTDOOR LOS); DASHED LINE WITH FULL SQUARES: SCENARIO 4 (OUTDOOR NLOS).)

On the other hand, the additional delay in PDU transmission introduced by the slotted time axis leads to higher delays, in agreement with [7], as shown in Fig. 4.

Note, however, that the increase in the delay as a function of the number of users is slower than that in the delay for Pure Aloha, indicating that, for higher numbers of users, the Slotted Aloha approach should eventually guarantee lower delays and thus better performance.

In a second set of simulations, we evaluated the impact of the user data rate R on performance. We considered a network of 10 nodes and measured throughput and delay for three different data rates: 10, 20, and 30 kb/s, respectively.

The throughput obtained in the case of the Pure Aloha approach is presented in Fig. 5.

Fig. 5 shows that performance of the $(UWB)^2$ degrades gracefully as the offered traffic increases: the throughput is above 90% in all considered cases, and is well above 95% for the LOS scenarios. This behavior is confirmed by data on delay, as shown in Fig. 6. LOS scenarios have very low delays, and delays for NLOS scenarios are below 1.75 ms in all cases.

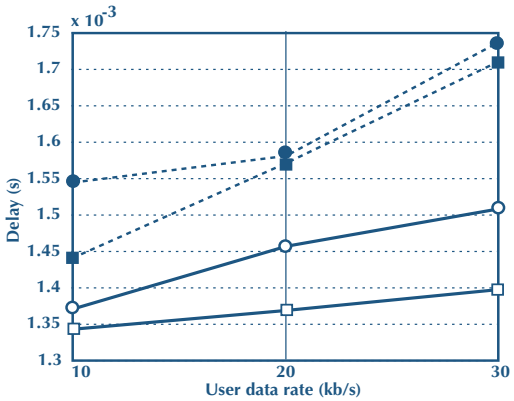

FIGURE 6: DELAY AS A FUNCTION OF USER DATA RATE FOR A NETWORK OF 10 NODES AND PURE ALOHA ACCESS (FILLED LINE WITH WHITE CIRCLES: SCENARIO 1 (INDOOR LOS); DASHED LINE WITH FULL CIRCLES: SCENARIO 2 (INDOOR NLOS); FILLED LINE WITH WHITE SQUARES: SCENARIO 3 (OUTDOOR LOS); DASHED LINE WITH FULL SQUARES: SCENARIO 4 (OUTDOOR NLOS).)

We measured throughput and delay for Slotted Aloha, presented in Figs 7 and 8, respectively.

Fig. 7 shows that Slotted Aloha provides slightly better results compared to Pure Aloha for high offered traffic loads, thanks to a lower probability of packet collision. The difference is more pronounced for NLOS scenarios, where Slotted Aloha

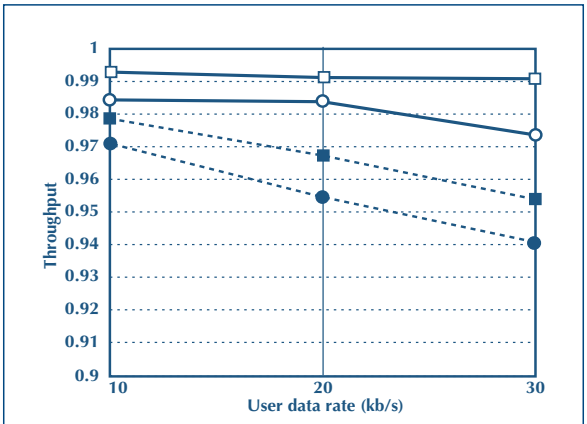

FIGURE 7: THROUGHPUT AS A FUNCTION OF USER DATA RATE FOR A NETWORK OF 10 NODES AND SLOTTED ALOHA ACCESS (FILLED LINE WITH WHITE CIRCLES: SCENARIO 1 (INDOOR LOS); DASHED LINE WITH FULL CIRCLES: SCENARIO 2 (INDOOR NLOS); FILLED LINE WITH WHITE SQUARES: SCENARIO 3 (OUTDOOR LOS); DASHED LINE WITH FULL SQUARES: SCENARIO 4 (OUTDOOR NLOS).)

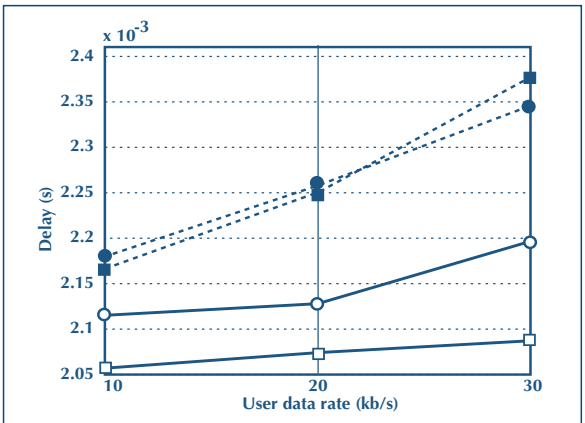

FIGURE 8: DELAY AS A FUNCTION OF USER DATA RATE FOR A NETWORK OF 10 NODES AND SLOTTED ALOHA ACCESS (FILLED LINE WITH WHITE CIRCLES: SCENARIO 1 (INDOOR LOS); DASHED LINE WITH FULL CIRCLES: SCENARIO 2 (INDOOR NLOS); FILLED LINE WITH WHITE SQUARES: SCENARIO 3 (OUTDOOR LOS); DASHED LINE WITH FULL SQUARES: SCENARIO 4 (OUTDOOR NLOS).)

guarantees in all cases a throughput higher than 94%, vs. 91% obtained by the Pure Aloha in the worst case.

The results are confirmed by Fig. 8, showing that the increase in delay as the offered traffic increases is proportionally lower than in the case of Pure Aloha, as a consequence of the higher robustness of the Slotted Aloha approach in high traffic scenarios.

6. CONCLUSIONS

In this work, the (UWB)^2 MAC protocol, originally introduced in [6], was revisited in view of its application to the future IEEE 802.15.4a standard. The structure of both control and DATA MACPDUs of the (UWB)^2 protocol was derived from the PDU structure of the existing 802.15.4 MAC, thus guaranteeing full support for the network topologies defined within the original standard. The (UWB)^2 protocol adopts Aloha for medium access and CDMA for multiple access, based on the use of Time Hopping codes. The protocol can operate in either a slot-free (pure) or a slotted fashion, and can thus fit both centralized and distributed network architectures. The protocol also includes a ranging procedure in order to enable the operation of location-based protocols at higher layers.

The performance of the (UWB)^2 MAC was evaluated in the presence of multipath-affected propagation channels, derived from the channel model proposed within the IEEE 802.15.4a Task Group. Performance in both pure and slotted modes of operation was analyzed by simulation in indoor and outdoor scenarios in both Line-Of-Sight and Non-Line-Of-Sight conditions. In order to properly take into account the impact of multipath and channel delay spread on network performance, we also introduced an ad-hoc MUI model based on the concept of Pulse Collision in this work.

Simulation results show that the (UWB)^2 MAC guarantees satisfactory network performance in both indoor and outdoor scenarios, even in presence of NLOS propagation conditions. Furthermore, network performance in case of high traffic loads can be improved by adopting a Slotted Aloha approach. Results suggest that, despite its simplicity, the (UWB)^2 MAC provides high throughput and low delays for bit rates up to several tens of kb/s and for networks composed of tens of terminals, thereby making it a viable solution for future UWB low data rate networks.

ACKNOWLEDGMENTS

This work was partially supported by the European Union within the Integrated Projects n. 506897 - PULSERS and n. 511766 LIAISON and by STMicroelectronics Srl (Italy) within the research contract "UWB Ranging and Positioning in Radio Communication Systems."

REFERENCES

[1] The IEEE 802.15.4 standard, available at http://www.ieee.org.

[2] P. Kinney, "ZIGBEE TECHNOLOGY: WIRELESS CONTROL THAT SIMPLY WORKS," (2003), Available at http://www.zigbee.org/en/resources/#WhitePapers.

[3] IEEE 802.15.TG4a page, http://www.ieee802.org/15/pub/TG4a.html.

[4] M.-G. Di Benedetto and Giancola G., **Understanding Ultra Wide Band Radio fundamentals**, Prentice Hall, 2004.

[5] L. De Nardis and G. M. Maggio, "LOW DATA RATE UWB NETWORKS," in Ultra Wideband Wireless Communications, John Wiley & Sons, Inc. (May 2006).

[6] M.-G. Di Benedetto, L. De Nardis, M. Junk and G. Giancola, "$(UWB)^2$: UNCOORDINATED, WIRELESS, BASEBORN MEDIUM ACCESS FOR UWB COMMUNICATION NETWORKS," MONET: Special Issue on WLAN optimization at the MAC and network levels, vol. 5, no. 10, pp. 663-674, October 2005.

[7] L. De Nardis, G. Giancola and M.-G. Di Benedetto, "PERFORMANCE ANALYSIS OF UNCOORDINATED MEDIUM ACCESS CONTROL IN LOW DATA RATE UWB NETWORKS," in Proceedings of the 1st IEEE/CreateNet International Workshop on "ULTRAWIDEBAND WIRELESS NETWORKING," within the 2nd International Conference on Broadband Networks, (invited paper), pp. 206-212, October 2005, Boston, Massachusetts, USA.

[8] L. De Nardis and M.-G. Di Benedetto, "MEDIUM ACCESS CONTROL DESIGN FOR UWB COMMUNICATION SYSTEMS: REVIEW AND TRENDS," KICS Journal of Communications and Networks, vol. 5, no. 4, pp. 386–393, December 2003.

[9] E. S. Sousa and J. A. Silvester, "SPREADING CODE PROTOCOLS FOR DISTRIBUTED SPREAD-SPECTRUM PACKET RADIO NETWORKS," IEEE Transactions on Communications, vol. COM–36, no. 3, pp. 272–281, March 1988.

[10] IEEE 802.15.4a Channel Model Final Report, Rev.1 (November 2004), available at: ftp://ieee:wireless@ftp.802wirelessworld.com/15/04/15-04-0662-00-004a-channel-model-final-report-r1.pdf.

[11] G. Giancola and M.-G. Di Benedetto, "A NOVEL APPROACH FOR ESTIMATING MULTI USER INTERFERENCE IN IMPULSE RADIO UWB NETWORKS: THE PULSE COLLISION MODEL," (invited paper) EURASIP Signal Processing Journal, Special Issue on Signal Processing in UWB Communications, vol. 86, no. 5, pp. 2172-2184, May 2006.

[12] G. M. Maggio, "802.15.4A UWB-PHY," IEEE 802.15.4a Document # IEEE 15-05-0707-01-004a, November 2005.

■ CONTACT: ST.JOURNAL@ST.COM ■

AN ULTRA-LOW ENERGY ASYNCHRONOUS PROCESSOR FOR WIRELESS SENSOR NETWORKS

L. Necchi (1), L. Lavagno (1)
D. Pandini (2), L. Vanzago (2)

(1) Politecnico di Torino
(2) STMicroelectronics

This paper describes the design flow used for an asynchronous 8-bit processor implementing the Atmel AVR instruction set architecture. The goal is to show dramatic reductions in power and energy with respect to the synchronous case by exploiting aggressive dynamic voltage scaling, while achieving almost the same area, delay, and power at the same voltage as normal synchronous operation, and while retaining a traditional design flow. The processor was implemented in a $130nm$ technology using desynchronization, starting from an initial design downloaded from OpenCores.org. It consumes $14pJ$ per instruction to deliver 170 MIPS at $1.2\ V$ and $2.7\ pJ$ per instruction to deliver 48 MIPS at $0.54\ V$. It thus dramatically improves the energy consumed per instruction with respect to previous results from the literature.

1. INTRODUCTION

Wireless sensor networks (WSNs) promise to offer a variety of low-cost measurement and actuation services for a very broad range of applications, from home and factory automation, to health care, to environment and building monitoring. Most WSN architectures rely on a multitude of small devices spread over the covered area in order to provide services in an unobtrusive, reliable, and economical manner. Hence, node cost, due to fabrication, deployment, and maintenance, is a key aspect of WSN proliferation. Reducing both power and energy requirements per computation activity (e.g., sensing operation, packet reception or transmission, routing action) permits the manufacturer to reduce the size of the battery and increase its lifetime, thus impacting both unit and deployment costs.

While a majority of current demo applications for WSNs spend most of the node power operating the radio, it is clear that requiring more and more services from WSNs in the future will mean increased power consumption for the digital portion of the node as well. Examples of these power-hungry services include network security [16], in-network aggregation [22], and distributed data processing [20]. While adoption of these services in real WSN deployments is still very preliminary, they are generally recognized as

very significant in the future, as they will improve WSN robustness and global network lifetime.

This paper describes an implementation of a typical WSN processor, the 8-bit Atmel AVR, using asynchronous techniques. The resulting processor is called YUPPIE, which stands for Yet another Ultra-low Power asynchronous Processor for wIreless sensor nEtworks. The *key contribution* of this exercise is to show that it is indeed possible to *simultaneously achieve*:

- design re-use, since the synthesizable VHDL model of the original microcontroller was downloaded from the OpenCores.org web site.
- implementation with a standard design flow and library, based on a 130 *nm* technology from STMicroelectronics.
- extremely short design time, since the complete design and simulation was completed in about four months of work by a person with no previous exposure to asynchronous design techniques and limited exposure to synchronous ones. Most of the time was actually spent learning to use the tools from the industrial partner's design flow.
- aggressive power management by dynamic voltage scaling, which in an asynchronous context naturally implies frequency scaling as well.
- virtually zero wake-up time, when returning from sleep mode, and frequency change recovery time. In synchronous circuits, both of these require from hundreds to thousands of clock cycles, due to the long settling time of Phase Locked Loops (e.g., the Atmel AVR requires 2 ms at 4MHz).
- minimization of energy consumption by working with a voltage supply very close to the process threshold.
- reduction of (estimated) electro-magnetic emission, with a promise of beneficial effects for the analog part of the circuit as well.

Our work uses the desynchronization approach [2] because it dramatically eases the transition from a synchronous to an asynchronous design style. This was a key requirement from the industrial partner of the project, STMicroelectronics.

It is well known that asynchronous circuits reduce electro-magnetic emission with respect to equivalent synchronous ones [11], [15], because they reduce the power consumption peaks in the vicinity of clock edges. Hence, they produce a flatter power spectrum and exhibit smaller voltage supply drops.

It is also well known that asynchronous circuits can be reliably operated at very low voltages when device characteristics exhibit second and third order effects, that is, when synchronous operation becomes very problematic [13]. This is, of course, especially true of quasi-delay insensitive or delay-insensitive circuits, since the datapath signals its completion reliably, no matter how slowly gates and wires switch. However, our experiments with a fabricated circuit [5] have shown that even cheaper bundled-data circuits can be operated at very low voltages. This is due to the fact that the delay characteristics of different gates, operating at the same low voltage, vary in a reasonably similar manner (see Fig. 7).

Asynchronous circuits also offer the possibility of dramatic increases in performance in the presence of extreme process and operating condition variability. This advantage, which was not a primary goal of this exercise due to the specific application domain, is permitted by the reduced margins that asynchronous design requires as compared to those required by synchronous. In the synchronous case, all design, manufacturing, and environment uncertainties must be handled by adding margins to the clock cycles. It is widely reported that these margins are currently in excess of 100% for a typical ASIC design flow and manufacturing process. This means that a device whose operating frequency is not determined a priori but is instead chosen at run time (depending on both fabrication and operating environment effects) can work twice as fast as an equivalent synchronous one.

We are thus confident that desynchronization can easily be used to perform initial experiments in the domain of aggressive power, energy, and EMI minimization. The hope is that more

radical asynchronous design techniques, which of course offer advantages along all these dimensions, will be embraced, once the advantages of asynchronicity have been shown.

This paper describes in detail the design flow that has been used and the various choices that have been made. While our design case study covers one specific architecture, we believe that the techniques that we propose can be applied to any digital architecture for low-power low-energy battery-powered applications.

The desynchronization flow has the following characteristics:

- It accepts as input a synchronous design, described with a VHDL or Verilog RTL specification.
- It uses standard EDA tools to synthesize, map, place, and route the design. Standard tools are also used for timing analysis, equivalence checking, local clock generation, extraction, scan insertion, automated test pattern generation, and so on. In this experiment, we used Synopsys Design Compiler and Cadence SOC Encounter for logic synthesis and physical design respectively, and Synopsys Nanosim for performance and power simulation.
- It fully automates the de-synchronization step [5], which removes the clock tree and replaces it with a network of asynchronous controllers.

2. PREVIOUS WORK

2.1 Wireless Sensor Network platforms and applications

Over the past few years, Wireless Sensor Networks have gained more and more attention as a basic technology enabling the vision of Ambient Intelligence. Nowadays, the advances of silicon processing technologies permit the miniaturization of low-cost devices that, by integrating sensing, computational, and communication capabilities, constitute the basic interface with the physical world. One of the most important constraints in Wireless Sensor Networks, mainly built with battery-powered nodes, is low power consumption, which has been driving research in the definition of both system and node architecture components [9]. Several device platforms, diversified in terms of storage, communication, and computational capabilities, have been identified as necessary for most Wireless Sensor Networks, corresponding to the layers of multi-tiered network architectures [8]. Berkeley Motes (i.e., the MICA Family) are a notable example of a general-sensing-class device with limited on-board capabilities. These motes, especially the MICA2 version, are extensively used by a huge number of research groups for Wireless Sensor Network prototyping in several typical ambient intelligence application scenarios. These include, for example, environmental monitoring [12], structural health monitoring [21], personal health monitoring [19], and tracking [18].

The most recently commercially developed version of the MICA family, i.e., the MICAZ, replaces the CC1000 radio used in MICA2 with a CC2420 radio that is compatible with the standard IEEE 802.15.4. This standard defines the physical and Medium Access Control layer specifications for low data rate wireless connectivity. Adopted by the Zigbee Industrial Consortium, the use of this standard is aimed at promoting the usage of low data rate networks in WSN application scenarios such as industrial and home automation.

Currently, much of the research on WSN design at the system level is based on reducing the quantity of raw sensing data transmitted by the nodes by increasing the amount of local computation performed on each node in order to reduce the overall data traffic on the network. This increases the computing capability requirements for each node and, as a consequence, increases the importance of reducing the power consumption of its digital components.

Our design example is thus oriented towards the definition of an application-independent architecture because it addresses a general purpose CPU design, namely, a desynchronized version of the same processor used in the MICA family motes. We also claim that desynchronization can also be profitably used to derive asynchronous implementations of dedicated hardware

unit for signal processing, encryption, and so on, in order to further reduce peripheral power and energy consumption.

2.2 Desynchronization

De-synchronization modifies the standard design flow for synchronous circuits [4] in three key steps:

1. *Conversion of the flip-flop-based synchronous circuit into a latch-based one (M and S latches in Fig. 1(b)).*
   D-flip-flops are conceptually composed of master-slave latches. To perform de-synchronization, this internal structure is explicitly revealed (see Fig. 1(b)) to decouple local clocks for master and slave latches (in a D-flip-flop, both derived from the same clock), so as to separate concerns between setup and hold time satisfaction. This step is essential in avoiding the low clock skew constraints that are inherent in the flip-flop-based design style.

2. *Generation of matched delays for the combinational logic* (denoted by rounded rectangles in Fig. 1(b)). Each matched delay must be greater than or equal to the delay of the critical path of the corresponding combinational block.
   Each matched delay serves as a completion detector for the corresponding combinational block.

3. *Implementation of the local controllers.*
   Each controller must open the latch after the following one has been closed (to avoid double latching or hold time violations) and close it after the previous one has been opened and the combinational logic has settled (to avoid setup time violations). Protocols and controller implementations are described more in detail in [2].

Fig. 2 depicts a synchronous netlist after the conversion into latch-based design. The shadowed boxes represent latches, whereas the white boxes represent combinational logic. Latches must alternate their phases.
Those with a label 0 (1) at the clock input represent the *even* (*odd*) latches. All latches are transparent when the control signal is high (CLK=0 for even, and CLK=1 for odd). Data transfers must always occur from even (master) to odd (slave) latches and vice-versa. Usually, this latch-based scheme is implemented with two non-overlapping phases generated from the same clock.

The correctness of the desynchronization approach, i.e., the fact that the asynchronous circuit implements exactly the same I/O functionality as its synchronous counterpart has been proved in [2].

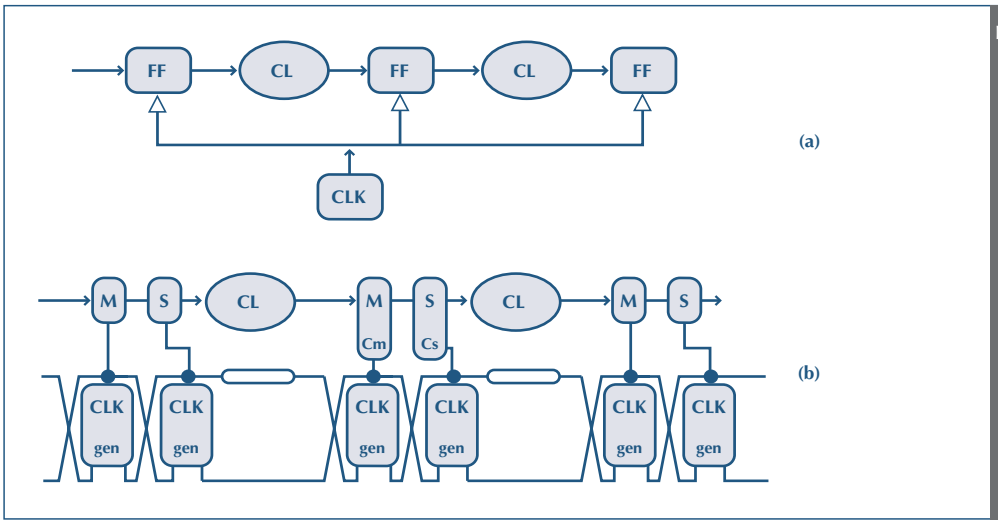

FIGURE 1: (A) SYNCHRONOUS CIRCUIT, (B) DE-SYNCHRONIZED CIRCUIT.

Interfacing with synchronous external peripherals and memories can be done in different manners, depending on their performance and clock precision requirements. Peripherals that require precise clocks (e.g., UARTs, timers) can be driven from an external clock, and then a synchronization interface can be added, or the external clock can be used directly to drive the desynchronized

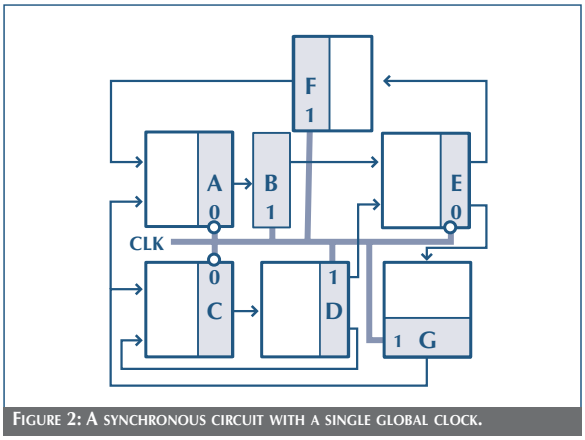

FIGURE 2: A SYNCHRONOUS CIRCUIT WITH A SINGLE GLOBAL CLOCK.

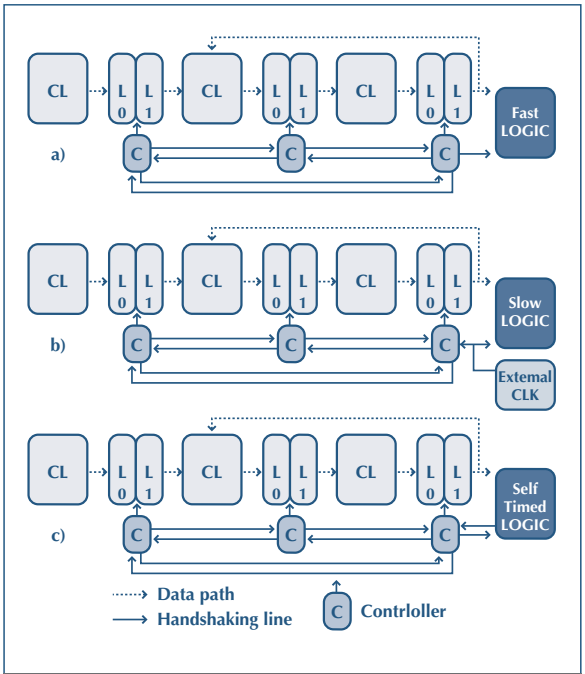

FIGURE 3: SYNCHRONOUS EXTERNAL LOGIC INTERFACING STRATEGY: (A) FAST LOGIC; (B) SLOW LOGIC OR PRECISE CLOCKS REQUIREMENTS; (C) SELF TIMED LOGIC.

controller inputs (see Fig. 3(b)), which are assumed to respond with an acknowledge faster than the clock. In the last case, only the EMI advantages of desynchronization can be achieved. Otherwise, peripherals which do not require a precise clock (see Fig. 3(a)) or which may be adapted to provide an acknowledge output (such as RAMs, see (Fig. 3(c)) can be driven by the clock output produced by the desynchronized controllers.

2.3. Dynamic voltage and frequency scaling

It is well-known that, even under normal operating conditions, the delay of a CMOS circuit scales almost linearly with its voltage supply, while its power scales quadratically. Thus, the normalized energy-per-cycle or energy-per-operation efficiency measure scales linearly with voltage supply. However, it is very difficult to use this optimization opportunity to the extreme, by operating very close to the threshold voltage (as we do in this paper), because:

1. library cells are seldom characterized by the manufacturer at such extreme operating conditions. Hence, the normal ASIC design flow is unsuitable for guaranteeing correct operation under the lowest voltage, energy, and power operating conditions.
2. the gate delay models deviate significantly from those used under nominal conditions and make a straightforward scaling of performance and power impossible, or at least very risky.
3. the effects of various random or hard-to-predict phenomena, such as threshold voltage variations, wire width variations, and local voltage supply variations due to IR drop are significantly magnified.

All this means that, even if one were able to use the traditional design flow for circuits to be operated at a voltage supply close to the transistor threshold voltage, the performance margins that one would have to use to ensure correct operation would be huge (they already exceed 100% under nominal conditions).

Two approaches have been proposed in the literature to tackle this problem with purely synchronous means. Both are based

on sampling the output of a signal which is forced to make a transition very close to the clock cycle and to slow down the clock frequency or increase the voltage supply if this critical sampling happens at the current voltage and frequency conditions.

The Razor CPU [7] is designed with double slave latches and an ex-or in each master-slave pair (thus increasing by over 100% the area of each converted latch). The second slave is clocked half a clock cycle later than the first slave. A difference in values between the slaves, as detected by the comparator, shows that the input changed very close to the falling edge of the clock of the first slave, consequently memorizing an incorrect value. The Razor in that case "skips a beat" and restarts the pipeline with the value of the second latch, which, assuming that environmental conditions change slowly, is always latched correctly. An external controller always keeps voltage and clock frequency very close to this "critical clocking" condition in order to operate the processor very close to the best *Vdd* point for the required clock frequency under the current temperature conditions.

The approach, while very appealing for processors, has an inherent problem that makes it inapplicable to ASICs. Due to the near-critical clocking, it is always possible that the first latch goes meta-stable. In that case, the whole detector and the clock controller may suffer from meta-stability problems. That case is detected with analogue mechanisms, and the whole pipeline is flushed and restarted. This is easy to do in a processor for which flushing and restarting is already part of a modern micro-architecture. It is, however, very difficult, if not impossible, to achieve automatically for a generic logic circuit, such as those found in modern ASICs.

Another technique that has been proposed to dynamically monitor the delay of the logic at the current voltage value and adjust the clock frequency accordingly is the PowerWise™ technology from National Semiconductors. It basically samples, with a high frequency clock, the output of a digital delay line that toggles once per system clock cycle. This is used, more or less as in Razor, to measure the delay of the line in the current environment conditions (temperature, *V dd*, etc.). The scheme is safer than that of Razor because it allows one to insert enough synchronizers after the delay line to reduce the meta-stability danger. However, it is an indirect measure and requires a complicated (patented) logic to monitor and control the clock speed.

Desynchronization achieves similar goals with much simpler logic because the delay line output is directly used to generate the clock period. Moreover, several controllers can be interspersed with the logic so that local conditions (e.g., temperature, *V dd*, transistor, and wire parameter variations) can be controlled much more closely, thus requiring potentially smaller margins.

Moreover, desynchronization reduces electro-magnetic emission (EMI) problems because the clock for every cluster of latches driven by a single controller is out of phase with respect to the others. This spreads the power spectrum, reducing both irradiated and on-chip power noise. It also can improve performance because it computes on-line the *useful skew* for each individual group of latches. The performance of the system is thus not given by the longest pipeline stage, as in the synchronous case, but by the maximum cycle mean[1]. This means that even unbalanced pipelines are automatically balanced at runtime by the handshaking mechanism.

2.4. Low-power asynchronous processors

Several asynchronous micro-processors reported in the literature exhibit very favorable power, energy, and EMI characteristics with respect to their synchronous counterparts. For example, the Lutonium from CalTech [14] implements the 8051 ISA and reportedly achieves 500 *pJ* per instruction at 1.8 *V* and MIPS, and 140 *pJ* per instruction at 0,9 *V* and 66 MIPS, with 180 *nm* a process.

According to [3], scaling CMOS technology to the next generation improves performance and reduces power

1. The maximum cycle mean is defined as the maximum overall cycles in the logic, which may include one or more latches, of the total cycle delay divided by the number of latches in the cycle.

consumption, thus reducing overall energy consumption by about 65% per generation. Based on this analysis, the Lutonium should achieve 175 *pJ* per instruction at full voltage supply and 49 *pJ* per instruction at minimum voltage supply, using a 130 *nm* process like our design.

The CalTech processors are based on Quasi-Delay-Insensitive (QDI) circuits, and are thereby extremely robust with respect to all sorts of variations, including *data-dependent computation delays*.

Desynchronization uses a bundled-data completion detection mechanism, which results in smaller circuitry but requires margins to ensure correct operation. As a result, it cannot exploit as effectively the inherent characteristics of asynchronicity (low power, modularity, etc.), but it is much simpler to implement, due to its use of a fairly standard design flow.

The Philips 8051 is another 8-bit micro-controller [17], which was synthesized starting from a Tangram specification. Unlike the Lutonium, the Philips 8051 used a bundled-data datapath, thus resulting in (presumably) lower area overheads. The main advantages over its synchronous counterpart, which led to its inclusion in some commercial products, were the low power consumption and, more significantly, the low EMI, which dramatically reduced the cost of implementing the analogue RF portion of a pager. A second-generation implementation of the Philips 8051, commercialized by Handshake Solutions using a 180 *nm* process, consumes 150 *pJ* per instruction delivering 6.3 MIPS at 1.8 *V*. If implemented with a 130 *nm* process, this micro-controller should achieve the energy consumption of 53 *pJ* per instruction.

The Tangram flow is very similar in spirit to the desynchronization flow because it uses mostly standard tools for physical design, design rule checking, and so on. However, it requires one to use a new proprietary language based on Communicating Sequential Processes, for specification and high-level simulation. Moreover, synthesis and design for testability tools are Tangram-specific. This is both an advantage, in that it improves the optimality of the result, and a disadvantage, because it requires extensive designer training and a partially new set of tools. In comparison, desynchronization is much easier for an experienced synchronous designer to pick up and use.

The SNAP/LE processor [6], [10] is an asynchronous low-power 16-bit processor that was designed specifically for Wireless Sensor Network applications. Its Instruction Set Architecture was optimized specifically for WSN applications because it operates in *event-driven* mode: basically, all code is executed as an interrupt service routine. It also features a set of instructions that directly control the various peripherals (timer, message handler) as co-processors rather than by using the normal memory-mapped register interface.
SNAP/LE, implemented in a process like the Lutonium, requires 218 *pJ* per instruction to deliver 240 MIPS at 1.8 *V* and 23 *pJ* per instruction to deliver 28 MIPS at 0.6 *V*. It was implemented, like the Lutonium, using the QDI implementation style, which is a significant departure from traditional synchronous design. If implemented with a process, this micro-controller should achieve the energy consumption of per instruction at full voltage supply and per instruction at minimum voltage supply.

Our processor, to be described more in detail in Section 4, was implemented in a 130 *nm* process (one generation after the other processors described above). It requires 14 *pJ* per instruction to deliver 170 MIPS at 1.2 *V* and 2.7 *pJ* per instruction to deliver 48 MIPS at 0.54 *V*. These results do not include program and data RAM access power. We estimated, from the data sheets of memory components for the process that we used, that using 4KB of RAM for code and data, as could be required by a small WSN application, would increase energy consumption by about 20%.

3. THE ASYNCHRONOUS AVR PROCESSOR

3.1. The synchronous RTL design

The project started from an open-source implementation of an 8-bit microcontroller that is compatible with the Atmel AVR architecture, the "AVR_CORE" available from the

`OpenCores.org` web site. We chose this processor because its Instruction Set Architecture is compatible with that of most WSN nodes of the MICA family.

Our aim was to demonstrate the possibility of asynchronously redesigning an existing standard CPU core for a WSN node in a very short time at reasonable cost. As mentioned above, we wanted to show reductions in terms of power consumption and noise emission as well as potential increases in performance, while keeping compatibility with the AVR instruction set. This permits the direct reuse of the binary software developed for this ISA.

The VHDL design from which we started provides a full implementation of the instruction set of the Atmel AVR microcontroller and includes various extra component, such as data memory, program memory, timers, and UART.

The design also includes a fully featured VHDL testbench that allows one to test the behavior of the entire platform. This was very important since it dramatically sped up all phases of design, from debugging the implementation of the flow all the way to measuring performance and power by simulation. (We used Mentor Graphics Modelsim and Synopsys Nanosim for this purpose).

3.2. Logic synthesis

After making sure via simulation that the downloaded "AVR_COR" indeed worked as advertised, we took the first step: the logic synthesis of the initial synchronous design.

For this step, we used Synopsys Design Compiler. A short script was written in order to:

- load the VHDL core file;
- define performance and area constraints;
- synthesize the netlist;
- map it on the chosen industrial standard cell library;
- modify the project hierarchy to suit the needs of the desynchronization step;
- estimate critical path delay; and
- save the resulting netlist on a Verilog file.

3.3. De-synchronization

The de-synchronization methodology [2] has been implemented, as described in [1], in the standard design flow of our industrial partner. The `desync` tool converts a technology-mapped synchronous netlist into an asynchronous one, performing the following steps:

- loading and parsing the Verilog input file;
- replacing registers with pairs of latches;
- automatically grouping the combinational logic into latch-separated islands;
- estimating, using Synopsys PrimeTime, the logic delay of each group;
- implementing the delay chains and the controllers;
- saving the de-synchronized output file in Verilog format.

In this case, we chose to use only one logic group (i.e., one controller for the master and one for the slave latches) because AVR_CORE only has 4500 gates organized as a 2-level pipeline (fetch and execute). This obviously prunes some of the de-synchronization advantages, principally, the electro-magnetic emission reduction.

The total area of the desynchronized core is about 5% larger than that of the synchronous one. Our results do not include the effect of placement and routing, which are unlikely to be significant for such a small design.

3.4. Logic simulations of the asynchronous core

The de-synchronized version of AVR_CORE was then simulated, again using Modelsim, in order to verify the correct behavior of the core. This required some minor modifications to the simulation settings and to the environment model. For example, we imported in Modelsim a new Standard Delay Format (SDF) file that contains the detailed gate-level delay information and library gate models.

Furthermore, we modified the VHDL testbench to properly connect the handshake lines added by the de-synchronization

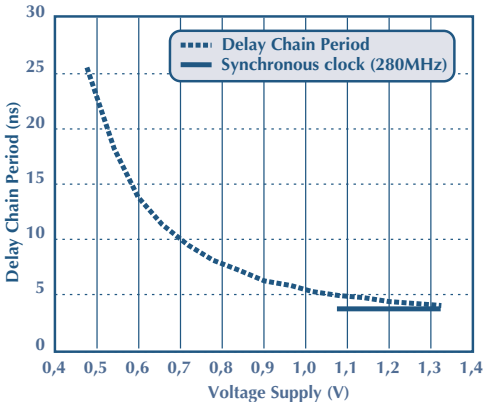

FIGURE 4: DESYNCHRONIZED PROCESSOR CYCLE PERIOD VERSUS SUPPLY VOLTAGE.

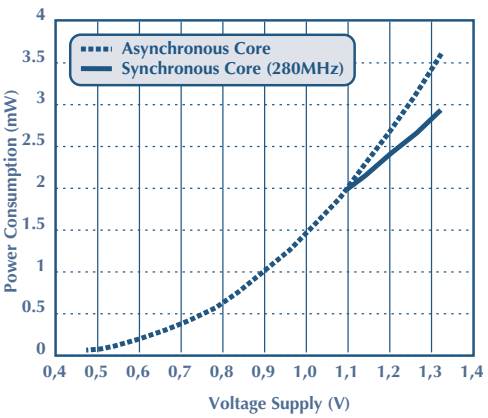

FIGURE 5: DESYNCHRONIZED PROCESSOR POWER VERSUS SUPPLY VOLTAGE.

tool instead of the externally generated clock. We connected each peripheral to the falling edge of the "Request Out" handshake signal of the desynchronized partition (we will call this signal "pseudo-clock"). After this, logical simulation was executed correctly.

3.5. Power estimations with Nanosim

Although Synopsys Design Compiler can perform power estimations for both synchronous and asynchronous circuits, using switching activity files (SAIF) derived from a simulator such as Modelsim, does not allow us to use it beyond the characterization corners of the standard cell library. Since we wanted to verify how power changes with very low supply voltages, something not usually considered among these corner cases, we had to perform transistor-level simulations.

For this purpose, we used Nanosim, an advanced transistor-level circuit simulation and analysis tool for analog, digital, and mixed-signal design verification. Nanosim uses the same electrical model parameters that simulators such as Spice use, but it has more simplified transistor and wire models that permit much faster simulation with a minimal loss of accuracy.

The inputs of the synchronous and desynchronized circuit were driven using test vectors generated by Modelsim, and the output vectors produced by both simulators under nominal conditions were compared against each other to check that no errors had occurred during the various synthesis, optimization, and translation steps. The next section describes the results of such simulations.

4. SIMULATION RESULTS AND ANALYSIS

The simulation results that we obtained indeed show the expected behavior when lowering the voltage supply of both the synchronous and the asynchronous circuit. The delay of each circuit decreases with increasing *Vdd* as shown in Fig. 4, while the power consumption increases with increasing *Vdd*, as shown in Fig. 5. The energy per instruction, shown in 6, increases almost linearly with the voltage supply. At the best power supply level, around 0.5 V, the desynchronized circuit is about 5 times more energy efficient than the synchronous one. Furthermore, the power consumption is reduced by more than one order of magnitude by scaling the voltage from the nominal level, 1.2 V, to about 0.5 V.

Of course, one could also scale the voltage supply for the synchronous circuit as well and obtain similar power reductions by using either the Razor or the PowerWise approaches mentioned above. However, this is much more difficult to do, as discussed in Section 2.3, because it amounts to essentially measuring the

timing of a delay line at a given supply voltage and then tuning the clock to match the circuit performance. Moreover, Razor would require one to change the AVR architecture drastically since it does not support pipeline squashing and restarting.

The desynchronized circuit offers the same advantages "for free." Moreover, it allows one to spread delay lines wherever they are needed to measure the effects of spatially correlated effects, for example, temperature and wire thickness[2]. It is also not subject to the meta-stability problems that are inherent in the near-critical sampling of a changing signal.

Fig. 7 shows the timing of the matched delay lines and of the actual critical path of the combinational logic, when varying *Vdd*. Remember that, in order to ensure correct operation of a bundled-data circuit, the delay lines must always be slower than the longest combinational path. The figure shows that the matched delays track the actual logic delay very well, and suggests that true measurement of logic delays, by dual-rail encoding, e.g., may not be necessary after all, at least for the technology under consideration, which is 130 *nm*. For this example, we did not try to fine-tune the delay line since performance was not the primary goal of the exercise, but we did want to keep the design fairly robust. Hence, we chose a margin of about 100%, which is close to what synchronous design would require under nominal conditions which is much better than what the same would require at low voltage supplies.

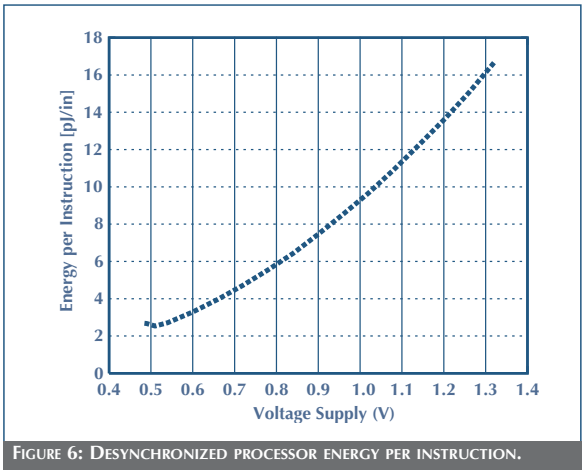

FIGURE 6: DESYNCHRONIZED PROCESSOR ENERGY PER INSTRUCTION.

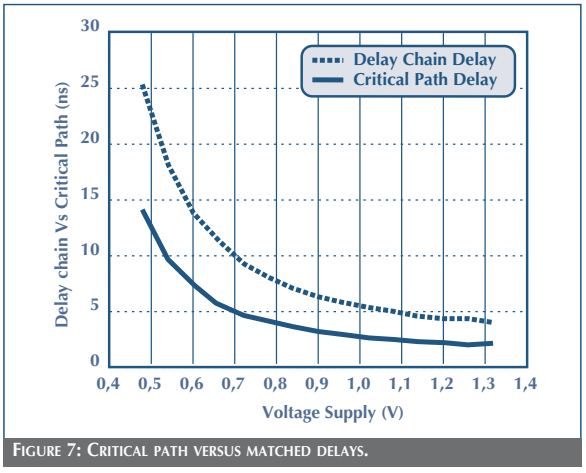

FIGURE 7: CRITICAL PATH VERSUS MATCHED DELAYS.

5. CONCLUSIONS

Our processor, which is ISA compatible with the Atmel AVR, requires 14 *pJ* per instruction to deliver 170 MIPS at 1.2 *V* and 2.7 *pJ* per instruction to deliver 48 MIPS at 0.54 *V*. These results seem to be much better than those of the other asynchronous processors reported in Section 2.4 (which were implemented with a 180 *nm* process), even when those are scaled to 130 *nm*. This means that, even after considering the improvement in technology over one generation, which reduces according to [3] energy consumption by 65%, and after considering RAM access energy, which would add approximately 20% energy consumption for 4 KB of RAM, the results are quite comforting with respect to previous work in both the synchronous and the asynchronous domain.

We can attribute them to the *extremely low area and performance overhead* that desynchronization imposes over synchronicity. Hence, the margins that are required due to the bundled-data

2. Wire thickness, which affects both resistance and side capacitance, is spatially correlated, due to the effects of the chemical-mechanical polishing step.

approach (about 2X, as discussed in Section 4) are still much less than the overhead due to dual-rail implementation of QDI circuits and to the extremely low granularity controllers required by the Tangram design flow.

Based on the results of this work, we can claim that the only *inherently robust* manner of achieving extreme Dynamic Voltage and Frequency Scaling is by using asynchronous techniques. Such techniques *measure reliably*, rather than either *estimate* or *measure with inherent meta-stability risks*, the actual delay of combinational logic, thus taking into account manufacturing process variations, environmental conditions, and so on.

The desynchronization technique is not perfectly and natively asynchronous because all blocks still operate in lockstep with respect to each other. However, it provides the means to measure logic delays by using matched delay lines. Its very easy adoption, which was also shown in this project by re-implementing an existing legacy RTL core *without any re-design and almost without any knowledge of asynchronous design techniques*, makes it an appealing first step in the direction of more widespread adoption by industry. We believe that Wireless Sensor Networks provide a very promising application domain for very low-power and low-energy digital electronics, due to the increasing computational requirements and the stringent constraints on battery life. In the future, we will explore in more detail how the low-energy and (hopefully) low-EMI characteristics of our processor can be exploited in actual WSN applications. We will also verify how well delay lines track the critical path delay in the presence of process variability.

ACKNOWLEDGMENTS

The authors would like to thank Nikos Andrikos and Christos Sotiriou from ICS-FORTH for the extensive support they provided us while using their desynchronization tool and Maurizio Tranchero from Politecnico di Torino for his help during the first phase of this project.

REFERENCES

[1] N. Andrikos. **Personal communication**.

[2] I. Blunno, J. Cortadella, A. Kondratyev, L. Lavagno, K. Lwin, and C. Sotiriou, "HANDSHAKE PROTOCOLS FOR DE-SYNCHRONIZATION," Proc. International Symposium on Advanced Research in Asynchronous Circuits and Systems, pages 149-158, 2004.

[3] Shekhar Borkar, "DESIGN CHALLENGES OF TECHNOLOGY SCALING," IEEE Micro, July-August 1999.

[4] D. Chinnery and K. Keutzer, editors, **Closing the Gap between ASIC and Custom: Tools and Techniques for High-Performance ASIC design**, Kluwer Academic Publishers, 2002.

[5] J. Cortadella, A. Kondratyev, L. Lavagno, and C. Sotiriou, "DESYNCHRONIZATION: SYNTHESIS OF ASYNCHRONOUS CIRCUITS FROM SYNCHRONOUS SPECIFICATIONS," IEEE Transactions on CAD, 2005.

[6] V. Ekanayake, C. Kelly, and R. Manohar, "AN ULTRA LOW-POWER PROCESSOR FOR SENSOR NETWORKS". In Proceedings of the international conference on Architectural support for programming languages and operating systems, (ASPLOS), 2004.

[7] D. Ernst, S. Das, S. Lee, D. Blaauw, T. Austin, T. Mudge, N. S. Kim and K. Flautner, "RAZOR: CIRCUIT-LEVEL CORRECTION OF TIMING ERRORS FOR LOW-POWER OPERATION," IEEE Micro, November 2004.

[8] J. Hill, M. Horton, R. Kling, and L. Krishnamurthy, "THE PLATFORMS ENABLING WIRELESS SENSOR NETWORKS," Communications of the ACM, 47(6), 2004.

[9] J. Hill, R. Szewczyk, A. Woo, S. Hollar, D. Culler, and K. Pister, "SYSTEM ARCHITECTURE DIRECTIONS FOR NETWORKED SENSORS," Architectural Support

for Programming Languages and Operating Systems, pages 93-104, 2000.

[10] Clinton Kelly, Virantha Ekanayake, and Rajit Manohar, "SNAP: A SENSOR NETWORK ASYNCHRONOUS PROCESSOR," Proc. International Symposium on Advanced Research in Asynchronous Circuits and Systems, pages 24-33, IEEE Computer Society Press, May 2003.

[11] Joep Kessels and Ad Peeters, "THE TANGRAM FRAMEWORK: ASYNCHRONOUS CIRCUITS FOR LOW POWER," Proc. of Asia and South Pacific Design Automation Conference, pages 255-260, February 2001.

[12] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, "WIRELESS SENSOR NETWORKS FOR HABITAT MONITORING," ACM International Workshop on Wireless Sensor Networks and Applications (WSNA'02), September 2002.

[13] Alain J. Martin, Andrew Lines, Rajit Manohar, Mika Nyström, Paul Pénzes, Robert Southworth, and Uri Cummings, "THE DESIGN OF AN ASYNCHRONOUS MIPS R3000 MICROPROCESSOR," Advanced Research in VLSI, pages 164-181, September 1997.

[14] Alain J. Martin, Mika Nyström, Karl Papadantonakis, Paul I. Pénzes, Piyush Prakash, Catherine G. Wong, Jonathan Chang, Kevin S. Ko, Benjamin Lee, Elaine Ou, James Pugh, Eino-Ville Talvala, James T. Tong, and Ahmet Tura, "THE LUTONIUM: A SUB-NANOJOULE ASYNCHRONOUS 8051 MICROCONTROLLER," Proc. International Symposium on Advanced Research in Asynchronous Circuits and Systems, pages 14-23, IEEE Computer Society Press, May 2003.

[15] J. McCardle and D. Chester, "MEASURING AN ASYNCHRONOUS PROCESSOR'S POWER AND NOISE," Synopsys User Group (SNUG), 2001.

[16] A. Perrig, J. Stankovic, and D. Wagner, "SECURITY IN WIRELESS SENSOR NETWORKS," Communications of the ACM, 47(6), June 2004.

[17] Philips Semiconductors, "P87CL888; 80C51 ULTRA LOW POWER (ULP) TELEPHONY CONTROLLER"

[18] C. Sharp, S. Schaffert, A. Woo, N. Sastry, C. Karlof, S. Sastry, C. Karlof, S. Sastry, and D.Culler, "DESIGN AND IMPLEMENTATION OF A SENSOR NETWORK SYSTEM FOR VEHICLE TRACKING AND AUTONOMOUS INTERCEPTION," In European Workshop on Wireless Sensor Networks, January 2005.

[19] V. Shnayder, B r. Chen, K. Lorincz, T. Fulford-Jones, and M. Welsh, "SENSOR NETWORKS FOR MEDICAL CARE," Technical Report TR-08-05, Division of Engineering and Applied Sciences, Harvard University, 2005.

[20] H. Wang, D. Estrin, and L. Girod, "PREPROCESSING IN A TIERED SENSOR NETWORK FOR HABITAT MONITORING", EURASIP JASP, special issue of sensor networks, 2003.

[21] N. Xu, S. Rangwala, K.K. Chintalapudi, D. Ganesan, A. Broad, R. Govindan, and D.Estrin, "A WIRELESS SENSOR NETWORK FOR STRUCTURAL MONITORING," Proceedings of the international conference on Embedded networked sensor systems, 2004.

[22] J. Zhao, R. Govindan, and D. Estrin, "COMPUTING AGGREGATES FOR MONITORING WIRELESS SENSOR NETWORKS" In International Workshop on Sensor Network Protocols and Applications, May 2003.

■ CONTACT: ST.JOURNAL@ST.COM ■

IMPLEMENTING ECC FOR 8-BIT SYSTEMS AND POWER CONSUMPTION CONSIDERATIONS

Guido Bertoni[1],
Luca Breveglieri[2], Matteo Venturi[2]

(1) STMicroelectronics
(2) Politecnico di Milano

Public key cryptography is widely considered as the fundamental building block for key exchange; different public key algorithms are standardized and used in many applications. Among them, ECC (Elliptic Curve Cryptography) is considered the best solution in terms of security, computational requirements, and storage needs for secret and public keys. Energy consumption is one of the main constraints to be considered in wireless sensor networks. In the case of sensor networks, the typical approaches of minimizing latency via a complete hardware coprocessor and of reducing area overhead via an efficient implementation of finite field operations might not provide the best solution. In this paper, two coprocessors for minimizing both additional resources and power consumption are presented. Costs and performances of such coprocessors are compared with known results, showing that space exists for the reduction of energy consumption without degrading the other performance figures.

1. INTRODUCTION

The implementation of security services, especially those related to secure data exchange, is becoming an indispensable requirement for sensor network applications [1]. In high-end systems, the cryptographic applications have already undergone intensive development, while the security services implemented in low-power low-cost systems are still in a primitive state. This is a strong limitation which could prevent sensor networks from covering a wider range of applications. A pervasive diffusion of this technology will occur only if security requirements are efficiently matched by the entire system. One of the major barriers to deploying security on sensor networks is the limited computation capabilities, particularly critical for public key primitives.

Future technological advances will probably push the networks towards lower cost and energy consumption rather than to higher performances in computation [2]. Thus, it is essential to find a smart solution to ensuring satisfactory security services which will minimally impact performance and energy consumption

while exploiting the few available system resources as much as possible.

Public key cryptography is a building block that allows secure communication in dynamic environments such as sensor networks without the need of a permanently on-line central authority [3]. The most efficient schemes that provide cryptographic protocols are the Elliptic Curve Cryptosystems (ECC) [4][5]. The most important operations performed by those systems is *the scalar multiplication, or kP*. The execution of one *kP*, for example, is necessary to compute a decryption transformation, or a digital signature generation. The calculation of two *kP* allows an encryption transformation, a signature verification, or a key exchange. The security protocols based on ECC are standardized by different institutions such as IEEE, ANSI, NIST, and ISO [4].

Recent publications have evaluated the latency of software implementations of ECC and the speed-up obtained with dedicated instructions, or with hardware multipliers, but none of them focused on the energy spent in ECC computations, where traditional metrics have been used, as latency and silicon area demand. Most of the previous scientific publications have explored the field of high performance, where the only constraint is the computational time.

We believe that, in applications such as sensor networks, the latency of public-key operations is relatively important compared to energy consumption because most sensor networks have a limited bandwidth and relatively low communication latency. The aim of this work is to evaluate the energetic cost of different ECC implementations for low-end systems and to propose a new hardware coprocessor architecture which not only speeds up cryptographic operations but minimizes energy consumption over all.

The paper is organized as follows: section 2 presents the main topics of public key cryptography; section 3 briefly introduces the ECC mathematical background; section 4 presents the previous work in this field; section 5 evaluates the energy cost of cryptographic operations; and section 6 presents the new coprocessors and its performances. The conclusions of the work can be found in section 7.

2. PUBLIC KEY CRYPTOGRAPHY

Cryptography offers techniques to protect communications and to provide security services. These techniques consist of applying an encryption scheme to transform data.

Encryption primitives are known in mathematics as *one-way functions with a trapdoor*. The idea is that once the encryption has transformed the information, it is computationally unfeasible to recover the original content without the *trapdoor information*. This piece of information, in cryptographic systems, is called a *key*. The use of keys in cryptographic protocols is very useful. When the security of the scheme relies on the secrecy of the encryption and decryption functions, the entire scheme has to be redesigned if these functions are revealed. On the contrary, when security relies only on a secret key and the transformations are publicly known, a simple key change can secure the whole system once again.

When two entities want to communicate in a secure way, one possibility is that they agree upon a unique secret key k and use k to encrypt messages. The systems following this idea are called *symmetric-key cryptosystems*, and the Advanced Encryption Standard (AES) is an example thereof. The use of a common secret key allows fast cryptographic operations, but the problems of key agreement and management turn out to be unsolvable in large and dynamic networks and in remote communications.

The problem of key management and agreement is solved by the so called *public-key systems* (proposed for the first time in [3]) at the expense of more complex operations. In public-key schemes, each entity owns a pair of keys: a public key k_p and a secret key k_s. The key k_p is publicly known, and it can be used by other entities when they want to send an encrypted message to

its owner. The only possibility to recover the original message is by using k_s, only known by the legitimate owner of k_p.

For obtaining the best performances from the two cryptosystems, a hybrid approach is generally adopted. A session key is exchanged using public-key cryptography and is then used in a symmetric-key cryptosystem for encrypting data.

Public key cryptography is commonly used in Internet security protocols such as SSL or IPSec, smartcard applications, and Digital Right Managements. The dynamic composition of a sensor network can benefit from the characteristics of public-key cryptography.

A typical use of a public key algorithm is the so called Diffie-Hellman key exchange. The protocol is a sequence of three steps: private/public key generation, public key exchange, and shared secret derivation. The result of the protocol is a function of the parameters and of the two private keys, but no secret has been exchanged over the insecure channel. A detailed explanation of the DH is given in the next section, using ECC as the public key primitives.

3. ELLIPTIC CURVE CRYPTOSYSTEMS

The theoretical security of a cryptographic system is strictly related to the hardness of the mathematical problem defining the system itself. There are three mathematical problems exploited in public key cryptography applications: the Integer Factorization Problem (IFP), the Discrete Logarithm Problem (DLP), and the Elliptic Curve Discrete Logarithm Problem (ECDLP). While sub-exponential algorithms are known to solve the IFP and the DLP, only fully-exponential algorithms have been proposed to solve the ECDLP [4]. This means that ECDLP-based systems can provide the same level of security as DLP- or IFP-based systems, using smaller key sizes. As the computing time of the basic operations is directly proportional to the operand size, the implementation of ECDLP schemes represents a great advantage, especially when calculations are executed by low-end processors. Elliptic curves (EC) are defined by the Weirestrass equation:

$$E : y^2 + a_1 x y + a_3 y = x^3 + a_2 x^2 + a_4 x + a_6. \quad (1)$$

One point $P(x_P; y_P)$ belongs to the curve E if its coordinates satisfy equation (1). In cryptographic applications, the coordinates x_P and y_P are elements of a finite field $GF(q)$, where q is the number of field elements. Curve points can be represented as well in projective coordinates $P(X_P : Y_P : Z_P)$. The curve point operations are expressed in terms of field operations that involve point coordinates. An exhaustive presentation of the EC theory can be found in [4].

The most important operation in Elliptic Curve Cryptosystems (ECC) is the scalar multiplication, briefly referred to as kP. This operation is defined as:

$$Q = kP = P + P + \ldots + P \ (k \ times). \quad (2)$$

Solving the ECDLP means to find the integer k such that $Q = kP$, knowing the starting point P, the resulting point Q, and the parameters of the curve. The execution time of EC cryptographic schemes is dominated by scalar multiplication. In this work, scalar multiplication is the key issue leading comparisons between different implementation strategies. Since scalar multiplication is difficult to be inverted by an attacker, it is easier to see that k is the secret key while Q is the public key.

The most efficient algorithm for calculating scalar multiplication without precomputation but with a minimum use of inversion in the finite field is the modification of the Montgomery Scalar Multiplication proposed by Lopez and Dahab [12]. This method is based on projective coordinates, executes only one inversion during a kP, and requires the use of two points in projective coordinates as temporary variables. In the case of small architectures, the advantages provided by this algorithm are remarkable. This explains why the Montgomery algorithm is

usually implemented on low-power devices [5][6][7] and why it is used in our implementations as well.

In this paper, we use a binary finite field, $GF(2^{163})$, for the definition of the elliptic curve. We considered binary fields because they have some arithmetical properties, such as the absence of carry propagation, which provides good implementation advantages. We selected a field size of 163 bits because it is widely considered the appropriate key size for today's security requirements [10].

ECC can be modeled using architecture with 4 different layers, an architecture which reflects the conceptual hierarchy of implemented operations. The first layer is represented by the hardware instructions available on the target device; the second level defines the operations over binary fields $GF(2^m)$ (addition, multiplication, etc.); the third level implements the operations on curve points *(P+Q, 2P, kP)*; and the fourth and upper level defines the cryptographic protocols. The security protocols that can be implemented (D-H, ElGamal, ECDSA, etc.) provide key exchange, encryption/decryption, and digital signature services.

Diffie Hellman is one of the most used public key protocols; it allows sharing a secret key through an insecure channel in order to exchange data using symmetric algorithms, which are faster and easier to be executed in software.

Let's consider two devices, *A* and *B*. Every device selects a random number smaller than the group order. This is the secret key, which we represent as k_a and k_b for the two devices. The private key is kept secret and not shared with any other device in order to protect the privacy of the communications.

Now every device computes its own public key, Q_a and Q_b, respectively. Up to this phase, all the computations could be done off-line and possibly only once during the life of the device.
Once the online computation starts, the device exchanges the public keys, and once public key is received, device A calculates $Q=k_aQ_b$ while B calculates $Q=k_bQ_a$.

The result of the calculation is equal for the two devices, as shown by $Q=k_aQ_b=k_bQ_a=k_ak_bP$.
No secret is exchanged through the unsecured channel. The only attack possible is the man-in-the-middle attack, where the attacker substitutes its public key with the two exchanged over the air. This problem is solved through the use of certificates or certification authorities. This problem is well known in the field of cryptography and is out of the scope of the paper. Reference [4] provides a description of known solutions and problems.

4. RELATED WORK

The number of papers concerning the implementation of ECC in low-end environments is increasing. Nevertheless, time performances often remain the only evaluation parameter throughout the research.

In [5], Gura et al. propose a software implementation that executes one *kP* on a 160-bit prime field in 0.81 s on the ATmega128 processor at 8 MHz. They propose a new multiply-&-accumulate instruction that reduces the calculation time to 0.59 s. In [6], the *kP* is calculated on the 163-bit binary field using the ATmega128. Its execution takes 4.14 s in software implementation and 0.29 s with new instructions designed for binary fields. In [7], Kumar and Paar present an FPGA hardware multiplier such as an ISE for the ATmega128. They measured 0.113 s to execute one *kP* over $GF(2^{163})$ at 4 MHz.

In [8], Savaş et al. introduce a coprocessor containing a multiplier and a divider for prime finite fields. The time required for a scalar multiplication is 31.9 ms at 20 MHz for a prime field *GF(p)* of 168 bits. They use the TSMC 0.13 μm ASIC library for synthesis and obtain a coprocessor size of 30 kgates and a power consumption of 0.99 mW at 20 MHz and 1 V power supply. In [9], Aigner et al. present a coprocessor for 8-bit architecture smart-cards. The data path mainly consists of leaf cells which integrate multiplication, addition, and inversion. The coprocessor has been synthesized by a 0.13 μm CMOS process, resulting in 25-kgate size. The scalar multiplication over $GF(2^{191})$ is executed in 34.1 ms at 10 MHz.

5. ENERGY CONSUMPTION DURING A *kP*

In this work, we considered the implementation of public-key ECC over the binary field GF(2^{163}). This cryptosystem has been implemented on the Atmel 8-bit AVR ATmega128. The choice of this processor is due to its technical characteristics and to the previously published work [5], which chooses it as the target processor.

ATmega128 has a RISC architecture and can work at a clock frequency up to 16 MHz. For our implementation, we fixed the clock frequency at 8 MHz. The CPU contains a Register File of 32 temporary 8-bit registers; the memory access latency is 2 clock cycles. The processor can typically access a 4-kbyte RAM. The limited resources of the processor yield a low-power consumption, which makes such a processor suitable for applications in scarce-energy environments. The average power consumption of ATmega128 is 12-15 mW [13].

Considering the results reported in the related literature, it is possible to make a rough evaluation of the energy consumed for executing one *kP* operation. In [6], scalar multiplication takes 4.14 s in a software implementation and 0.29 s after an ISE of the CPU core. The new instructions want to improve the field multiplication efficiency because it is the most time-consuming operation during a *kP* (89% of the total execution time for software implementations). This happens because the integer arithmetic instructions of standard processors support binary field arithmetic rather inefficiently. These machine instructions are obtained with few modifications to the integer multiplier usually included in the processor ALU; the hardware overhead required for the ISE, compared to the actual multiplier size, is estimated at about 5%. It is reasonable to assume that the cost overhead after the ISE is quite negligible compared to the cost of the entire processor. Also the power consumption of the processor should remain unchanged after the extension of the multiplier structure.

An important remark for energy consumption is that cryptographic calculations normally require multiple precision operations and that the size of the operands can be very large when they are represented in small architectures. This results in a very large number of memory accesses performed during one kP. Compared to the average power consumption of a 12-15 mW 8-bit CPU, the power consumption of one RAM access is remarkable. In fact, we estimated the power consumption of an embedded 4-kbyte RAM working at 8 MHz produced by ST Microelectronics, and we determined that, during one memory access, the power absorption is of 70 mW. For this reason, the number of memory accesses performed during program execution can represent an important parameter for evaluating energy consumption.

Having taken into account the figures reported in [5], we calculated that over 4.15 million memory accesses are performed during a software execution of the Montgomery algorithm on 8-bit platforms and over 0.43 million with binary field instructions.

It is possible to estimate the energy spent for executing one *kP* by summing up the energy consumed by the CPU and all memory accesses. It turns out that approximately 120 mJ are spent in the software implementation and 11.1 mJ after the ISE.

6. DEDICATED COPROCESSOR: DESIGN AND RESULTS

The performance of the systems can be largely increased if the cryptographic calculations are committed to a dedicated coprocessor capable of managing data blocks much larger than the existing CPU architecture. In the specific case of our implementation, the operands have a size of 163 bits, while the existing data path is only 8 bits. We have designed three different hardware devices in VHDL, and we have synthesized them using the 0.18 um CMOS technology library designed by STMicroelectronics. Using the Synopsys tools Design Compiler and PrimePower, we obtained the area occupancy, the critical path, and the power consumption for each implementation.

As the most time consuming operations during the *kP* are field

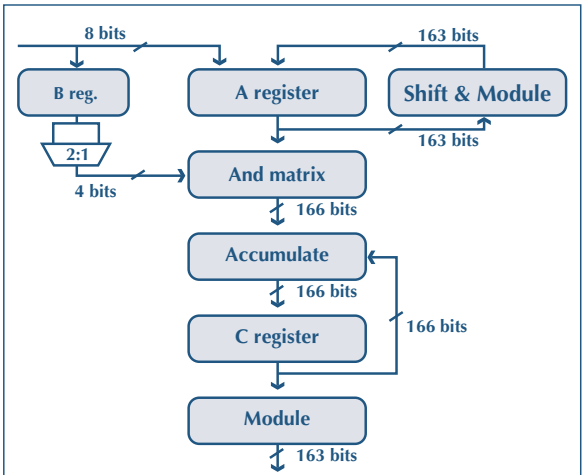

FIGURE 1: STRUCTURE OF THE HARDWARE MULTIPLIER.

multiplications, the first coprocessor we implemented was a serial-digit hardware multiplier for the binary field $GF(2^{163})$. Its structure is shown in Fig. 1. The multiplication is carried out in several iterations by multiplying a 163-bit multiplicand (register A) by a few bits of the second operand (register B). The partial result is accumulated into register C at each multiplication iteration. Thanks to the hardware multiplier, it is possible to reduce the number of memory accesses performed during a *kP* to 255,000 I/O instructions. The time performances provided by this multiplier depend also on the implemented digit size *d*: by choosing d=4 bits, we are able to reach a good trade-off between additional costs and performances.

After designing a hardware multiplier, we were able to improve the efficiency of field operations considerably. At this point, further improvements should be obtained by optimizing memory operations, in particular by reducing the intensive data flow between RAM and coprocessor. The expected advantages concern the time performances and, more importantly, the reduction of energy consumption.

The second coprocessor we designed targets this goal. In order to reduce the memory operations, we added to the previous multiplier three temporary registers of 163 bits each. These registers can store the last three multiplication results so that it is no longer necessary to write a result into the RAM if it is an operand of the following multiplication.

In order to improve data flow efficiency, the coprocessor executes a hardware logical XOR between two register outputs. It is thus able to execute a time-costless field addition without performing any memory operations. This 3-register coprocessor autonomously executes one iteration of the Montgomery algorithm, with an optimized memory access strategy. The number of memory accesses performed during a kP drops to 47,628.

Further reductions of data flow are possible by adding four new 163-bit registers, thus obtaining a 7-register coprocessor. Two of these registers store the *x* coordinate of *P* and the coefficient *b*, both used throughout the calculation of *kP* (registers *Rx* and *Rb*). The other two registers store the coordinates X_b and Z_b at the beginning of each iteration (registers RX_b and RZ_b). This coprocessor only has to access 84 RAM words, containing the *x* and *z* coordinates of points *A* and *B*. Inverting the point addition and doubling procedures, these coordinates are loaded and stored from and into RAM only once at each iteration. The additional

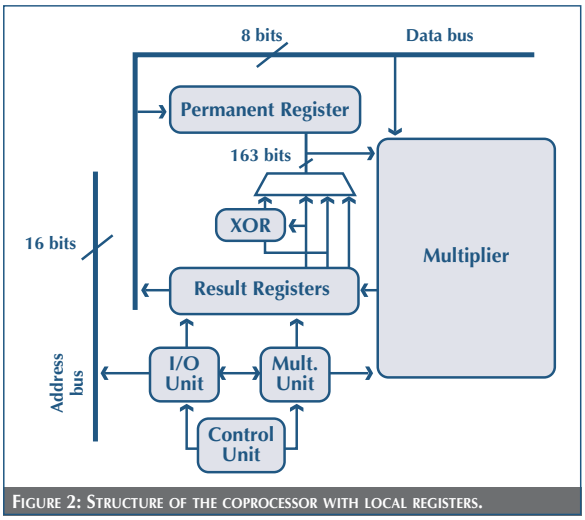

FIGURE 2: STRUCTURE OF THE COPROCESSOR WITH LOCAL REGISTERS.

resources available to this coprocessor do not aim mainly at timing improvements,; they rather intend to increase the efficiency of data transport, reducing energy consumption as much as possible. A detailed presentation of the sequence of operations performed by the 7-register coprocessor can be found in appendix A. The structure of the coprocessor is depicted in Fig. 2.

The silicon surface and the power consumption of the hardware implementations are reported in Table 1. As a term of comparison, we report the same parameters for an 8-bit processor. As it often comes in the implementation of a cryptographic coprocessor, the silicon area required by dedicated devices is comparable or larger than the area of the processor. Nevertheless, it is interesting to point out the good energetic properties of the new coprocessors, which match the needs of low power environments.

Architecture	Area (gates)	Consumption
4-bit digit multiplier	5,425	172 μW
3-reg coprocessor	11,957	305 μW
7-reg coprocessor	18,550	503 μW
Typical 8bit processor	6,500	12 mW

Table 1: Silicon area and power consumption of hardware implementations.

In Table 2, we report the timing results and the memory operations performed by the two coprocessors. Compared with the previously analyzed implementations, the advantages deriving from the use of a dedicated device are evident. The performance gap of the two hardware devices shows how the intensive data flow affects the cryptographic operations. Using

Coprocessor	Execution Time	Number of Memory operations	Energy consumption
Software impl.	2.07	4,148,496	120 mJ
ISE implem.	0.29	435,618	11.1 mJ
4-bit digit mult	97.56 ms	255,550	5.95 mJ
3-reg coprocessor	17.05 ms	47,62	1.09 mJ
7-reg coprocessor	14.68 ms	27,216	0.66 mJ

Table 2: Execution time, memory operations, and energy consumption.

Coprocessor	Additional Area (gates)	Execution Time
Software implementation	-	2.07 s
ISE implementation	5,425	0.29 s
4-bit digit multiplier	11,957	97.56 ms
3-reg coprocessor	18,550	17.05 ms
7-reg coprocessor	6,500	14.68 ms

Table 3: Comparison of area time trade off.

the results reported in Tables 1 and 2, we calculated an energy consumption of 5.95 mJ for calculating one kP on the hardware multiplier and 1.09 mJ on the 3-register coprocessor, while the consumption is reduced to 0.66 mJ for 7 registers. Table 3 shows the trade off between area increase and execution time.

7. CONCLUSIONS

Elliptic Curve Cryptography is the best candidate for effective and efficient implementation of public key primitives in a wireless sensor network. Other research has been already presented for the implementation of ECC in software and hardware, but none of them has focused on the compromise of low-energy consumption, reduced silicon area requirements, and significant speed-up compared to software solutions.

Different hardware support has been proposed for improving timing and power consumption. The coprocessor with 3 register is probably the best compromise for low-end nodes like those used in sensor networks; it computes a scalar multiplication in 17 ms @ 8 MHz with a request of 12 Kgates and an energy consumption of 1.09 mJ. The solution seems very attractive compared with previously known figures.

9. APPENDIX A

Point Doubling: $B \leftarrow 2B$

1. load Z_b into register RZ_b;
2. load X_b into register RX_b (and simultaneous calculation of RZ_b^2);
3. calculate RX_b^2;

4. calculate $Z_b \leftarrow RZ_b{}^2 \otimes RX_b{}^2$;
5. store Z_b into memory (and simultaneous calculation of $RZ_b{}^2 \otimes RZ_b{}^2$);
6. calculate $RX_b{}^2 \otimes RX_b{}^2$;
7. calculate $Rb \otimes RZ_b{}^4$;
8. store into memory $RX_b{}^4 \oplus (Rb \otimes RZ_b{}^4)$.

Point Addition: $A \leftarrow A + B$

1. load X_a into register B (and simultaneous calculation of $X_a \otimes RZ_b$);
2. load Z_a into register B (and simultaneous calculation of $Z_a \otimes RX_b$);
3. calculate $Z_a \leftarrow [(X_a \otimes RZ_b) \oplus (Z_a \otimes RX_b)]^2$;
4. store Z_a into memory (and simultaneous calculation of $(X_a \otimes RZ_b) \otimes (Z_a \otimes RX_b)$);
5. calculate $Px \otimes Z_a$;
6. store $X_a \leftarrow (Px \otimes Z_a) \oplus [(X_a \otimes RZ_b) \otimes (Z_a \otimes RX_b)]$ into memory.

REFERENCES

[1] A. Perring, H. Stankovic, and D.Wagner. "SECURITY IN WIRELESS SENSOR NETWORKS," in Communications of the ACM, volume 47, pages 53-57, 2004.

[2] J. Hill, R. Szewczyk, A. Woo, S. Hollar, D. Culler, and K. Pister. "SYSTEM ARCHITECTURE DIRECTIONS FOR NETWORKED SENSORS," in ACM, editor, ASPLOS IX, pages 93-104, November 2000.

[3] W Diffie and M. Hellmann. "NEW DIRECTIONS IN CRIPTOGRAPHY," in Transaction in Information Theory, IT-22, pages 644-654. IEEE Computer Society, 1976.

[4] D. Hankerson, A. J. Menezes, and S. A. Vanstone. "GUIDE TO ELLIPTIC CURVE CRYPTOGRAPHY," Springer-Verlag, 2004.

[5] N. Gura, A. Patel, A. Wander, H. Eberle, and S. Chang Shantz. "COMPARING ELLIPTIC CURVE CRYPTOGRAPHY AND RSA ON 8-BIT CPUS," in M. Joye and J. J. Quisquater, editors, Cryptographic Hardware and Embedded Systems — CHES 2004, volume LNCS 3156, pages 119–132. Springer-Verlag, 2004.

[6] H. Eberle, A. Wander, N. Gura, and S. Chang-Shantz. "ARCHITECTURAL EXTENSIONS FOR ELLIPTIC CURVE CRYPTOGRAPHY OVER GF(2^m)," Sun Microsystems Laboratories, 2005. research.sun.com/sunlabsday/docs.2004/Micro.pdf.

[7] S. Kumar and C. Paar. "RECONFIGURABLE INSTRUCTION SET EXTENSION FOR ENABLING ECC ON AN 8-BIT PROCESSOR," Chair for Communication Security, Ruhr-Universität Bochum, Germany, 2004. www.crypto.ruhr-uni-bochum.de /imperia/md/../fpl04/final.pdf.

[8] E. Öztürk and B. Sunar and E. Savaş. "LOW-POWER ELLIPTIC CURVE CRYPTOGRAPHY USING SCALED MODULAR ARITHMETIC," in M. Joye and J. J. Quisquater, editors, Cryptographic Hardware and Embedded Systems — CHES 2004, volume LNCS 3156, pages 133–147. Springer-Verlag, 2004.

[9] H. Aigner and H. Bock and M. Hütter and J.Wolkerstorfer. "LOW-COST ECC COPROCESSOR FOR SMART CARD," in M. Joye and J. J. Quisquater, editors, Cryptographic Hardware and Embedded Systems — CHES 2004, volume LNCS 3156, pages 107–118. Springer-Verlag, 2004.

[10] A.K. Lenstra and E.R. Verheul. "SELECTING CRYPTOGRAPHIC KEY SIZES," in Journal of Cryptology, 2000.

[11] D. Hankerson, J. López Hernandez, and A. Menezes. "SOFTWARE IMPLEMENTATION OF ELLIPTIC CURVE CRYPTOGRAPHY OVER BINARY FIELDS," in Ç. Koç and C. Paar, editors, Second International Workshop on Cryptographic Hardware and Embedded Systems — CHES 2000, volume LNCS 1965, Berlin, 2000. Springer-Verlag.

[12] J. López and R. Dahab. "FAST MULTIPLICATION ON ELLIPTIC CURVES OVER GF(2^m) WITHOUT PRECOMPUTATION," in Jr. Ç. K. Koç and C. Paar, editors, Cryptographic Hardware and Embedded Systems — CHES 1999, volume LNCS 1717, pages 316 – 327. Springer-Verlag, 1999.

[13] R. Venugopalan and A. Dean. "IMPROVING ENERGY-EFFICIENCY IN SENSOR NETWORKS BY RAISING COMMUNICATION THROUGHPUT USING SOFTWARE THREAD INTEGRATION," www.cesr.ncsu.edu/agdean/ODES 04/ODES Venugopalan slides.pdf.

STMicroelectronics
Corporate Headquarters
39, Chemin du Champ-des-Filles
C.P. 21
CH-1228 Geneva
Plan-Les-Ouates - Switzerland

www.st.com