ICONS 2011

The Sixth International Conference on Systems

January 23-28, 2011 - St. Maarten,

The Netherlands Antilles

ICONS 2011 Editors

Leszek Koszalka, Wroclaw University of Technology, Poland

Daniela Dragomirescu, LAAS-CNRS / University of Toulouse, France

ICONS 2011

Foreword

The Sixth International Conference on Systems (ICONS 2011), held on January 23-27, 2011 in St. Maarten, The Netherlands Antilles, covered a broad spectrum of topics.

The conference covered fundamentals on designing, implementing, testing, validating and maintaining various kinds of software and hardware systems. Several tracks are proposed to treat the topics from theory to practice, in terms of methodologies, design, implementation, testing, use cases, tools, and lessons learnt. ICONS 2011 provided an international forum for discussions between researchers, practitioners and students interested in new developments targeting all the above areas.

We were very pleased to receive a large amount of top quality contributions. The accepted papers covered a wide range of networking related topics spanning Mobile Learning, safety requirements, privacy, complex systems, and embedded systems. We believe that the ICONS 2011 papers offered a large panel of solutions to key problems in all areas of systems and set challenging directions for industrial research and development

We take here the opportunity to thank all the members of the ICONS 2011 technical program committee as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also thank all the authors that dedicated much of their time and efforts to contribute to this ICONS 2011. We truly believe that thanks to all these efforts, the final conference program consisted of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations and sponsors. We are grateful to the members of the ICONS 2011 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

ICONS 2011 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in research on systems.

The beautiful places of St. Maarten surely provided a pleasant environment during the conference and we hope you had a chance to visit the surroundings.

ICONS 2011 Chairs

Raimund Ege, Northern Illinois University, USA
Hermann Kaindl, Vienna University of Technology, Austria
Leszek Koszalka, Wroclaw University of Technology, Poland
Daniela Dragomirescu, LAAS-CNRS / University of Toulouse, France

ICONS 2011

Committee

ICONS Advisory Chairs

Raimund Ege, Northern Illinois University, USA
Hermann Kaindl, Vienna University of Technology, Austria
Leszek Koszalka, Wroclaw University of Technology, Poland
Daniela Dragomirescu, LAAS-CNRS / University of Toulouse, France

ICONS 2011 Technical Program Committee

Giner Alor Hernández, Instituto Tecnológico de Orizaba - Veracruz, México
César Andrés Sánchez, Universidad Complutense de Madrid, Spain
Rafic "Ray" Bachnak, Texas A&M International University - Laredo, USA
Lubomir Bakule, Institute of Information Theory and Automation of the ASCR, Czech Republic
Nicolas Belanger, Eurocopter Group, France
Ateet Bhalla, NRI Institute of Information Science and Technology - Bhopal, India
Jun Bi, Tsingua University - Beijing, China
Mietek Brdys, University of Birmingham, UK
Keith Burnham, Coventry University, UK
Diletta Romana Cacciagrano, Università di Camerino, Italia
Mario Cannataro, University "Magna Græcia" of Catanzaro - Germaneto, Italy
Andrzej Chydzinski, Silesian University of Technology - Gliwice, Poland
Maria de las Mercedes Garcia Merayo, Universidad Complutense de Madrid, Spain
Yezyd Enrique Donoso Meisel, Universidad de los Andes - Bogotá, Colombia
Athanasios Doukas, University of Patras, Greece
António Dourado, University of Coimbra, Portugal
Daniela Dragomirescu, LAAS-CNRS / University of Toulouse, France
Raimund Ege, Northern Illinois University, USA
Daniel Federico Flueckinger, Scuola Universitaria Professionale della Svizzera Italiana, Switzerland
Miguel Franklin de Castro, Federal University of Ceará, Brazil
Eva Gescheidtová, Brno University of Technology, Czech Republic
Laurent George, University of Paris-Est Creteil Val de Marne, France
Vittoria Gianuzzi , University of Genova, Italy
Julien Henaut, LAAS-CNRS, France
Marko Jäntti, University of Eastern Finland, Finland
Heinz Jürgen Müller, Baden-Wuerttemberg Cooperative State University - Mannheim, Germany
Hermann Kaindl, Vienna University of Technology, Austria
Andrzej Kasprzak, Wroclaw University of Technology, Poland
Leszek Koszalka, Wroclaw University of Technology, Poland
Ondrej Krejcar, VSB - Technical University of Ostrava, Czech Republic
Radek Kuchta, Brno University of Technology, Czech Republic
Johannes Loinig, Graz University of Technology, Graz // NXP Semiconductors GmbH Austria, Gratkorn, Austria
Jia-Ning Luo (羅嘉寧), Ming Chuan University, Taiwan

Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
D. Manivannan, University of. Kentucky, UK
Fabrice Mourlin, Paris 12 University - Créteil, France
Namje Park, Arizona State University, USA
Marek Penhaker, VŠB - Technical University of Ostrava, Czech Republic
Pawel Podsiadlo, The University of Western Australia - Crawley, Australia
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland
Juha Röning, University of Oulu, Finland
Demetrios G. Sampson, University of Piraeus & CERTH, Greece
Rainer Schönbein, Fraunhofer IOSB, Germany
Florian Segor, Fraunhofer-Institut für Optronik - Karlsruhe, Germany
Vilém Srovnal, VŠB - Technical University of Ostrava, Czech Republic
Pavel Šteffan, Brno University of Technology, Czech Republic
Miroslav Sveda, Brno University of Technology, Czech Republic
Stanislav Tarasiewicz, Université Laval - Québec City, Canada
Denis Trcek, Univerza v Ljubljani, Slovenia
Elena Troubitsyna, Abo Akademi University, Finland
Theo Tryfonas, University of Bristol, UK
Dario Vieira, ENSIIE, France
Elisangela Vieira, Alcatel-Lucent Entreprise - Colombes, France
M. Howard Williams, Heriot-Watt University - Edinburgh, UK
Heinz-Dietrich Wuttke, Ilmenau University of Technology, Germany
Xiaodong Xu, Beijing University of Posts and Telecommunications, China
Yanyan Yang, University of Portsmouth, UK
Patrick Meumeu Yomsi, Université Libre de Bruxelles (ULB), Belgium
Soraya Zertal, University of Versailles, France

Copyright Information

Table of Contents

Implementation of MCU Invariant I2C Slave Driver Using Bit Banging

Arindam Halder, Ranjan Dasgupta
Innovation Lab, TATA Consultancy Services, Ltd.
Kolkata, India
arindam.halder@tcs.com,ranjan.dasgupta@tcs.com

Jayakar Chepada, Nrusingh Prasad Dash
Innovation Lab, TATA Consultancy Services, Ltd.
Kolkata, India
jayakar.ch@tcs.com,nrusingh.dash@tcs.com

Abstract—**The paper gives an overview of programming I2C slave device using bit banging method implemented in C programming language. In Microcontroller Unit (MCU) based tiny embedded system, with no built in universal serial communication (e.g., I2C, SPI, etc.) hardware engine support, bit banging method is the most efficient technique for handling any such communication. Finally, the paper states and demonstrates a solution of I2C communication between two devices as a case study**.

Keywords- I2C; Bit Banging; SPI; MCU; EEPROM.

I. Introduction

Philips Semiconductors had developed the I2C protocol over 20 years ago and has an extensive collection of specific usage across several general purpose devices. The I2C-bus supports two wires, serial data (SDA) and serial clock (SCL), it carries the information between the devices connected to the bus. Each device is recognized by a unique 7-bit address and can operate as either a transmitter or a receiver, depending on the function of the device [2], [6].

Sometimes, processors do not have built in hardware support for the universal serial communication. In such case, design your own code to implement serial communication, which is known as bit banging. For example, the MSP430F1232 MCU does not have any kind of built in hardware serial communication support. Therefore, to add any I2C serial device in a project based on this MCU, you have to create code to handle the communication. So, the challenge is to write bit banging I2C slave driver using C programming language. The Slave will be synchronized by the master clock, and the data portion will be driven by either the master or the slave. The synchronization part is taken care by the port pin interrupts of SCL and SDA. This implementation is briefly given in the case study and in the code implementation (Appendix). The advantage to design your own code is that, you can add the I2C serial communication to any microcontroller. The data transfer between any two microcontrollers can be achieved by using this code. A microcontroller can also use this code to communicate with any other devices, which are on the same single board (e.g., EEPROM, etc.).

II. Protocol specification

In I2C protocol, a master is the device which initiates a data transfer on the bus and generates the clock signals to permit that transfer. At that time, any device address is considered as a slave. During the I2C communication, unique situations arise, which are defined as START and STOP conditions (Figure 1). A HIGH to LOW transition on the SDA line, while SCL is HIGH, is one such unique case. This situation indicates a START condition. A LOW to HIGH transition on the SDA line, while SCL is HIGH, defines a STOP condition.

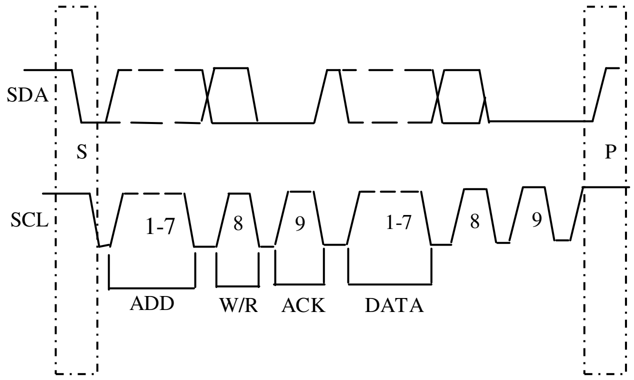


Figure 1. Data transfer on I2C bus

Every byte put on the SDA line must be 8-bit long. Each byte has to be followed by an acknowledgement bit. The data is transferred with the most significant bit (MSB) first. The transmitter releases the SDA line (HIGH) during the acknowledge clock pulse. The receiver must pull down the SDA line during the acknowledge clock pulse, so that, it remains stable LOW during the HIGH period of this clock pulse (Figure 1).

After START condition, a slave address is sent. In 7 bit long, followed by an eighth bit, which is a data direction bit (R/W) - a ‘zero’ indicates a transmission (WRITE), a ‘one’ indicates a request for data (READ) (Figures 2 and 3). After the address byte, the data to be read or write according to the data direction bit, are sent. This way, data transmission happens until a stop condition occurs.

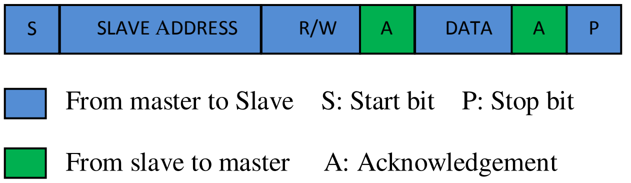

S
SLAVE ADDRESS
R/W
A
DATA
A
P
From master to Slave S: Start bit P: Stop bit
From slave to master A: Acknowledgement

Figure 2. A master – transmitter addressing a slave receiver with 7 bit address

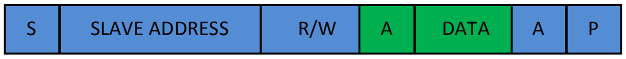


Figure 3. A master reads a slave immediately after the first byte

III. CASE STUDY

In our implementation, TI Davinci (DM6446) processor acts as an I2C master and MSP430 as an I2C slave device. The MCU I/O port pin P1.1 is configured as SCL and P2.1 as SDA. Both the pins are interrupt enabled. The master uses standard-mode of I2C communication, with its data transfer rate of up to 100 kbps and 7-bit addressing mode. To maintain the same bit rate in the slave side, we use Timer A of MSP430. Timer A has interrupt capabilities and it may be generated from the counter on overflow conditions.

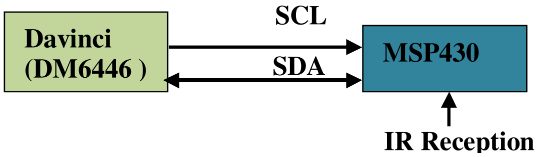


Figure 4. I2C bit banging implementation between Davinci processor and MSP430

IV. CODE IMPLEMENTATION

We have written the source code using C programming language. Most of the time, the assembly language code is more difficult to be understand, where as in C programming language it is much easier. But the most challenging thing is the performance. It should perform like the microcontroller has inbuilt hardware I2C engine. In source code, we call function init_i2c(i2c_read_byte). This function initializes the port pin for i2c slave and registers a callback function i2c_read_byte, which will process the data after receiving or before transmitting (see Appendix).

MSP430_SWI2CSV_init() calls a function named resetSWI2C. resetSWI2C function is used several times in the code to reset the I2C slave. The purpose of this function is to reinitialize the SCL and SDA pins. Here both the SCL and SDA are configured as input pin, where SCL will report an interruption on the transition from low to high, and SDA will report from high to low transition. This condition will be treated as a start condition of the I2C communication. The rest of the code will execute on the interrupt context.

As per the I2C specification, the master has the bus control over I2C bus and it is a synchronous protocol, so the clock (SCL) will be controlled by the master only, and the dataline (SDA) will be controlled by the master or the slave as per the requirement. Here SCL will be captured by the __interrupt void Port_1(void) interrupt function and SDA will be taken care by the __interrupt void Port_2(void) interrupt function (see Appendix).

SDA interrupt is used for detecting start and stop conditions of the I2C communication. This also detects repeated start condition.

There are 4 states considered for the slave device

1. SLAVE_ADDRESS_RECEIVE
2. SLAVE_NOTMY_ADDRESS
3. SLAVE_DATA_RECEIVE
4. SLAVE_DATA_TRANSMIT

Here, both the high to low transition and low to high transition on the SCL line are taken care. The initial state is SLAVE_ADDRESS_RECEIVE. Two counters, rising edge counter and falling edge counter are used to call specific function after specific number of bits received or transmitted. In low to high transition of the SCL line, PORT 2 (SDA) values are captured, enabling SDA interrupt to capture repeated start or stop condition from the master, and the SCL line are configured for high to low transition. In high to low transition of the SCL line, captured value is processed and the SDA interrupt is disabled.

IAR Embedded workbench IDE (Integrated Development Environment) is used to develop and compiles the code. MSP430 USB-Debug-Interface is used for porting the source code to the microcontroller.

V. CONCLUSION

The paper analyzes the simplicity concerns of writing I2C slave driver using bit banging method. An optimal and robust solution is implemented with a trade-off between speed and reliability. This generic approach can easily be adapted to any embedded device. Apart from MSP430F1232, other microcontrollers, like PIC16F5x or AT89C2051, also do not have universal serial communication interface [7], [8]. This generic code can also be used to achieve serial I2C communication just by changing the timer of specific MCU, to maintain the bit rate and interrupt driven PORT pins of the SCL and the SDA.

REFERENCES

[1] http://www.dwhoffman.com/bit_banging [retrieved: December 2010]

[2] http://www.nxp.com/acrobat_download2/literature/9398/39340011.pdf [retrieved: December 2010]

[3] http://focus.ti.com/mcu/docs/mcuprodtechdoc.tsp?sectionId=95&tabId=1204&familyId=911&techDoc=6&docCategoryId=6&viewType=mostrecent [retrieved: December 2010]

[4] http://processors.wiki.ti.com/index.php/TMS320DM6446 [retrieved: November 2010]

[5] http://focus.ti.com/lit/wp/spry136/spry136.pdf [retrieved: November 2010]

[6] http://ics.nxp.com/support/documents/interface/pdf/an10216.pdf [retrieved: December 2010]

[7] http://ww1.microchip.com/downloads/en/devicedoc/41213C.pdf [retrieved: December 2010]

[8] http://www.atmel.com/dyn/resources/prod_documents/doc0368.pdf [retrieved: December 2010]

APPENDIX

```
#define init_i2c(callback)              \
do {                                    \
  MSP430_SWI2CSV_init();                \
  regI2CCallBack(callback);             \
} while(0)

void resetSWI2C(){
    set_scl_output_low();
    set_sda_output_low();
    set_sda_input();
    set_scl_input();
    set_scl_rising_intr();
    set_sda_falling_intr();
    clear_scl_intr();
    enable_scl_intr();
    clear_sda_intr();
    enable_sda_intr();
    i2c_data=0;
}

#pragma vector=PORT1_VECTOR
__interrupt void Port_1(void)          // SCL
{
    if(P1IN&BIT0)      // Low to High (Rising Edge)
  {
    rising_edge_counter++;
    switch(i2c_flag)
    {
    case SLAVE_ADDRESS_RECEIVE:
    case SLAVE_NOTMY_ADDRESS:
    case SLAVE_DATA_RECEIVE:
       address_low_high1to9();
    break;
    default:           // SLAVE_DATA_TRANSMIT
        if(rising_edge_counter<= BYTE_LENGTH)
        transmit_low_high1to8();
        else
        address_low_high1to9 ();
    }
    if(rising_edge_counter==9)
      rising_edge_counter=0;
    }
  else
  {                  // High to Low (Falling Edge)
    falling_edge_counter++;
    switch(i2c_flag)
    {
      case SLAVE_ADDRESS_RECEIVE:
        if(falling_edge_counter<=SEVENTH_BIT)
          address_high_low1to7();
        else if(falling_edge_counter==BYTE_LENGTH)
          address_high_low8();
        else
          address_high_low9();
      break;
      case SLAVE_NOTMY_ADDRESS:
        if(falling_edge_counter<= BYTE_LENGTH)
          notmyaddress_high_low1to8();
        else
          notmyaddress_high_low9();
     break;
case SLAVE_DATA_RECEIVE:
      if(falling_edge_counter<= SEVENTH_BIT)
        receive_high_low1to7();
      else if(falling_edge_counter==BYTE_LENGTH)
        receive_high_low8();
      else
        receive_high_low9();
     break;
    default:       // SLAVE_DATA_TRANSMIT
     if(falling_edge_counter<=SEVENTH_BIT)
       transmit_high_low1to7();
     else if(falling_edge_counter==BYTE_LENGHT)
       transmit_high_low8();
     else
       transmit_high_low9();
    }
     if(falling_edge_counter==NINTH_BIT)
       falling_edge_counter=0;
    }
}

#pragma vector=PORT2_VECTOR
__interrupt void Port_2(void)            //    SDA
{
  if(SCL pin is low)                  //  SCL Low
      clear_sda_intr();            // Clear SDA interrupt flag
  else
  {
    if((P2IES&SDA))                 //  start
    {
     /*Initialize all counters*/
     /*Reload timers*/
     disable_sda_intr();          // Disable SDA interrupt
     if(i2c_flag==SLAVE_DATA_RECEIVE)
     {
        /*Set repeated start condition*/
      }
     i2c_flag=SLAVE_ADDRESS_RECEIVE;
     index=0;
    }
    else                         //  Stop
    {
     set_scl_input();                //  (set SCL as input)
     set_sda_input();                //  (set SDA as input)
     set_sda_falling_intr();      // SCL for 1->0 interrupt edge
```

```
      clear_sda_intr();             // Clear SCL interrupt flag
      set_scl_rising_intr(); // SCL for 0->1 interrupt edge
      clear_scl_intr();             // Clear SCL interrupt flag
      /*Clear timer*/
      i2c_flag=SLAVE_ADDRESS_RECEIVE;
    }
  }
}

void address_low_high1to9(){
   P2IES=(P2IN&BIT0);     // Set SDA interrupt edge
   set_scl_falling_intr();    // Set SCL for 1->0 interrupt edge
   clear_scl_intr();            // Clear SCL interrupt flag
   clear_sda_intr();           // Clear SDA interrupt flag
   enable_sda_intr();        // Enable SDA interrupt
 }

void address_high_low1to7(){
        set_scl_rising_intr();
        clear_scl_intr();
        if(repeat_start==1){
        /*all counters = 0, repeat_start =0*/
         }
        /*store the bit value*/
       /*Left shift carry bit into I2C val*/
       disable_sda_intr();
}

void notmyaddress_high_low9(){
        set_scl_output();
        set_sda_input();
        set_scl_rising_intr();
        clear_scl_intr();
        /*Set time out*/
        clear_scl_intr();
       i2c_flag=SLAVE_NOTMY_ADDRESS;
}

void  notmyaddress_high_low1to8(){
       set_scl_rising_intr();
       clear_scl_intr();
       disable_sda_intr();
}

void address_high_low9(){
        set_scl_output();
        if(i2c_data&0x01)  // Slave Transmit
       {
          i2c_flag=SLAVE_DATA_TRANSMIT;
          /*store the transmit data bitwise
          to SDA line*/
         /*Left shift transmit data*/
       }
       else{
            set_sda_input();
            i2c_flag= SLAVE_DATA_RECEIVE;
```

```
       }
       set_scl_rising_intr();
       clear_scl_intr();
     disable_sda_intr();
     set_scl_input();
}

void transmit_high_low9(){
       set_scl_output();
       if(No address match){
          i2c_flag = SLAVE_NOTMY_ADDRESS;
          set_scl_rising_intr();
          clear_scl_intr();
          disable_sda_intr();
          set_scl_input();
       }
       else{
        i2c_flag =  SLAVE_DATA_TRANSMIT;
        if(transmit bit value is 1)
           set_sda_output_high();
        else
           set_sda_output_low();
        set_sda_output();
        /*Left shift i2c val*/
        set_scl_rising_intr();
        clear_scl_intr();
        /*Reload timer*/
       disable_sda_intr();
       set_scl_input();
      }
}

void receive_high_low1to7(){
       set_scl_rising_intr();
       clear_scl_intr();
       /*store the bit value*/
      /*Left shift carry bit into I2C val*/
      disable_sda_intr();
}

void receive_high_low8(){
       set_sda_output_low();
       set_sda_output();
       set_scl_rising_intr();
       clear_scl_intr();
       /*store the bit value*/
       disable_sda_intr();
       set_scl_input();
}

void receive_high_low9(){
      set_scl_output();
      set_sda_input();
      set_scl_rising_intr();
      clear_scl_intr();
      disable_sda_intr();
      i2c_flag=SLAVE_DATA_RECEIVE;
      /*Stop Timer*/
      /*Call callback  function*/
     /*reset i2c data, and reload timer*/
     set_scl_input();
}

void transmit_low_high1to8(){
      /*Set SDA edge interrupt direction*/
      /*store SDA out data*/
      set_scl_falling_intr();
      clear_scl_intr();
      clear_sda_intr();
      enable_sda_intr();
}

void transmit_high_low8(){
    set_sda_input();
    set_scl_rising_intr();
    clear_scl_intr();
    disable_sda_intr();
}

void transmit_high_low1to7(){
       set_scl_rising_intr();
       clear_scl_intr();
       if(transmit bit value is 1)
          set_sda_output_high();
       else
          set_sda_output_low();
          set_sda_output();
          /*Left shift i2c val*/
          disable_sda_intr();
}
```

A Programmable Interconnection Network for Multiple Communication Patterns

Václav Dvořák, Jiří Jaroš
Dept.of Computer Systems
Faculty of Information Technology BUT,
Brno, Czech Republic
dvorak, jarosjir@fit.vutbr.cz

Abstract–**Application-specific or embedded systems with less than 16 processing cores are too small to use some kind of network on chip (NoC) for interconnection. On the other hand, a crossbar and related circuitry (arbiters, memory elements) are too expensive in terms of chip area. As only few pair-wise and collective communication patterns are mostly used in specific applications, we explore an interconnection network that can support only selected communication patterns and no others. The main contribution of the paper is designing of such networks without routers or arbiters, in a form of programmable combinational logic, with limited crossbar functionality. The interconnection network can be implemented by multiplexers or block RAMs on the FPGA chip at a very low cost. A functional decomposition of the related multiple-output Boolean function into a cascade of block RAM devices is aided by multi-terminal binary decision diagrams and is illustrated on examples.**

Keywords*– *multiprocessor SoCs; programmable interconnection; on-chip interconnects; crossbar switch; logic decomposition; multi-terminal BDDs

I. INTRODUCTION

Multiprocesor systems-on-chip (MPSoC) consist of multiple, usually heterogenous, processing elements (PEs) with local memory, and I/O components. They are usually targeted for embedded applications such as multimedia, telecommunication architecture, network security, and the like. In the implementation of MPSoC, an on-chip network comes to the forefront because of its impact on the performance of the system. Design of MPSoC relies at present mostly on point-to-point connections rather than on shared buses. Buses are not scalable beyond some limit and may not provide required performance because the available communication bandwidth is shared among all the units connected to the bus. Scalability and reusability were two features that led to the network on chip (NoC) paradigm for on-chip communication [1].

The topologies of choice for NoCs have been ring, mesh, fat tree, crossbar and spidergon [2]. Performance of these networks in pair-wise as well as in collective communications is well understood. The lower bounds for time complexity of collective communications are known and can be reached in some cases by optimum scheduling of communications [3].

Communication operations can be either point-to-point, with one source and one destination, or collective, with more than two participating processes. Collective communications (CCs) are invoked by nodes to distribute, gather, and exchange data. Some embedded parallel applications, like network or media processors, are characterized by independent data streams or by a small amount of inter-process communications [1]. However, many special-purpose parallel applications display a bulk synchronous processing (BSP) behavior: the processing nodes access the network according to a global, structured communication pattern.

A collective operation is usually defined in terms of a group of processes. The operation is executed when all processes in the group call the communication routine with matching parameters. We classify collective operations into three types according to their purpose: CCs (One-to-All, OA, All-to-One, AO, All-to-All, AA), global computation (reduction AOR or AAR and scan) and synchronization (barrier). The CCs are most important, as other collective operations are closely related to them. In a broadcast (OAB), one process sends the same message to every group member, whereas in a scatter (OAS), one process sends a different message to each member. Gather (AOG) is the dual operation of scatter, in that one process receives a message from each group member. These basic operations can be combined to form more complex operations. In all-to-all broadcast (AAB), every process sends a message to every other group member. In complete exchange, also referred to as all-to-all scatter-gather (AAS), every group member sends a different message to every other group member. Permutations, and partial permutations (i.e., permutations in which some source to destination pairs are missing) are important CCs that can be used as building blocks to create more complex all-to-all CCs. Since complexities of some communications are similar (AOG ~ OAS, AOR ~ OAB, AAR ~ AAB), we will focus only on 4 basic types (OAB, OAS, AAB, AAS).

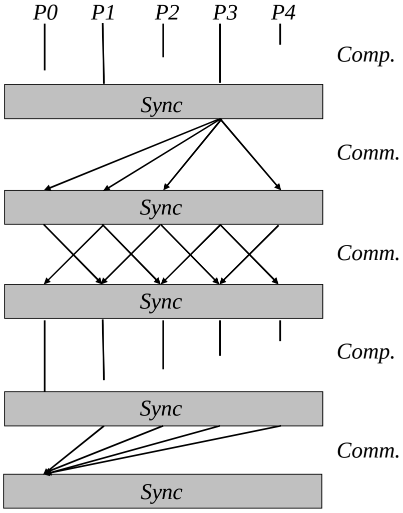


Figure 1. BSP algorithm with five supersteps executed on five processors.

The distributed-memory BSP model proposed by Valiant is essential to our discussion. It comprises computer architecture, a class of algorithms, and a function for charging costs to algorithms [6]. A BSP computer consists of a collection of processors, each with private memory, and a communication network that allows processors to access memories of other processors. A BSP algorithm consists of a sequence of super-steps, which contain a number of either computation steps or communication steps, followed by the global barrier synchronization. In a computation super-step, each processor performs a sequence of operations on local data. In a communication super-step, each processor sends and receives a number of messages (Figure1).

The time, or cost, of a communication super-step is

$$T_{comm} = hg + l, \qquad (1)$$

where each processor sends/receives to/from other processors h or less data words, g is the time needed to transfer one data word under the continuous message traffic, and l is a fixed overhead (latency) of communication and of global synchronization.

Further on we will assume that application-specific systems of interest in this paper are compatible with the BSP model described above. In the following Section II, we map the state of the art in application-specific interconnects and formulate the problem to be dealt with. The main results are in Sections III. We first consider complexity of different implementations of a crossbar network and arrive at its low-cost, application-specific version in Section IIIA, and then present logic design of a programmable interconnect on a small example in Section IIIB. Synthesis of the application-specific NoC for multiple communication patterns is done on a small set of patterns in Section IV. The presented technique is assessed in Conclusions.

II. STATE OF THE ART AND THE PROBLEM STATEMENT

In the context of application-specific MPSoCs, communication architecture, more often than not does not have to support *all* pair-wise communications and efficient implementation of *all* collective communications such as broadcast, multicast, gather, scatter, and others. Quite a few applications running on MPSoCs use only a limited set of deterministic communication patterns and a general NoC infrastructure suitable for general-purpose computing is not needed. By taking advantage of the known application communication behavior, special-purpose networks may be designed for well-behaved communication requirements, resulting in networks that are more resource/ performance effective. We will therefore try to simplify the interconnection network by providing a support just for the required communication patterns.

The application-specific optimization of interconnection network with respect to performance and power consumption can be obtained by removal of some links. For example, the spidergon topology is opened for such optimization [4]. A design methodology using a recursive bisection technique for generating optimum topology for applications with well-behaved communication patterns has been introduced in [5]. In contrast, our approach treats the interconnection as combinational logic and thus covers not only the single instance of a problem, but a wide class of applications with limited number of CPU cores and with limited communication needs. The starting point is a non-blocking crossbar, which is the ideal on-chip interconnect, but it is prohibitively expensive for larger MPSoCs.

Beside static networks, it may also be of interest to allow the network to be reconfigurable, at run-time. For example, reconfigurable computing paradigms (e.g., FPGAs) have increasingly become more practical alternatives recently. Field-programmable interconnect devices (FPIDs), acting as SRAM-based switching matrices, can be reconfigured dynamically, in the same way as standard SRAM-based FPGAs – by means of SRAM cells controlling switching elements (pass transistors or transmission gates). Reconfiguration times are therefore much too long and prevent dynamic reconfiguration before each communication pattern.

The problem to be addressed in this paper is to find as simple message-passing communication structure as possible for MPSoC with a small number (≤ 16) of PEs and with a set of deterministic communication patterns customized for a specific application. Unidirectional, one bit wide links will be assumed.

III. Main Results

We will focus on interconnect programmable in run-time, with performance comparable to a crossbar. For the fastest operation, the control of a programmable intercon-nection network must come from inside the chip, e.g. from the master PE. Such in-system reconfiguration can achieve some of the flexibility of software with the performance of dedicated hardware. This idea is elaborated further on in two following sub-sections.

A. Crossbar Implementation

For multiprocessor SoCs of smaller size, a crossbar (Xbar) switch is utilized for an efficient on-chip network solution, Figure 2. Arbiters provided for each X-bar output select one request from all coming in and set the switch to the appropriate position. There is a CAD tool for generation of round-robin arbitration and $N \times M$ X-bar switch logic for MPSoCs [7] based on user specifications. To evaluate the hardware complexity, two possible implementations have to be considered.

A crosspoint-based implementation of a square $N \times N$ crossbar makes use of N columns of N crosspoints realized by transmission gates, pass transistors or tri-state inverters. A crosspoint-based crossbar implies a memory element per crosspoint (N^2 elements altogether) that lets a row signal be propagated or not to some column. Contents of memory elements are determined by (round-robin) arbiters, one per column.

A second approach to implement the crossbars is to use logic multiplexers. A full $N \times N$ crossbar requires N N-input multiplexers and N arbiters (one per each multiplexer). As $\log_2 N$ memory elements are required for each column multiplexer, a total of $N \log_2 N$ configuration memory elements are needed. This approach thus leads to a significant reduction in the number of memory elements.

As regards a support for pair-wise and collective communication, the X-bar in Figure 2 is able to implement any permutation of inputs or its subset, broadcast from any node, multicast, and also several non-conflicting parallel multicasts. Whereas a source PE can send up messages to all piers in parallel, the target can receive only a single message. Multiple messages targeted for a single PE are filtered out by arbiters.

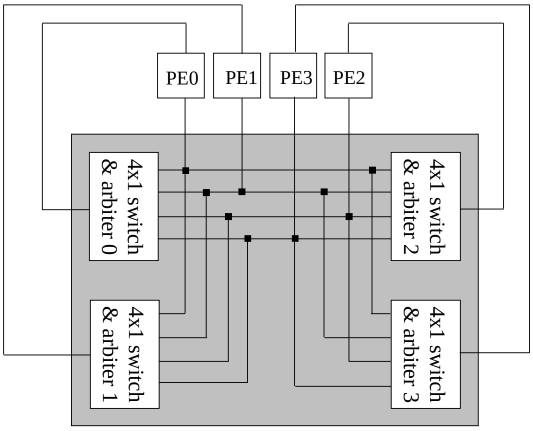


Figure 2. The example of 4×4 X-bar with four processors

The above limitation determines the number of communication steps needed for typical collective communications. Broadcast and multicasts need only a single step, gather (and reduction operations) require $\log_2 N$ communication steps. All-to-all communica-tions such as broadcast or scatter are implemented as a sequence of N-1 permutations.

If only a specific set of pair-wise and collective communications is needed in a certain BSP application, we can simplify the X-bar in Figure 2 a great deal. We can get rid of arbiters entirely and use switches of size $P \times 1$, where P is the number of communication patterns. All switches have identical control, so that the number of control signals is typically much lower than that in ordinary $N \times N$ crossbars,

$$\log_2 P < N \log_2 N. \qquad (2)$$

Multiplexers with 16 to 32 data inputs are quite common and the number of required communication patterns may nicely fit into this range.

B. Logic Design of a Programmable Interconnect

The programmable interconnect can be implemented as a network of programmable switches. The switch itself is a logic device that can connect some or all inputs, one-to-one, to some or all outputs. Multiplexer and crossbar switches are building blocks of more complex programmable interconnection networks. E.g., an elementary 2 × 2 crossbar is used in multi-stage interconnection networks that are much cheaper than crossbars.

We will illustrate logic design of programmable interconnect in more detail on a class of programmable Bit-Masking and Shifting interconnect devices with 4 to 8 inputs (BMS4 to BMS8). These units are useful when implementing multi-way branching in micro-programmed controllers; they enable efficient allocation of microcode memory to a cluster of multi-way dispatch tables [8]. For example, the task of the 16-way BMS4 is to shift 4 or less active inputs, selected by a 4-bit mask, to the lowest positions of the 4-bit output vector and reset the rest of outputs. The output vector then serves as an offset from the base address of a dispatch table; this way the dispatch tables of various size can be stored in control memory in a compact form.

Cube specification of BMS units has been generated automatically. For example, the BMS4 function was specified by 81 cubes:

```
.i 8
.o 4
.ilb m0 m1 m2 m3 x0 x1 x2 x3
.ob y0 y1 y2 y3
.type fr
.p 81
0000---- 0000
0001---0 0000
0001---1 1000
0010--0- 0000
0010--1- 1000
   …
   …
11111011 1011
11110111 0111
11111111 1111
.e
```

Synthesis of the related combinational logic for the 4-input BMS can be done in several ways. The simplest solution would be a single 256 × 4-bit look-up tables LUTs (ROM or Block RAM). However, if only smaller LUTs were available, we can decompose the single LUT into a cascade of two or more smaller LUTs. The latency will increase, but with possible pipeline operation the throughput will remain the same. Various decompositions are easily found from a Multi-Terminal BDD (MTBDD) representing the BMS4 function. The optimal variable ordering of the MTBDD can be found e.g., by the Heuristic Iterative Decomposition Tool HIDET [9] and is shown in Figure 3. The suitable cut of the MTBDD generates a cascade of two LUTs (Figure 4a) with the resulting capacity in bits less than a half of the single LUT capacity.

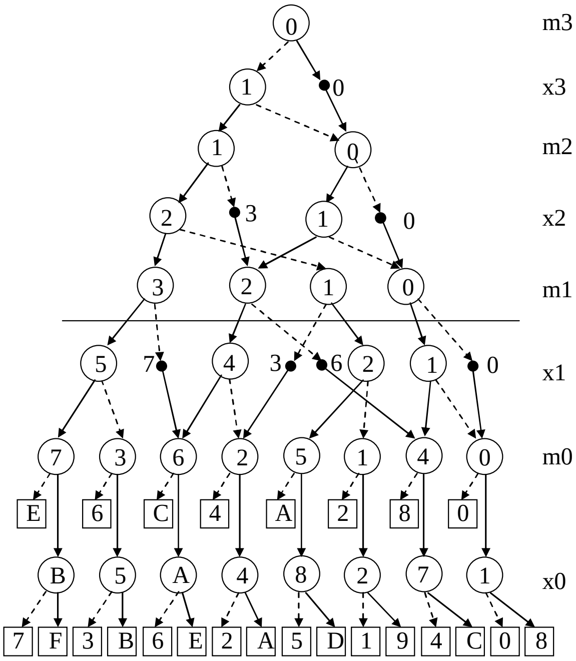


Figure 3. MTBDD of the 4-input BMS.

The traditional design of BMS module in a form of multiplexer network is shown in Figure 4b. Comparing both designs, the LUT cascade wins in the area size devoted to interconnections and in flexibility to implement other communication patterns. As for the performance, delays few ns per stage have been demonstrated for an experimental LUT cascade [10]. Note that the ordering of variables found by HIDET is the same as the optimal ordering of traditional design in Figure 4b.

Parameters of logic design obtained by Xilinx FPGA synthesis tool for BMSs with 4 to 8 inputs and parameters of MTBDDs obtained by the HIDET are shown in Table 1. The local LUT cascade width *x* relates logarithmically to the local values of MTBDD width *w* ($x = \lceil \log_2 w \rceil$) between neighbor LUTs. BMS units could be implemented as a cascade of LUTs eliminating messy wiring and reducing chip area for the interconnect. The delay of such switch-boxes is adjustable by the cascade length. If we take the delay of FPGA’s 4-input LUTs plus wiring delay approximately equal to cascaded LUTs´ delay, we should use not more LUTs in a cascade than it is given in FPGA column “levels” for the same or better performance. Note that communication from PE’s outputs to PE’s inputs is now supported by regular wiring from addresses to data outputs of multi-bit memory modules and by external regular wiring among these modules.

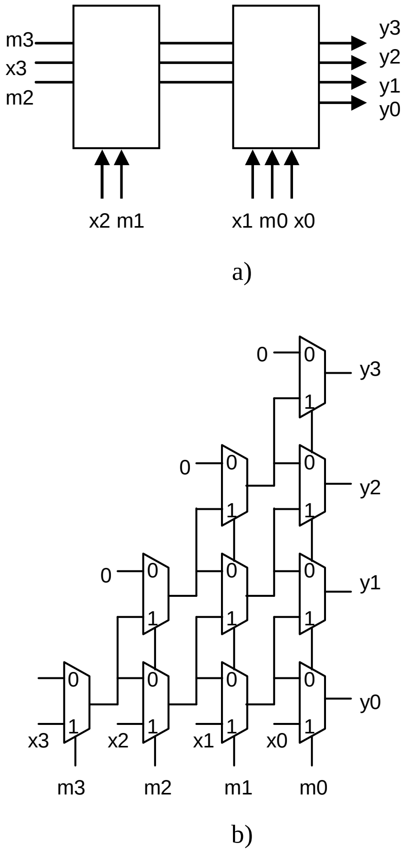


Figure 4. Implementation of the 4-input BMS a) as the 2-LUT cascade b) as a MPX network

TABLE I. PARAMETERS OF FPGA AND MTBDD-BASED DESIGNS

	FPGA		MTBDD	
	#4-LUTs	levels	cost	width
BMS4	8	3	30	≤ 4
BMS6	20	5	126	≤ 6
BMS7	32	6	254	≤ 7
BMS8	80	7	447	≤ 8

IV. SYNTHESIS OF APPLICATION-SPECIFIC NOCS FOR MULTIPLE COMMUNICATION PATTERNS

The $N \times N$ multiplexer-based crossbar network without arbiters can also be visualized as a combinational logic network with $N + N\log_2 N$ inputs and N outputs. E.g. for $N = 8$ we get 32 inputs and 8 outputs. Out of 2^{24} programmable configurations, only a negligible fraction may be utilized in a certain MPSoC. The multiplexer-based crossbar as an interconnection network is thus for many MPSoC a luxury. A simpler way how to implement a low cost programmable unidirectional interconnect is to replace individual control of M crossbar multiplexers by identical control. Apparently, the number of multiplexer inputs P will now be determined by the number of required communication patterns. For example, barrel shifters can be implemented as crosspoint-based crossbars with common diagonal control [11] or as multiplexer-based crossbars with one multiplexer per output and a common control [12].

Let us design for illustration the application-specific NoC connecting 8 PEs (labeled 0 to 7) and supporting the following 7 communication patterns (encoded in 3 configuration bits):

1. Broadcast from node 0
2. Cyclic shift from node i to node (i+1)mod 8
3. Cyclic shift from node i to node (i+2)mod 8
4. Skew (0↔7, 1↔6, 2↔5, 3↔4)
5. Gather1 (7→6, 5→4, 3→ 2, 1→ 0)
6. Gather2 (6→4, 2→ 0)
7. Gather3 (4→ 0).

A crossbar-like implementation of the above set of communications without arbiters would require eight 8-input wide multiplexers and 8 arbiters. The same network can also be implemented as a regular ROM (multi-bit LUT) with 8+3 = 11 address bits and 8-bit wide output (single Xilinx BRAM block 2048 × 8 bit). The interconnection network is thus embedded in the regular ROM structure. If 2048 × 8 bit is too large an array, and a higher latency is tolerable, we can split it into 2 or more ROMs in a cascade by means of splitting the related MTBDD; one such decomposition is shown in Figure 5.

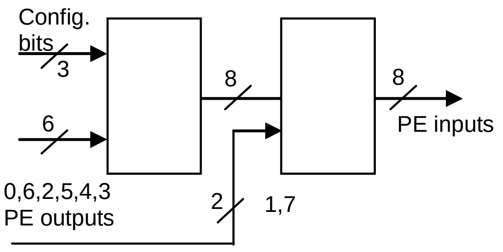


Figure 5. Implementation of the specific interconnection network as the cascade of two multi-bit LUTs

Similar cascade decompositions will be handy especially for larger problems (for example, 16 PEs, 32 patterns, that is 21 inputs). The communication bandwidth for larger messages will not be decreased when we use pipelining. The size of partial LUTs and their interconnection are again minimized by the HIDET [9] as in the previous section.

VI. CONCLUSIONS

Application-specific multiprocessor SoCs with a restricted set of deterministic communication patterns and not more than some 16 processing elements can operate with a low cost customized communication network synthesized as a logic subsystem. We assume that applications running on the system are well behaved (i.e., have similar communication patterns on every run) and data independent (i.e., have similar communication patterns for any data set). Contrary to a full crossbar network, $N \times 1$ switches (multiplexers) share control inputs and are not controlled by arbiters, but by one selected PE, for example by the first PE reaching the barrier before a communication step. Thus the arbiters may be eliminated completely. The width of multiplexers is given by the number of communication patterns in the application: one pattern is assigned to one multiplexer input. Instead of reconfiguring topology for a certain application, the communication module is programmed for communication patterns repeatedly in the runtime.

Beside multiplexers, other devices can be used to implement the specialized interconnection network. BRAM devices or cascaded BRAMs are another option. The kind of programmable interconnect suggested in the paper is run-time programmable much faster than FPGAs or FPIDs, because programming is reduced to processing a single store (output) machine instruction to a pattern holding register. Freedom from contention is an additional favorable side-effect of the presented approach.

Performance of the suggested communication module should be comparable to or better than crossbar performance, because of absence of arbitration logic. Prediction of overall overhead of a BSP algorithm is possible by means of (1). Other components of communication architecture like communication protocols or interface design between PEs and a communication module will be a subject of future research.

ACKNOWLEDGMENT

This research has been carried out under the financial support of the research grants GP103/10/1517 “Natural Computing on Unconventional Platforms”, GA 102/09/H042 "Mathematical and Engineering Approaches to Developing Reliable and Secure Concurrent and Distributed Computer Systems", GA102/08/1429 “Safety and security of networked embedded system applications”, MSM 0021630528 “Security-Oriented Research in Information Technology”, and FIT-S-10-1 "Secured, reliable and adaptive computer systems", Brno University of Technology.

REFERENCES

[1] Jantsch, A. and Tenhunen, H. Networks on Chip, Kluwer Academic Publ., Boston, 2003.

[2] Karim, F. and Nguyen, A.: An Interconnect Architecture for Networking Systems on Chips. IEEE Micro, 2002, pp.36-45.

[3] Jaroš J. and Dvořák V.: Evolutionary-Based Conflict-Free Scheduling of Collective Communications on Spidergon NoCs, In: Proceedings of 2010 Genetic and Evolutionary Computation Conference, GECCO 2010, New York, US, ACM, 2010, pp. 1171-1178.

[4] Palermo, G. et al.: Application-Specific Topology Design Customization for STNoC. Proc. of the 10th Euromicro Conference on Digital System Design Architectures, Methods and Tools, DSD 2007, pp.547-550.

[5] Ho, W.H. and Pinkston, T.M.: A Methodology for Designing Efficient On-Chip Interconnects on Well-Behaved Communication Patterns. Proc. Of the 9th Int. Symposium on High Performance Computer Architecture, Anaheim, 2003, pp. 377-388.

[6] Bisseling, R.H.: Parallel Scientific Computation. Oxford Univ. Press, New York, 2004.

[7] Shin, E.S.: Automated generation of round robin arbitration and crossbar switch logic. Ph.D. thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, November 2003.

[8] Dvořák, V.: LUT Cascade-Based Architectures for High Productivity Embedded Systems, In: International Review on Computers and Software, Vol. 2, No 4, Naples, Italy, pp. 357-365, 2007.

[9] Mikušek P. and Dvořák V.: On Lookup Table Cascade-Based Realizations of Arbiters, In: 11th EUROMICRO Conference on Digital System Design DSD 2008, Parma, IT, IEEE CS, 2008, pp. 795-802.

[10] K. Nakamura, T. Sasao, M. Matsuura, K. Tanaka, K. Yoshizumi, H. Qin, and Y. Iguchi, "Programmable logic device with an 8-stage cascade of 64K-bit asynchronous SRAMs," Cool Chips VIII, IEEE Symposium on Low-Power and High-Speed Chips, April 20-22, 2005, Yokohama, Japan.

[11] Asano, D. K.: Computer Architecture, Shift Circuits. 2001. http://www-comm.cs.shinshu-u.ac.jp/public/comparch/node45.html

[12] Gigliotti, P.: Implementing Barrel Shifters Using Multipliers. Xilinx Application note, XAPP195 (v1.1), 2004.

A Cyber-Physical System Design Approach

Miroslav Sveda
Faculty of Information Technology
Brno University of Technology
Brno, Czech Republic
e-mail: sveda@fit.vutbr.cz

Radimir Vrba
Faculty of Electrical Engineering and Communication
Brno University of Technology
Brno, Czech Republic
e-mail: vrbar@feec.vutbr.cz

***Abstract*— This paper exemplifies principles of cyber-physical systems design using original smart data acquisition systems capable to store and present measured data wirelessly. The presented temperature data logger stands for an example of flexible, mobile and intelligent appliances fitting various industrial or medical applications. Similarly, the discussed sensor network represents a system architecture stemming from wireless smart pressure sensors connected by Bluetooth and from a network concentrator, which is based either on PDA personal digital assistant or on GSM SmartPhone. Two pilot software implementations were developed for IPAQ PDA 5450 and Nokia 3650 SmartPhone. The paper describes a cyber-physical system design approach using novel data acquisition systems, which can serve as components of public or technological process monitoring systems and allow collecting data also from locations difficult to reach, e.g., from sensors located on rotating parts.**

Keywords- Embedded system design, smart sensor, wireless communication, temperature and pressure measurement.

I. Introduction

The charter for the CPS (Cyber-Physical System) Summit in April 2008 [6] introduced the following: “The integration of physical systems and processes with networked computing has led to the emergence of a new generation of engineered systems: cyber-physical systems. Such systems use computations and communication deeply embedded in and interacting with physical processes to add new capabilities to physical systems. These cyber-physical systems range from miniscule (pacemakers) to large-scale (the national power grid). Because computer-augmented devices are everywhere, they are a huge source of economic leverage. … it is a profound revolution that turns entire industrial sectors into producers of cyber-physical systems. This is not about adding computing and communication equipment to conventional products where both sides maintain separate identities. This is about merging computing and networking with physical systems to create new revolutionary science, technical capabilities and products.” Embedded computers allow designers to add capabilities to physical systems that they could not feasibly add in any other way. By merging computing and communication with physical processes and mediating the way we interact with the physical world, cyber-physical systems bring many benefits: they make systems safer and more efficient; they reduce the cost of building and operating these systems; and they allow individual machines to work together to form complex systems that provide new capabilities.

The kernel of this paper consists of a general part, covered by Section II, of case studies presenting the real-world applications discussed in Sections III and IV, and of conclusions. The general part is devoted to the review of state of the art in frame of CPS design. The first introduced application discusses a new mobile temperature data logger with RFID (Radio Frequency Identification) capabilities. The second application deals with a developed sensor network that embodies the pressure sensing system consisting of wireless smart pressure sensors connected by the Bluetooth, and of a network concentrator, which is based either on PDA personal digital assistant or on GSM SmartPhone.

II. Cyber-Physical Systems Desicn

Many of the embedded systems-related studies and efforts in the past have focused on the challenges the physical environment brings to the scientific foundations of networking and information technology, see [2] and [4]. However, the full scope of the change enabled by introducing CPS as a new branch of science and technology provides much more than restructuring inside this domain. The new approach can turn entire industrial sectors into producers of CPS. Actually, CPS is about merging computing and networking with physical systems to create new capabilities and improve product quality [9].

Cyber-physical systems denote a new modeling paradigm that promotes a holistic view on real-world – and therefore complex – systems. These systems have been studied before from various particular perspectives using paradigms like ubiquitous and distributed computing or embedded and hybrid systems. The above mentioned facts require also another approach to the design of such systems respecting from the beginning of design process the application domain that influences quality-of-service requirements such as real-time behavior, safety and security [12], [13] and [14], but also precision, reliability and other non-functional properties and contentment affecting attributes specified usually by official standards [8].

In a CPS application, the function of a computation is defined by its effect on the physical world, which is in this case not only a system environment, but evidently also a component of the designed application system. Therefore, proper design environments should be used to improve or at

ISBN:978-1-61208-114-4

least to enable efficiency of the design process. In cyber-physical systems the passage of time becomes a central feature — in fact, it is this key constraint that distinguishes these systems from distributed computing in general. Time is central to predicting, measuring, and controlling properties of the physical world: given a (deterministic) physical model, the initial state, the inputs, and the amount of time elapsed, one can compute the current state of the plant. This principle provides the foundations of control theory. However, for current mainstream programming paradigms, given the source code, the program's initial state, and the amount of time elapsed, we cannot reliably predict future program state. When that program is integrated into a system with physical dynamics, this makes principled design of the entire system difficult. Instead, engineers are stuck with a prototype-and-test style of design, which leads to brittle systems that do not easily evolve to handle small changes in operating conditions and hardware platforms. Moreover, the disparity between the dynamics of the physical plant and the program seeking to control it potentially leads to errors, some of which can be catastrophic.

III. Wireless temperature data logger

Knowledge of temperature course during a certain time is needed in scientific, medical and industrial applications. In some applications, however, the recorded temperature course should be read wirelessly. This section describes main principles applied in a set of mobile temperature data logger and portable reader/writer with wireless transfer of digitized temperature values. The pilot concepts of the system were originally introduced in [7].

A. *Application basics*

A mobile temperature data logger with RFID features was designed for applications, where portability and wireless data transfer is inevitable. Communicating reader/writer can be mounted on the wall or can also be portable.

Two main modes of operation during temperature data logging may be remotely chosen for a tag:

- Mode 1 - normal data collecting method in preprogrammed regular acquisition time intervals (100 milliseconds up to 2 hours) with number of samples limited only by a data EEPROM memory size.
- Mode 2 - more memory size reducing method, when only breaking lower and upper temperature limits initiates storing the date and time stamp.

In mode 2, the following date and time stamp is stored only when the temperature returns back into the temperature band between lower and upper temperature predefined limits. Also enhanced mode can be set when the maximum or minimum temperature between breaking and returning points of a sampling temperature course is stored, too. This is a typical example for monitoring of food transport, where the time stamp and maximum temperature after breaking the limit help to identify that offender who damaged the transported goods.

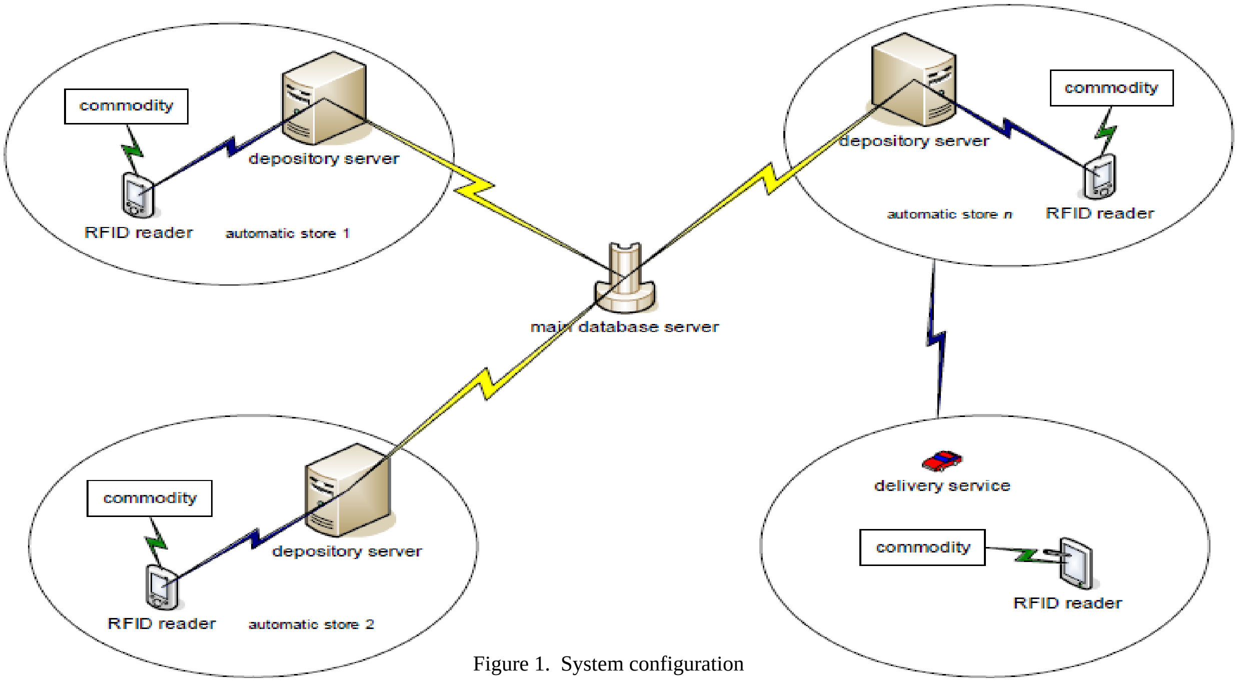


Figure 1. System configuration

Wireless RFID systems generate and radiate electromagnetic waves. This is the reason why those systems are legally classified as radio systems. The function of other radio services must under no circumstances be disrupted or impaired by the operation of RFID systems. For this reason, it is usually only possible to use frequency ranges that have been dedicated specifically for industrial, scientific or medical applications. These are the frequencies classified for worldwide as ISM (Industrial – Scientific – Medical), and they can also be used for RFID applications. The most important frequency ranges for RFID systems are therefore 135 kHz, 27.125 MHz, 40.68 MHz, 433.92 MHz, 869.0 MHz, 915.0 MHz, 2.45 GHz, 5.8 GHz and 24.125 GHz.

B. *Application constraints*

The range below 135 kHz is heavily used by other radio services because it has not been reserved as an ISM frequency range. The propagation conditions in this long wave frequency range permit the radio services that occupy this range to reach wide areas at a low technical cost. In order to prevent collisions, the future Licensing Act for Inductive Radio Systems in Europe, 220 ZV 122, will define a protected zone of between 70 and 119 kHz, which will no longer be allocated to RFID systems. The main block diagram of designed tag and reader/writer system is shown in Fig. 1.

Preferences for frequency range below 135 kHz allow reaching large ranges with low cost transponders. High level of power is available to the transponder. The transponder has low power consumption due to its lower clock frequency and often sleeping a standby mode of operation. Miniaturized transponder formats can be achieved due to the use of ferrite coils in transponder. Low absorption rate or high penetration depth in nonmetallic materials and water are available due to lower frequencies. Basic block diagram is shown in Fig. 2.

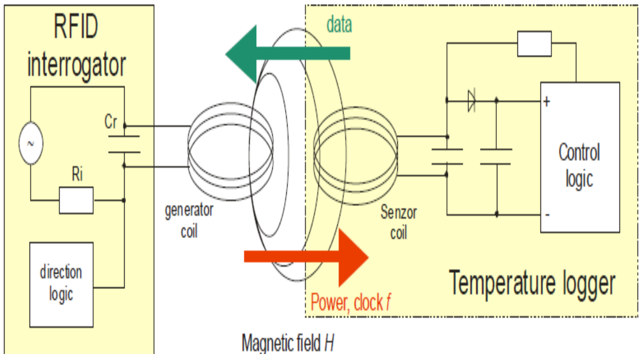


Figure 2. Basic block diagram

An inductively coupled data logger comprises an electronic data-logging device and a large area coil that functions as an antenna. Inductively coupled logger is almost always operated in passive mode of data transmitting between an RFID tag and reader. This means that all the energy needed for the data transfer to the temperature logger has been provided by the reader. For this purpose, the reader’s antenna coil generates a strong, high frequency electromagnetic field, which penetrates the cross-section of the coil area and the area around the coil.

C. *Data-logger design*

Temperature is recorded using a temperature tag at user defined time intervals. The temperature tag can be programmed so that when the memory is full it either stops further recording or continues recording by overwriting the earliest of the previously recorded data. Typical stored information contains: date and time stamp, temperature, temperature tag unique ID. Recorded information can be transferred to a reader/writer and then to a PC or a PDA with wired or wireless connection to a reader/writer. Temperature can be displayed graphically and the zoom functions allow focus on time periods where the temperature exceeds parameters.

The tag is a self-powered facility working like a wireless temperature sensor. It consists of a temperature probe and an active part with active RFID technology powered by an internal battery. The tag transmits an RF signal on demand at a pre-set time-interval. Tag life is estimated at several months depending on pre-set period of a transmission, where the related transmission interval can be configured via wireless connection by a reader/writer node. Each lifespan of the tag ends when the battery life is exhausted. Battery status can be inferred by interrogating the tag’s internal status value. The lifespan of the tag can be increased by delayed switch on by the first communication attempt of a reader/writer. A portion of this tag is used to measure actual temperature, to store measured data and to implement real time clock for timing. Distinct temperature values are acquired in pre-set intervals. The discussed tag can work in two basic modes as Data Logger and Out-of-Limit-Values Logger:

- Data Logger - Temperature is measured in pre-set intervals. All data is stored into the internal memory. In this mode there is stored only a start time. Number of measurements depends on memory size. Data logging principle is shown in Fig. 3.
- Out-of-Limit-Values Logger - If temperature is out of range, time stamp and temperature data are stored into the internal memory. Needless to say that the number of stored values depends on memory size.

This temperature tag is designed for usage in shipping containers, dairy industry, medical applications, fuel industry, refrigerated loads, agricultural industry, refrigeration monitoring, dangerous goods areas or anywhere, when temperature monitoring is required. Of course, the collection of possible applications is currently increasing for the reason that temperature is a critical process and quality assurance factor for many industries.

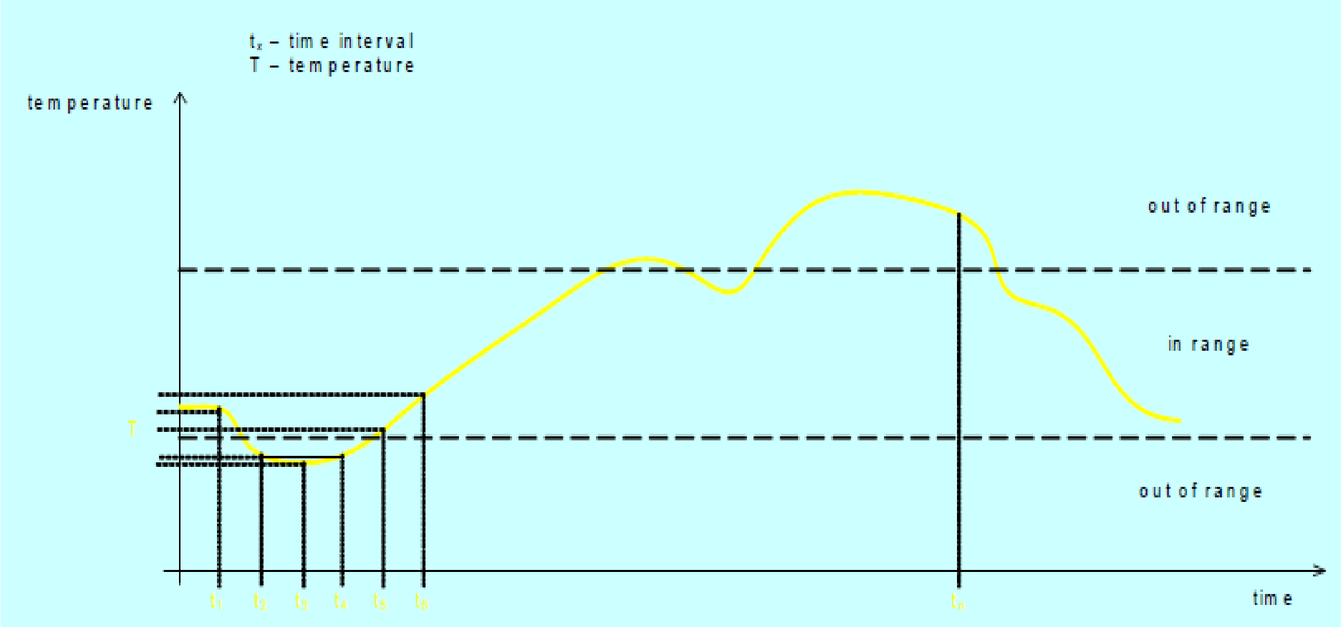


Figure 3. Data logging principle

IV. WIRELESS SMART SENSOR NETWORK

Many industrial systems cannot operate on cable connections among distributed components, which can be geographically remote, temporarily installed, allocated on rotating elements, etc. In such circumstances it is sometimes possible to connect smart sensors and actuators via RF wireless physical layer on ISM (Industrial, Scientific, and Medicine) band. For small range network with perimeter up to 100m the wireless connection standard Bluetooth can be used. This standard also enables to assure high safety and security of data by encryptions and related crypto-graphical techniques. The current section, which stems from partial results published earlier [15], proposes not only how to represent a system’s solution as an application example, but also how to generalize reached research results.

The described novel wireless smart sensor network system represents the system architecture formed by a set of the wireless smart pressure IEEE 1451 sensors. This original prototype is connected using a Bluetooth and a network concentrator/controller is based on (see Fig. 4):

- A PDA personal digital assistant (IPAQ PDA 5450 in this case study) or
- A GSM SmartPhone (Nokia 3650 SmartPhone in this case study).

Particular software applications are developed for both IPAQ PDA 5450 and Nokia 3650 SmartPhone. Those SW applications fit system requirements for data monitoring, parameter setting and exploiting all embedded functions of the novel smart pressure sensor with Bluetooth wireless interface in the slave role mode. Another SW application for standard PC enables to process data in MS Office formats. Such wireless system virtual port enables to configure the related sensor, to sense and to record data by remote reading and/or writing via a Bluetooth SPP supported port by a PC, a PDA or a SmartPhone equipped with a Bluetooth standard access module.

A. Application basics

Many industrial systems may not use cable connection between components and blocks of the system. Components are geographically isolated, temporarily installed or allocated on rotating elements etc. However, it is appropriate in such case to realize interconnection between components by wireless communication network. The conclusions enable to connect smart sensors and actuators via RF wireless physical layer on ISM (Industrial, Scientific, and Medicine) band. For small range network with perimeter up to 100 m an industrial Bluetooth wireless connection standard can be used, where full Bluetooth stack and basic necessary functions have been defined.

This standard also accepts requirements to assure high safety and security of data by using encryptions and cryptography. Bluetooth stack and defined protocols are also compliant to Internet protocols.

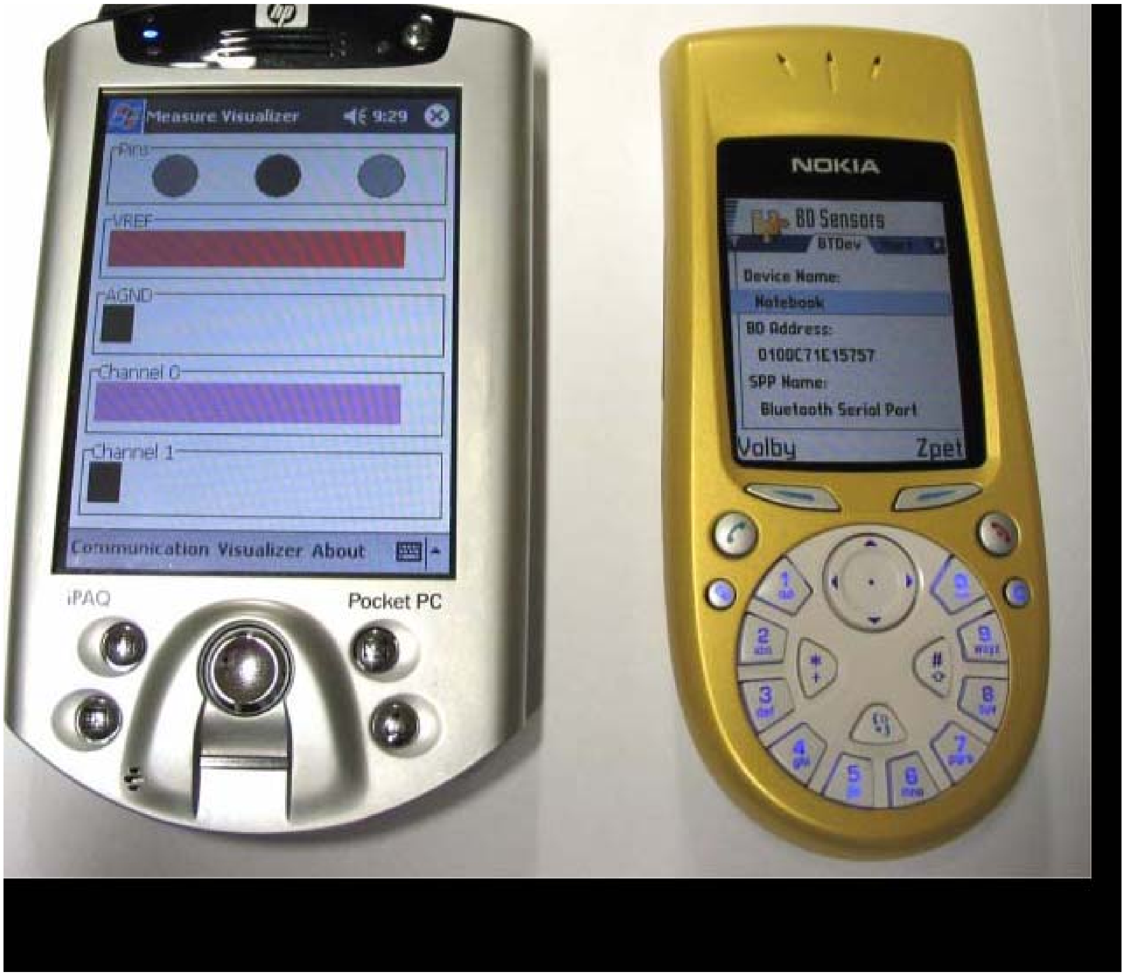


Figure 4. IPAQ PDA 5450 visualizing bar graphs of measured quantities
and GSM Nokia 3650 SmartPhone with a sample window for smart sensor parameter settings

Most of current control and measurement applications deploy galvanic connection of sensors and actuators to central control system. That scheme can cause some problems and limitations in cases, where it is not possible to use standard cable connection between control system (wiring station) and local sensors and actuators. Galvanic interconnection is possible in cases, when the configuration is permanent and it is possible to make cable interconnection between components.

This contribution, which stems from partial results published earlier, proposes not only how to represent a system's solution as an application example, but also how to generalize reached research results.

B. *Bluetooth*

The basic concept for connection of distributed smart sensor to the central data acquisition station is based on Bluetooth standard specification [1]. This standard defines data exchange among remote stations using RF wireless interface. The Bluetooth operates in the Industrial-Scientific-Medicine Band (ISM), which is in most countries defined as a band ranging from 2 400 MHz to 2 483.5 MHz.

In many countries, including the Czech Republic, the radio transmissions in the ISM band are not licensed. The Bluetooth standard defines 79 channels with the frequency width of 1 MHz in the ISM band. The position of any channel in the ISM band can be calculated as follows

$$f = 2402 + k,\ k = 0\ ...\ 78\ \text{MHz}.$$

Devices corresponding to the Bluetooth standard are subdivided according to the transmitted power into three power classes: 100 mW, 2.5 mW and 1 mW.

Bluetooth based systems consist of the following components: radio transmitter (2.4 GHz Bluetooth radio), link controller (controls the transmitter), and link manager & I/O (provides terminal interface). The Bluetooth standard defines two different data transmission methods. The first one defines synchronous data channel (Synchronous Connection Oriented -SCO), which is intended mostly for audio transmission, while the second one defines asynchronous data channel (Asynchronous Connectionless - ACL). Both SCO and ACL utilize the same RF line.

The wireless network is built on the PICONET, which is the simplest interconnected Bluetooth network that can consist of up to eight nodes. In every PICONET one node acts as a MASTER, the rest as SLAVEs. More PICONETs can form a higher-level entity called as SCATERNET. The formation of SCATTERNET is permit for more PICONETs in the same location, of course, respecting some restrictions on transmission capacity.

A channel is represented by a pseudo-random sequence defining change of used frequency (hop sequence). All PICONETs share the same RF band; however, each PICONET has its own hop sequence. Thus at one time each PICONET uses its own 1 MHz wide channel. For the actual data acquisition the capabilities of transmission lines are very important. Bluetooth standard defines two different transmission lines that provide different throughput: asynchronous ACL and synchronous SCO. According to the demanded transmission capacity, it is possible to select the required type of the transmission line on the fly; so, the type of the transmission line can be negotiated as needed.

The SCO line allows synchronous 64 Kbit/s transmission. The ACL line allows transmissions asynchronous and PICONET-wide broadcasts. As the ACL deploys multi-slot system, it is possible to achieve transmission speed of 721 Kbit/s in one direction and 57.6 Kbit/s in the opposite direction (presented transmission speed expects zero error corrections).

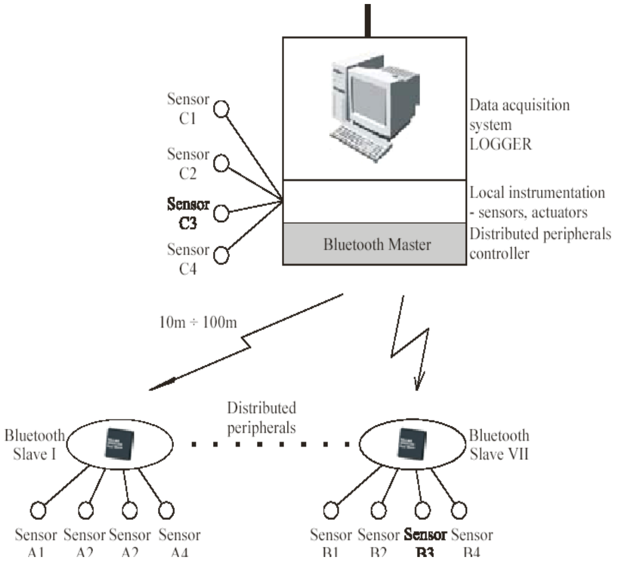


Figure 5. Structure of data acquisition system with data logger

Central data acquisition node (data logger) utilizes both integrated local sensory instrumentation, and distributed remote sensors connected to wireless Bluetooth network, see Fig.5. The data logger acts as a communication master, while the remote sensors act as slaves. In such network, there can be up to 7 distributed sensors. However, the distance between the master node and the slave cannot exceed 10/100 meters.

C. Data acquisition system

The number of possible Bluetooth masters in the central node defines the limitation of the entire distributed data acquisition system. The number of Bluetooth modules located in the same area is limited by the Bluetooth specification; hence, there should be guaranteed communication capacities that fit the application. The presented data acquisition system enables to include wide range of sensors with different requirements on their channel throughputs, e.g. simple temperature and humidity sensors as devices with narrow-band requirements, and CCD cameras as sensors with wide-band requirements on data throughput.

V. CONCLUSIONS

The paper, in its general part, is devoted to a brief review of state of the art of CPS domain, including important properties that can be implemented for applications, and to the typical features of CPS design, which are demonstrated in the rest of the paper. The presented real-world CPS applications deal with two designed sensory systems focusing on their physical design principles, system structures and communication architectures.

Our research group is currently launching a related continuation research that aims at the formal tools support of CPS design [11], [12], [13]. Evidently, this new research domain requires not only formal specification and verification techniques extensions and modifications, but also novel approaches and adaptations of such general methods as model checking and proving, see e.g., [1], [2], [3], [4], [8], [9] and [14].

ACKNOWLEDGMENT

The authors acknowledge contributions of their colleagues, namely Radek Kuchta, Pavel Steffan, Ondrej Sajdl and Zdenek Barton.

This project has been carried out with a financial support from the Czech Republic state budget through the CEZ MMT projects no. MSM0021630528: *Security-Oriented Research in Information Technology* and no. MSM 0021630503: *MIKROSYN - New Trends in Microelectronic Systems and Nanotechnologies*, by the Grant Agency of the Czech Republic through the grant no. GACR 102/08/1429: *Safety and Security of Networked Embedded System Applications*, by the Ministry of Industry and Trade CR through the grant no. FR-TI1/037 of Ministry of Industry and Trade: *Automatic Attack Processing*, and by the Brno University of Technology, Faculty of Information Technology through the specific research grant no. FIT-10-S-1: *Secured, Reliable and Adaptive Computer Systems*.

REFERENCES

[1] R. Akella and B.M. McMillin, Model-checking BNDC Properties in Cyber-Physical Systems, *Proceedings of the33rd INternational Computer Software and Applications Conference COMPSAC 2009*, IEEE CS, New York, NY, US, 2009, pp. 660-663.

[2] Borzoo Bonakdarpour, Challenges in Transformation of Existing Real-Time Embedded Systems to Cyber-Physical Systems *IEEE Symposium on Real-Time Systems RTSS RTSS - Ph.D. Forum on Deeply Real-Time Embedded Systems*, Tucson, Arizona, 2007, 2pp.

[3] M.C. Bujorianu and H.Barringer, An Integrated Specification Logic for Cyber-Physical Systems, *Proceedings of the 14th IEEE International Conference on Engineering of Complex Computer Systems*, Potsdam, Germany, 2009, pp. 91-300.

[4] J. C. Eidson, E. A. Lee, S. Matic, S. A. Seshia, and J. Zou, Time-centric Models For Designing Embedded Cyber-physical Systems, EECS Department, University of California, Berkeley, *Technical Report No. UCB/EECS-2009-135*, October 9, 2009.

[5] J. Fraden, *Handbook of Modern Sensors.* New York, AIP Press, 1997.

[6] B.H. Krogh, E. Lee, I. Lee, A. Mok, R. Rajkumar, L.R. Sha, A.S. Vincentelli, K. Shin, J. Stankovic, J. Sztipanovits, W. Wolf, and W. Zhao, *Cyber-Physical Systems, Executive Summary*, CPS Steering Group, Washington D.C., March 6, 2008. [http://www.nsf.gov/pubs/2008/nsf08611/nsf08611.htm] [accessed: June 30, 2010]

[7] R. Kuchta, P. Steffan, Z. Barton, R Vrba, and M. Sveda, Wireless Temperature Data Logger, *Proceedings of the 2005 Asian Conference on Sensors, and International Conference on new Techniques in Pharmaceutical and Biomedical Research*, 5-7 Sept. 2005 pp. 208-212.

[8] E.A. Lee, Computing Needs Time, *Communications of the ACM*, Vol.52, No.5, pp. 70-79, May 2009.

[9] National Science Foundation, *Cyber-Physical Systems Program Solicitation, NSF 10-515*, Arlington, VA, US, March 11, 2010

[10] J.A. Stankovic, I. Lee, A. Mok, and R. Rajkumar, Opportunities and obligations for physical computing systems, *IEEE Computer*, November 2005, pp. 23-31.

[11] M. Sveda and R. Vrba, An Embedded Application Regarded as Cyber-Physical System, *Proceedings of the Fifth International Conference on Systems ICONS 2010*, Les Menuires, FR, IARIA, 2010, pp. 170-174.

[12] M. Sveda and R. Vrba, Meta-Design with Safe and Secure Embedded System Networking, *International Journal On Advances in Security*, Vol. 2, No. 1, 2009, US, pp. 8-15.

[13] M. Sveda and R. Vrba, Specifications of Secure and Safe Embedded System Networks, *8th International Conference on Networks Proceedings ICN 2009*, New York, NY, US, IARIA, IEEE CS, 2009, pp. 220-225.

[14] H. Tang and B.M. McMillin, Security Property Violation in CPS through Timing, *Proceedings of the 28th on Distributed Computing Systems IDCS 2008, Workshops*, IEEE CS, New York, NY, US, 2008, pp. 519-524.

[15] R. Vrba, O. Sajdl, O., R. Kuchta, and M. Sveda., Wireless Smart Sensor Network System. In Proceedings of the ICSE & INCOSE 2004 Conference. Las Vegas, Nevada: CRC Press LLC, 2004, pp. 104-109.

Event Driven Programming for Embedded Systems - A Finite State Machine Based Approach

Nrusingh Prasad Dash*, Ranjan Dasgupta†, Jayakar Chepada‡ and Arindam Halder§
Innovation Lab, Tata Consultancy Services Ltd. Kolkata, 700091, India
*E-Mail:*nrusingh.dash@tcs.com,†ranjan.dasgupta@tcs.com, ‡jayakar.ch@tcs.com, §arindam.halder@tcs.com*

***Abstract*—The paper gives a brief overview of event driven program and its relationship with finite state machine (FSM). It proposes a FSM-based framework implemented in C programming language. In Microcontroller Unit (MCU) based tiny embedded system, FSM based software design and event-driven programming techniques are efficient in handling concurrent and asynchronous events usually occur. Finally, the paper states and demonstrates a solution of a system power sequence problem using the same framework as a case study.**

Keywords - Event Driven Programming; FSM; States; State Transitions; MCU.

I. Introduction

Most of the tiny embedded systems respond to external or internal events in some or other way. The external event can be an interrupt, or change of signal level at I/O pins, a message packet coming from other part of the system through some interface, e.g., serial peripheral interface(SPI), inter-integrated circuit(I2C), two wire interface(TWI), or simply an expiry of internal timer. This paper initially discusses the prior art on software implementation of FSM in Section II, subsequently provides the theoretical background of event driven programming paradigm and how the event driven programming problems can be solved using finite state machines(FSM) in Sections III, and IV respectively. In Section V, the paper proposes a FSM framework and narrates a case study where, the same framework has been used to implement an event driven application efficiently and easily on a TI MSP430F1232 MCU based system. Finally, the paper discusses the performance figures of the case study in Section VI

II. State of The Art

There are various approaches taken for the software implementation of finite state machines (FSM). The works presented in [1] and [3] are switch-case driven FSM implementations where, several comparisons are required before execution of the event handler. The number of comparison increases with the number of states and events. The more is the number of comparisons the more CPU cycle is consumed. In the work [4] a table driven event handler hashing approach has been taken to implement the FSMs, but does not separate out the FSM framework and FSM implementation, therefore lacks re-usability. The works [2] and [5], emphasize on model driven FSM generation techniques, but, to maintain the genericness and re-usability, generated code for FSM would require high memory foot-print as well as more CPU cycles and therefore may not be suitable for the embedded systems with very tight memory and cpu horse-power budget. The current work aims at an re-usable, simple and compact FSM framework, which takes minimal CPU cycles and less memory foot-print to implement an FSM problem.

III. Event Driven Programming

The events are mostly generated when user actions are done on a system. The user actions can be a press of a push button or a key pad, touch a touch screen, move or click of a mouse. The events can also be generated from the sensors or devices connected to a system (may be through interrupts or may be form of message packets through a physical interface). Sometimes event may be generated internally, e.g., timeout event or a software exception. Irrespective of the source or type of events, the event driven programming talks about a programming paradigm in which the flow of the program is determined by the events. The actual implementation of event driven programming can be done with any programming language, like C/C++ etc. Broadly these implementations have following sections of programs.

- Event Capture Section
- Event Despatch Section
- Event Handlers Section

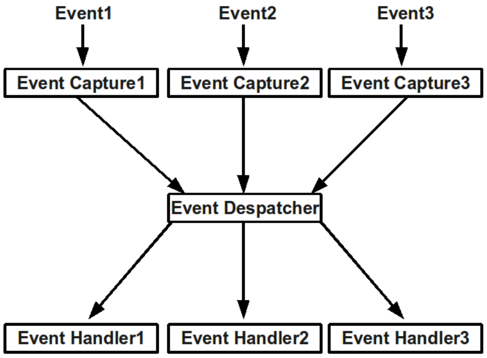


Figure 1. Sections of Event Driven Program

A. *Event Capture Section*

This section of program is responsible for capturing the events, do pre-processing and identifying the event type. The event capture section can be at one place or may be distributed across several interrupt handlers and the main background loop.

B. *Event Despatch Section*

The responsibility of the dispatcher is to map the events with the respective handler and calls the handler. If there is no associated handler with an event, dispatcher either drops the event or raises an exception.

C. *Event Handlers Section*

The event handlers implement the activities; those should take place on occurrence of an event.

Many event driven programs are state less, which means when any application finishes processing an event, the application does not require to maintain its earlier event. When the event occurs, the respective handler is just executed. It means it is a state less event driven program where the execution flow is not dependent on the earlier events. On the contrary the other category of event driven programs, where the execution flow is dependent on not only the current event but also the sequence of prior events, called as state full event driven programs. This article discusses about the latter category and how FSM can be used to solve state full event driven programming problems.

IV. Finite State Machine (FSM)

Finite State Machine (FSM) is a model behavior composed of a finite number of states, transitions between those states, and actions.

Finite state machines (FSM) consist of 4 main elements:

- **States** - Define behavior and may produce actions
- **State Transitions** - Switching of state from one to another
- **Conditions** - Set of rules which must be met to allow a state transition
- **Input Events** - Triggers which are either externally or internally generated, which may possibly invoke conditions and upon fulfilling the conditions lead to state transition.

Every FSM has an initial state, which is the starting point. The input events act as triggers, which cause an evaluation of the conditions imposed. On fulfilling those, the current state of the system switches to some other state, which is called as state transition. State transitions, may happen along with the associated actions in most of the cases. The actions can happen, before entering to a state or at exiting the state or while being in the state itself.

V. Case Study - A System Power ON/OFF Sequence

Problem Statement - There is a power key in a system. Initially the system is assumed to be off. When the system is off, if the power key is pressed for 2 seconds, it switches on. When the system is on, if power key is pressed for 2 seconds it switches off. But if the key press time is less than 2 seconds while system is either on or off state, it remains in same state which means there is no state transition. The Fig. 2 is the unified modeling language(UML) state chart representation of the problem stated above.

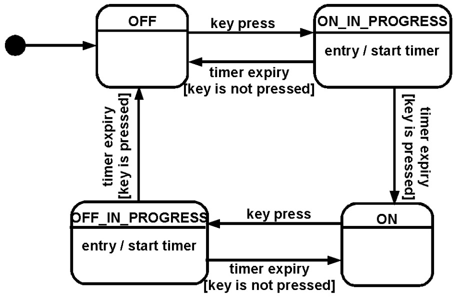


Figure 2. UML State Chart

The problem stated above can be very easily implemented using a FSM framework, that has been proposed in this work.

A. *The FSM Framework Implementation*

The C program Listing 1 in the appendix is a state transition table-driven implementation of FSM framework, which can be used for a quick and easy implementation of a FSM problem. The framework exposes following interfaces for programmer.

sm_declare_states(fsm_name, list_of_states) - Interface for declaring states. The first parameter is the name of the state machine and the rest of the parameters are the list of states separated by coma.

sm_declare_events(fsm_name, list_of_events) - Interface for declaring events. The first parameter is the name of the state machine and the rest of the parameters are are the list of events separated by coma.

sm_declare_state_machine(fsm_name, initial_state, list_of_state_handlers) - Interface for declaring the state transition table and initializing the state to initial_state.

sm_handle_event(fsm_name, event) - Interface is used for handling the event after the event is captured.

sm_set_state(fsm_context, state) - Interface is used inside the event handlers to set the next state.

sm_set_private_data(fsm_context, private_data) - Interface is used inside the event handlers to set the problem specific private data.

sm_get_private_data(fsm_context) - Interface is used inside the event handlers to obtain the reference to the problem specific private data.

sm_define_handle_event(fsm_name) - This is not an interface for the programmer. This macro defines the state machine handling function for the events. Note: It is customary to use this macro without putting a semicolon after it.

sm_declare_handle_event(fsm_name) - This is not an interface for the programmer. This macro declares the state machine handling function. It is customary to use this macro in the header file of the actual state machine implementation ending with a semicolon.

B. FSM Implementation of System Power ON/OFF Sequence

The FSM framework described in previous subsection is used to implement the FSM problem of system power on/off sequence described earlier.

The program Listing 2 in the appendix demonstrates how easily the states and the events can be declared. It is just a matter of using two macros, i. e., **sm_declare_states(fsm_name, list_of_states)** and **sm_declare_events(fsm_name, list_of_events)**. In this case the states are **OFF**, **ON_IN_PROGRESS**, **ON**, **OFF_IN_PROGRESS** and the events are **KEY_PRESS**, and **TIMER_EXPIRY**.

The program Listing 3 in the appendix demonstrates the implementation of the state transition table and the event handlers. The initial state setting and the state transition table definition is done with **sm_declare_state_machine(fsm_name, initial_state, list_of_state_handlers)** interface. The guard conditions, e.g., the check for the key is still pressed at the timer expiry or not is implemented as a condition check within the event handlers. The transition to next state is also done inside event handler using **sm_set_state(fsm_context, state)** interface.

The event capture sections are distributed. In the current problem the event **KEY_PRESS** is captured as polling of the respective pin in the main background loop and the event **TIMER_EXPIRY** is captured in timer interrupt context as timer expiry callback function. The event capture and despatch is demonstrated in program Listing 4 in the appendix. After the events are captured they are despatched to respective event handles using **sm_handle_event(fsm_name, event)**.

Note: The timer implementation and the power key state check program listings are not included or described to maintain the focus on the FSM implementation.

The FSM framework presented in this paper has theoretical commonality with the FSMs presented in [1] and [3] but has novelty in its implementation. It uses a pre-hashed event handling approach where, the current state and event are used as hash-keys to fetch the event handler from table, without requiring comparisons and saves the CPU cycles consumed. Unlike [4], the work presented in this paper separates out the FSM framework and FSM implementation, enhancing the re-usability of the framework. The Listing 1 in the appendix implements a reusable framework, which can be reused for implementing other FSM problems. The CPU cycle consumption and memory foot-print figures are very minimal as discussed in Section VI which proves its suitability for embedded systems applications.

However, the framework presented in this paper is suitable for the FSM problems where, the most of the state and event combinations are handled. Otherwise, the respective table entry consumes memory without doing any useful activity.

VI. Conclusion

The FSM framework discussed is not only a very quick and easy way to implement a FSM problem, but also the memory footprint of the generated code is very less and being a function table driven event handling implementation, the execution is pretty fast. These characteristics make the framework very much suitable for the tiny embedded systems application where the memory and processor resources are very scarce. The framework is used for the power sequence problem described in the paper on a MSP430F1232 MCU and the memory footprint is as below.

- 164 bytes of code memory
- 8 bytes of data memory
- 24 bytes of constant memory

The FSM consumes approximately 18 instruction cycles between the event despatch and the event handler is called. The above figures are reported using IAR Workbench [6] v4.

References

[1] Miro Samek, *Practical UML Statecharts in C/C++*, 2nd Edition, Newnes.

[2] Ilija Basicevic, Miroslav Popovic, and Ivan Velikic *"Use of Finite State Machine Based Framework in Implementation of Communication Protocols A Case Study"* Sixth Advanced International Conference on Telecommunications, May 9 - 15, 2010

[3] Andrei Drumea and Camelia Popescu, *"Finite State Machines and their applications in software for industrial control"*, 27th International Spring Seminar on Electronics Technology: Meeting the Challenges of Electronics Technology Progress, May 13 -16, 2004

[4] Johannes Weidl, Ren6 R. Klosch, Georg Trausmuth, and Harald Gall, *"Facilitating program comprehension via generic components for state machines"*, Fifth Iternational Workshop on Program Comprehension, March 28 - 30, 1997

[5] Chung-Shyan Liu, and Kuo-Hua Su, *"An FSM-Based Program Generator for Communication Protocol Software"*, Eighteenth Annual International Computer Software and Applications Conference, November 9-11, 1994

[6] http://www.iar.com/website1/1.0.1.0/220/1/

APPENDIX

Listing 1. FSM framework

```
#ifndef _SM_FRAMEWORK_H_
#define _SM_FRAMEWORK_H_
/*
File:sm_framework.h
Description: A framework for finite state machine implementations
*/
typedef struct sm_context sm_context_t;
typedef void (*sm_handler_t) (sm_context_t *);

#define sm_declare_states(name, ...)      \
typedef enum {                            \
    __VA_ARGS__,                          \
    name##_STATE_COUNT                    \
} name##_state_e

#define sm_declare_events(name, ...)     \
typedef enum {                           \
    __VA_ARGS__,                         \
    name##_EVENT_COUNT                   \
} name##_event_e


#define sm_declare_state_machine(name, st_init, ...) \
struct sm_context { \
    void                    *priv; \
    name##_state_e          state; \
    name##_event_e          event; \
    sm_handler_t            *handler; \
}; \
static const unsigned int \
name##_handler[name##_EVENT_COUNT][name##_STATE_COUNT] = {__VA_ARGS__}; \
static sm_context_t name##_context \
= {0, st_init, (name##_event_e) 0, (sm_handler_t *)name##_handler}

#define sm_set_state(c, s) (c)->state  = (s)
#define sm_set_private_data(c, p) (c)->priv = (p)
#define sm_get_private_data(c) (c)->priv

#define sm_define_handle_event(name) \
void name##_handle_event(name##_event_e ev) \
{   sm_context_t *c = &name##_context; \
    c->event = (ev); \
    c->handler[c->event * name##_STATE_COUNT + c->state](c); \
    return; \
}

#define sm_declare_handle_event(name) \
void name##_handle_event(name##_event_e ev)

#define sm_handle_event(name, ev) name##_handle_event(ev)

#endif /*_SM_FRAMEWORK_H_*/
```

Listing 2. Events and states declaration

```
#ifndef __POWER_STATE_MACHINE__
#define __POWER_STATE_MACHINE__
#include
/*
* File: power_state_machine.h
* Description: States and events declared
*/
sm_declare_states(power,
                  OFF,
                  ON_IN_PROGRESS,
                  ON,
                  OFF_IN_PROGRESS
                  );

sm_declare_events(power,
                  KEY_PRESS,
                  TIMER_EXPIRY
                  );

sm_declare_handle_event(power);

#endif /*__POWER_STATE_MACHINE__*/
```

Listing 3. Power state machine

```
#include <power_state_machine.h>
static void no_action_handler(sm_context_t * context);
static void on_key_press_when_off(sm_context_t *context);
static void on_timer_expiry_when_on_in_progress(sm_context_t *context);
static void on_key_press_when_on(sm_context_t * context);
static void on_timer_expiry_when_off_in_progress(sm_context_t *context);
sm_declare_state_machine(power,
/* initial state */
OFF,
/* event = KEY_PRESS state = OFF */
(unsigned int) on_key_press_when_off,
/* event = KEY_PRESS state = ON_IN_PROGRESS */
(unsigned int) no_action_handler,
/* event = KEY_PRESS state = ON */
(unsigned int) on_key_press_when_on,
/* event = KEY_PRESS state = OFF_IN_PROGRESS */
(unsigned int) no_action_handler,
/*
* event = TIMER_EXPIRY
* state = OFF
*/
(unsigned int) no_action_handler,
/*
* event = TIMER_EXPIRY
* state = ON_IN_PROGRESS
*/
(unsigned int) on_timer_expiry_when_on_in_progress,
/*
* event = TIMER_EXPIRY
* state = ON
*/
(unsigned int) no_action_handler,
/*
* event = TIMER_EXPIRY
* state = OFF_IN_PROGRESS
*/
(unsigned int) on_timer_expiry_when_off_in_progress);

/*don't put semicolon at the end of the line bellow*/
sm_define_handle_event(power)
```

```

static void no_action_handler(sm_context_t *
    context)
{
    return;
}
static void on_key_press_when_off(sm_context_t
    *context)
{
    timer_start(TIMER_ID);
    sm_set_state(context, ON_IN_PROGRESS);
    return;
}
static void on_timer_expiry_when_on_in_progress
    (sm_context_t *context)
{
    if(is_power_key_pressed())
    {
      sm_set_state(context, ON);
    }
    else
    {
      sm_set_state(context, OFF);
    }
    return;
}
static void on_key_press_when_on(sm_context_t *
    context)
{
    timer_start(TIMER_ID);
    sm_set_state(context, OFF_IN_PROGRESS);
    return;
}
static void
    on_timer_expiry_when_off_in_progress(
    sm_context_t *context)
{
    if(is_power_key_pressed())
    {
        sm_set_state(context, OFF);
    }
    else
    {
        sm_set_state(context, ON);
    }
    return;
}
```

Listing 4. Event capture and dispatch

```
#include <power_state_machine.h>
/*
* File: main.c
*/
#define TIMER_ID            1
#define TIMER_DURATION      2
/*
* Timer handler called in interrupt context
* captures the timer expiry event
*/
void power_key_timer(void *data)
{
     sm_handle_event(power, TIMER_EXPIRY);
}
/*
* Background loop for capturing key press event
*/
int main(void)
{
    timer_init(TIMER_ID,
               TIMER_DURATION,
               TIMER_ONE_SHOT,
               power_key_timer,
               NULL);
    while(1) {
        if(is_power_key_pressed()) {
            sm_handle_event(power, KEY_PRESS);
        }
    }
}
```

Optimizing Collective Communications on the K-port Spidergon Network

Jiri Jaros
Dept. of Computer Systems
Faculty of Information Technology, BUT
Brno, Czech Republic
e-mail: jarosjir@fit.vutbr.cz

Vaclav Dvorak
Dept. of Computer Systems
Faculty of Information Technology, BUT
Brno, Czech Republic
e-mail: dvorak@fit.vutbr.cz

***Abstract*—The paper investigates an impact of using *k* ports in the direct communication model of collective communications on the overall performance of the Spidergon interconnection network. Since the higher number of *k* internal ports can improve performance but increase the cost of interconnection network, the performed analysis introduces the ideal performance-cost tradeoff on slim- and fat-node Spidergon networks.**

Keywords-collective communications; k-port model; Spidergon; slim-nodes; fat-nodes; router usage; latency

I. INTRODUCTION

With an increasing number of processor cores, memory modules and other hardware units in the latest chips, the importance of communication among them and of related interconnection networks is steadily growing. The memory of many-core systems is physically distributed among computing nodes that communicate by sending data through a Network on Chip (NoC) [1].

Communication operations can be either point-to-point, with one source and one destination, or collective, with more than two participating processes. Some embedded parallel applications, like network or media processors, are characterized by independent data streams or by a small amount of inter-process communications [2]. However, many general-purpose parallel applications display a bulk synchronous behavior: the processing nodes access the network according to a global, structured communication pattern.

The performance of these collective communications (CC for short) has a dramatic impact on the overall efficiency of parallel processing. The most efficient way to switch messages through the network connecting multiple processing elements (PEs) makes use of wormhole (WH) switching. Wormhole switching reduces the effect of path length on communication time, but if multiple messages exist in the network concurrently (as it happens in CCs), contention for communication links may be a source of congestion and waiting times. To avoid congestion delays, CCs are necessary to organize into separated steps in time and to put into each step only such pair-wise communications whose paths do not share any links. The contention-free scheduling of CCs is therefore important.

The port model of the system defines the number *k* of PE ports that can be engaged in communication simultaneously. This means that beside 2*d* network channels, there are 2*k* internal unidirectional (DMA) channels, *k* input and *k* output channels, connecting each local processor core to its router that can transfer data simultaneously. Always $k \leq d$, where *d* is a node degree; a one-port model (k=1) and an all-port router model (k=d) are most frequently used. Typically, higher number of ports reduces communication overhead, but on the other hand, increases the complexity of routers and duplicates network interfaces in connected PEs.

In the most common one-port system, a PE has to transmit (and/or receive) messages sequentially (using only one local channel). The messages may block on occupied injection channel, even when their required network channels are free. These systems are very easy to implement and are often used in computer clusters equipped with only one network interface.

Architectures with multiple ports alleviate this bottleneck. In the all-port router architecture, there are as many local PE channels as there are network channels that reduce the message blocking latency during CC operations. On the other hand, an addition of internal ports requires more complex router and makes the system more expensive. Such all-port routers can be often found in systems on a chip. Fig. 1 illustrates the differences between one-port and all-port switches.

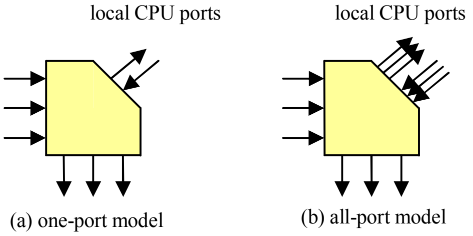


Figure 1. Port models for 3-regular Spidergon network.

The *k*-port model is a generalization of the port models and has been widely used, e.g., in [3] and [4]. An appropriate number of internal ports can boost the performance and keep the router complexity at reasonable level.

One example of successful *k*-port NoCs implementation is presented in [5] and [6]. The authors investigate the speedup of broadcast communication inside the Cell Broadband Engine processor [7], [8] and prove that using multi-ports (up to four) significantly reduces the broadcast latency of short messages. Unfortunately, no idea about other communication patterns was given there.

The paper [9] presents a novel analytical model to compute communication latency of multicast as a widely used collective communication operation in wormhole routed Spidergon [10] and Quarc [11] network. This model can predict the latency of broadcast communication in asynchronous multi-port wormhole networks. Unfortunately, the paper does not present any advantages of multi-port model against one-port model.

K-port model can also be effectively used in high-end workstations that are more and more often equipped with two network interfaces. An example implementation of *k*-port model can be found in SuperMicro SuperBlade servers [12]. These servers are equipped with 2-port Gigabit Ethernet connectors and thus can be connected into the interconnection network, e.g., Spidergon, using two ports.

The influence of using *k*-port on CC overhead and complexity of on-chip design has not been deeply investigated as yet. This paper deals with examining the advantages and disadvantages of *k*-port model on Spidergon network [10] with slim and fat nodes. It introduces the optimal number of ports for several Spidergon configurations and discusses their overhead.

The paper is structured as follows. Section II describes the time complexity of CCs and derives their lower bounds. Section III introduces the Spidergon topology and its slim-node and fat-node configurations. The implementation requirements of Spidergon are discussed here. Section IV presents known schedule techniques for CC on the Spidergon network and identifies their weakness. The new CC schedules obtained by means of evolutionary algorithms are presented in section V. The comparison of slim-node and fat-nodes simultaneously with various number of ports are outlined there. The Conclusion summarizes the achieved results and introduces the most suitable configurations.

II. Time Complexity of Collective Communications

A collective communication is usually defined in terms of a group of processes. The operation is executed when all processes in the group call the communication routine with matching parameters. We classify collective operations into three types according to their purpose:

- CCs (OA-One-to-All, AO-All-to-One, AA-All-to-All),
- global computation (reduction AOR or AAR and scan)
- synchronization (barrier).

The CCs are most important, as other collective operations are closely related to them. In a broadcast (OAB), one process sends the same message to every group member, whereas in a scatter (OAS), one process sends a different message to each member. Gather (AOG) is the dual operation to scatter, in that one process receives a message from each group member. These basic operations can be combined to form more complex operations. In all-to-all broadcast (AAB), every process sends a message to every other group member. In complete exchange, also referred to as all-to-all scatter-gather (AAS), every group member sends a different message to every other group member. Permutation operations, such as shift and transpose, are also CCs. Since complexities of some communications are similar (AOG ~ OAS, AOR ~ OAB, AAR ~ AAB), we will focus only on four basic types (OAB, OAS, AAB, AAS). Also, from now on, when we refer to “collective communications”, then we will assume only CCs involving the group of all processes.

The simplest time model of point-to-point communication in direct WH networks takes the communication time composed of a fixed start-up time t_s at the beginning (SW and HW overhead of a sender and a receiver), a serialization delay, i.e., the transfer time of m message units (words or bytes), and of a component that is a function of distance h (the number of channels on the route or hops a message has to do):

$$t_{WH} = t_s + mt_1 + ht_r \qquad (1)$$

where t_1 is per unit-message transfer time and t_r includes a routing decision delay, switching and inter-router latency. A relatively small dependence on h may be taken into account by including $h_{\max}t_r$ into t_s, so that only two parameters t_s and mt_1 are sufficient.

In the rest of the paper we assume that the CC in WH networks proceeds in synchronized steps. In one step of CC, a set of simultaneous packet transfers takes place along complete disjoint paths between source-destination node pairs. If the source and destination nodes are not adjacent, the messages go via some intermediate nodes, but PEs in these nodes are not aware of it; the messages are routed automatically by the routers attached to PEs.

Complexity of collective communication will be determined in terms of the number of communication steps or equivalently by the number of “start-ups” τ^{CC} (upper bound). Provided that the term $h_{\max}t_r$ is included in t_s and excluding contention for channels, CC times can be obtained approximately as the sum of start-up delays plus associated serialization delays m_it_1 in individual communication steps.

$$t_{CC} = \sum_{i=1}^{\tau^{CC}} (t_S + m_it_1) = \tau^{CC}[t_S + mt_1] \qquad (2)$$

The above expression assumes that the nodes can only re-transmit/consume original messages, so that the length of messages $m_\mathrm{i} = m$ remains constant in all communication steps. This is true in the so called direct model of communication; on the contrary, in the combining model the nodes can combine/extract partial messages with negligible overhead. The direct/combining model influences CC performance and either one can outperform the other in some cases. Further on we will consider the direct model only.

Possible synchronization overhead involved in communication steps, be it hardware or software-based, should be included in the start-up time t_s. Let us note that with uniform messages and a single clock signal domain on NoC, one barrier synchronization before CC might be sufficient to synchronize the whole CC. Communication steps would then follow in the lockstep. According to frequency of CCs and an amount of interleaved computation

(BSP model) in a certain application, efficiency of parallel processing can be estimated.

Further, the lower bound on the number of steps τ^{CC} depends on a channel type; we have to distinguish between unidirectional (simplex) channels and bi-directional (half-duplex HD, full-duplex FD) channels. Typically τ^{CC} will be twice as large for HD channels than for the FD ones. Further on we will consider FD channels. Finally, the lower bounds depend on number of internal ports k and node degree d.

The number of communication steps is in the first place influenced by the topology of an interconnection network. Generally the lower bounds $\tau_{CC}(G)$ for the network graph G depend on node degree d, number of internal ports k, number of processing elements P, and bisection width B_C, Table I.

TABLE I. LOWER BOUNDS ON THE NUMBER OF COMMUNICATION STEPS τ_{CC} (WH, K-PORT, DIRECT NETWORKS).

CC	τ_{CC} [steps]
OAB	$\lceil \log_{k+1} P \rceil$
AAB	$\lceil (P-1)/k \rceil$
OAS	$\lceil (P-1)/k \rceil$
AAS	$\max(\lceil P^2/(2B_C) \rceil, \lceil \Sigma/(Pd) \rceil, \lceil (P-1)/k \rceil)$

As far as the broadcast communication (OAB) is concerned, the lower bound on the number of steps $\tau_{OAB}(G) = s = \lceil \log_{k+1} P \rceil$ is given by the number of PEs informed in each step, that is initially 1, $1 + 1 \times k$ after the first step, $(k+1) + (k+1) \times k = (k+1)^2$ after the second step, etc.,…, and $(k+1)^s \geq P$ processing elements after step s. Since the broadcast message is the same for all the PEs, each PE once informed can help with distributing of the message in following steps.

In case of AAB communication, since each PE has to accept $P-1$ distinct messages, the lower bound is $\lceil (P-1)/k \rceil$ steps. A similar bound applies to OAS communication, because each PE cannot inject into the network more than k messages in one step.

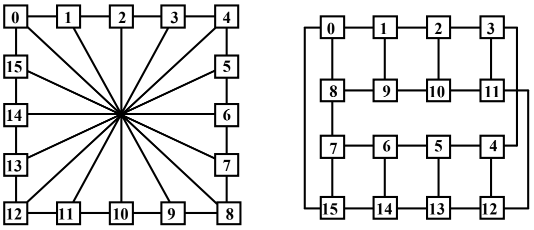
The lower bound for AAS can be obtained considering that half the messages from each PE cross the bisection, whereas the other half do not. There will be altogether $\lceil 2(P/2)(P/2)/B_C \rceil$ of such messages in both ways, where B_C is the channel bisection width [11]. Sometimes a stronger lower bound may be obtained considering the count of channels from all sources to all destinations (Σ) and the limited count Σ_1 of channels available for one step. In regular networks with constant node degree $\Sigma_1 = Pk$. As each PE has to accept $P-1$ distinct messages, $\lceil (P-1)/k \rceil$ bound has to be also obeyed.

Which lower bound takes effect depends on a particular network topology and the port model.

III. SPIDERGON TOPOLOGY AND ITS CONFIGURATIONS

Classical logarithmic diameter networks, e.g., hypercubes, butterflies and fat trees, provide enough bandwidth for all-to-all communications, but do not map well into the two dimensions provided by a silicon chip: the length of some interconnection wires increases proportionally to the number of nodes. This will decrease the clock frequency dramatically and degrade the performance. In this work we therefore restrict our attention to the Spidergon NoC topology with mostly local interconnection among processors.

The Spidergon depicted in Fig. 2 is the novel interconnection network architecture suitable for the on-chip communication demands of SoCs in several application domains [2]. The Spidergon NoC first reported in [10], and later in [11], has been recently adopted by STMicroelectronics [13] with the objective to realize low cost multiprocessor SoC implementation with topology opened for application-specific optimization. Spidergon is somewhere between the ring and mesh topologies: an even number of nodes is connected into a bidirectional ring and pairs of nodes are connected by a cross connection. Each edge in Fig. 2 represents two unidirectional physical links, one for each direction. In order to avoid deadlock, two virtual channels are multiplexed on each physical link. Fig. 2 depicts the 16-node Spidergon topology and its layout on a chip resembling a sparse mesh. Each node represents a router/switch (Fig. 2) and a PE.



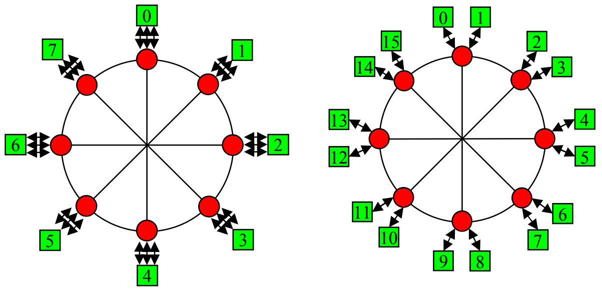
(a) basic spidergon topology (b) the on-chip layout

Figure 2. Isomorphic graphs of the 16-node Spidergon topology.

The nodes placed in direct interconnection networks can be divided into slim and fat ones. Slim nodes contain one PE and a router connecting it into the network. The PE and the router are connected with k internal ports. Slim nodes provide the highest communication performance but lower scalability of the network. Fat nodes with a few PEs connected with separate k internal ports to the router could provide cheaper solution for Spidergon networks of a larger size, similarly as fat hypercubes do. However, it is a trade-off between such measures as cost and performance.

Fig. 3 shows two examples of Spidergon configurations: the slim node all-port Spidergon with 8 nodes; and the 2-fat node one-port Spidergon with 8 nodes carrying 16 PEs.

Finding the optimal ratio between PEs connected to a single router and number of ports used to interconnect them is still an open question. Table II shows the total router port requirements for a few node configurations targeted to Spidergon networks. 1-port, 2-port and 3-port model with slim and 2-fat nodes are compared here. The number of total router ports (including internal and three external ones) is calculated for all configurations. Utilizations of prefabricated 8-port and 12-port routers that could be used for NoC implementation are shown here too.



(a) slim all-port 8-gon

(b) 2-fat one-port 8-gon

Figure 3. Two different 8-node Spidergon networks.

From the Table II, it can be seen that the 2-port 2-fat nodes offer the highest utilization of available links in 8-port router. Only one link remains idle here. Very popular 3-port slim nodes utilizing 6 links also offer an acceptable utilization of router resources.

If the bigger routers are available in NoC, the 3-port 2-fat nodes can be used but at the cost of higher complexity of NoC. Let us note that for one-port slim Spidergon, 4-port router suits best because it utilizes all router channels.

TABLE II. ROUTER UTILIZATION FOR TARGET SPIDERGONS NODES

Port model	Node type	Router ports	Utilization 8p	Utilization 12p
1-port	Slim	4	50%	33%
2-port	Slim	5	62.5%	41.6%
3-port	Slim	6	75%	50%
1-port	2-fat	5	62.5%	41.6%
2-port	2-fat	7	87.5%	58.3%
3-port	2-fat	9	--	75%

IV. SCHEDULING CCS ON K-PORT SPIDERGONS

The Spidergon, as well as the bidirectional ring topology, though very simple, is not free from routing deadlock, because the channel dependency graph is not acyclic [13], [14]. This can be seen on a common permutation called the cyclic shift. The problem can be solved by the introduction of virtual channels [13] and by implementing rules on channel use. However, in conflict-free CCs all source to destination paths are disjoint and therefore there is no competition for shared resources, no danger of deadlock and no need for escape virtual channels. When implementing CCs, we therefore use either one of two virtual channels.

The deterministic shortest path routing algorithms proposed for the Spidergon architecture are so called Across First (aFirst) and Across Last (aLast) [13], [15]. Both algorithms are minimal source routing. An analytical performance model has been analyzed in [16] and the average message latency evaluated. Regarding CCs, only the broadcast and multicast CCs on Spidergon were studied in the past[11]. Other CCs, especially all-to-all communications have not been analyzed in the literature as yet.

In this work, we want to improve the performance of Spidergon NoC by designing such communication schedules that prevent any possible link contention. We also want to investigate the influence of the number of internal port *k* on time complexity of designed schedules and the space overhead of corresponding routers. Optimized CC schedules can be uploaded into switch routing tables and boost the performance of many parallel algorithms. For this reason, four common CC patterns based on broadcast and scatter services will be analyzed.

Further, it should be noted that the lower bounds for fat Spidergon cannot be theoretically estimated like in the case of slim Spidergon, see Table I. This phenomenon can be explained on the 2-fat one-port Spidergon. In this case, neither lower bounds for one-port nor for all-port model apply here. The reason is that we cannot assign 3 network ports of a node explicitly to internal cores (PEs).

Let us also note that the optimal schedules are not known for the *k*-port fat-node Spidergon networks so far.

The optimization of CC scheduling is based on evolutionary algorithms (EA). These techniques applied already to CC scheduling problem on hypercubes of medium size (tens of nodes) [17] were able to find the already known optimum solutions obtained analytically. A schedule (chromosome) is encoded as a set of pair-wise transfers determined in space and time. The fitness function checks the validity of candidate CC schedules. A valid CC schedule for a given number of communication steps must be conflict-free. There are no shared resources (links/ports) in such a schedule. Valid schedules are either optimal (the number of steps equals the lower bound) or suboptimal. Evolution of a valid schedule for the given number of steps is completed as soon as fitness (number of conflicts) drops to zero. If it does not do so in a reasonable time, the prescribed number of steps must be increased of one step and lunched again.

However, for some CCs studied in this work analytic methods to find optimum schedules do not exist, so that the results can be compared only to theoretical lower bounds. The evolution gives us the upper bounds of time complexity that can be attained. It should be noted, that it is not clear if the lower bound can ever be reached.

V. REACHED UPPER BOUNDS ON TIME COMPLEXITY ON SPIDERGON NETWORKS

In this work, slim-node Spidergons with 6, 8, 10, 12, 14, and 16 nodes were examined. Fat-node Spidergons were represented by 6, 8, 10, 12, 14 and 16 nodes Spidergons with 12, 16, 20, 24, 28, and 32 PEs. The near optimal schedules for varying number of internal ports were sought using evolutionary algorithms. Table III and Table IV summarize the time complexity of designed schedules in terms of communication steps (upper bounds).

Two integers in one cell separated by a slash indicate that the lower bound (a smaller integer) has not been reached. A single integer represents both the lower and the upper identical bounds reached by an EA, or the lower bound cannot be determined.

A. Experimetnal Results on Slim-node Spidergons

Table III illustrates the upper bounds of one-to-all CCs are identical with the theoretically derived lower bounds in all cases. The upper bounds are proportional to the number of internal ports *k*. The OAB communication does not depend on *k* too strongly. On the other hand, OAS makes profit from higher number of internal port in full.

A slightly different situation can be seen in the case of all-to-all CCs. The lower bounds were not reached for AAB in all cases, especially for 12- and 14-node Spidergon. There were achieved only suboptimal solutions with one step worse time complexity here. The number of communication steps of AAB depends on number of internal ports significantly, and so, using higher number of port is always better.

The most complex AAS communication shows only a small dependence on number of internal ports. It is given by saturating the network with messages injected even using one port.

B. Experimetnal Results on Fat-node Spidergons

The lower bound on time complexity can be derived only for one-to-all communication in the case of *k*-port fat-node Spidergon. The reason is that we cannot assign 3 network ports of a node explicitly to internal cores. Evolutionary algorithms reached the lower bounds in all cases and designed as fast schedules as possible.

The upper bounds on time complexity are shown in Table IV. The lower bounds of OAB are not very dependent on port model. Further, these values are very close to the results obtained for slim-node spidergons. There are only one step differences in most cases, but with twice more connected PEs. The results reached for OAS show double upper bounds, which is caused by double number of PEs.

The results of evolution produced for AAB show the strong dependency on number of internal ports. The increase of upper bound is more than linear for 2-fat Spidergon than in the case of slim Spidergon.

Finally, Table IV illustrates an insignificant influence of *k* on AAS upper bounds. In most cases, the one-port model is sufficient. Let us note, the lower bound for *k* ports cannot be derived exactly, and so, the upper bounds reached by evolutionary algorithms give us the most accurate estimation.

C. Comparison of Slim-node and Fat-node Spidergons

In this subsection, we would like to mutually compare slim-node 16-Spidergon and 2-fat 8-Spidergon. These topologies connect the same number of PEs but in different manners. In the case of slim-node 16-Spridergon, there are 16 nodes placed around the ring and interconnected using cross links. In addition, each PE holds its own router. In the case of 2-fat 8-Spidergon, there are only 8 nodes around the ring and a router is shared between two PEs.

Looking at Table III and Table IV it is evident that upper bounds for one-to-all CCs are the same. The slim-node Spidergon is slightly better for AAB, but on the other hand, it is outperformed by fat-node one for AAS.

Similar observation can be done comparing slim-node 12-Spidergon and 2-fat 6-Spidergon that shows the fat topology gives the same performance but employing only a half of routers.

Taking into account router utilizations presented in Table II, it can be concluded that the optimal tradeoff between performance and router utilization is represented by 2-port fat-node Spidergons. These configurations bring a utilization of 87.5% of 8-port router with sufficient performance. Usage of 3-port slim-node Spidergons lead to lower router utilization, but can bring desired speed-up. On the other hand, 1-port slim-node Spidergon utilize routers in full, but this solution limit the performance dramatically. Finally, 3-port fat-node Spidergons require more complex routers and thus it is not attractive for NoC.

VI. Conclusions

We addressed the problems "is it better to use slim-node or fat-node Spidergon and what number of the internal ports should be implemented?" The lower bounds on time complexity cannot be mathematically derived for some Spidergon configurations. For this reason, an evolutionary algorithm was employed to find the lowest possible upper bounds and simultaneously corresponding conflict-free schedules that have not been known so far. The original contribution of the paper is an assessment of upper bounds of CCs on Spidergon network with fat-nodes and *k* internal ports. The assessments done with evolutionary algorithms are presented in Table III and Table IV.

Taking into account router ports utilization and number of interconnection links, fat-node spidergons seem to be more suitable for networks on chip. The performance degradation using fat nodes is very low and even higher for all-to-all scatter CC pattern.

The experimental results also indicate that CCs scale well with the number of internal ports. The only one exception is the AAS communication where the upper bound is given by interconnection network topology.

Considering limited resources on chip and router utilization, the most suitable Spidergon configurations use two PEs in one node, each connected by two internal ports to a shared router. This statement can be generalized to all 3-regular topologies (three output links), e.g., 3D hypercube.

Future research will be oriented toward optimizing CCs on Spidergons with more PEs in a node and also on complex comparison of slim and fat-node Spidergon. Next research will be oriented on investigation of the influence of port model on networks with higher number of external links like K-ring.

Acknowledgment

This research has been partially supported by the research grants "Safety and security of networked embedded systems applications", GA 102/08/1429 of Czech Science Foundation (2008-10), "Natural Computing on Unconventional Platforms", GP103/10/1517, Czech Science Foundation (2010-13), "Secured, reliable and adaptive computer systems", BUT FIT grant FIT-10-S-1 (2010) and the research plan "Security-oriented research in information technology", MSM 0021630528 (2007-13).

REFERENCES

[1] D. N. Jayasimha, B. Zafar, and Y. Hoskote, “On-Chip Interconnection Networks: Why They are Different and How to Compare Them”, Platform Architecture Research, Intel Corporation, 2006.

[2] A. Jantsch and H. Tenhunen, “Networks on Chip”, Kluwer Academic Publ., Boston, 2003.

[3] Q. F. Stout and B. Wagar, "Intensive hypercube communication: prearranged communication in link-bound machines", in Journal of Parallel and Distributed Computing, vol. 10, pp. 167-181, 1990.

[4] J. Bruck, Ching-Tien Ho, S. Kipnis, E. Upfal, and D. Weathersby: “Efficient Algorithms for All-to-All Communications in Multiport Message-Passing Systems”, in IEEE Transactions on parallel and distributed systems, vol. 8, no. 11, pp. 1143-1156, 1997.

[5] F. Khunjush and N. J. Dimopoulos, “Characterization of single-port and multi-port collective communication operations on the Cell BE processor”, in IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp. 624-630, 2009.

[6] Y. Qian and A. Afsahi, “High Performance RDMA-based Multi-port All-gather on Multi-rail QsNet II”, in High Performance Computing Systems and Applications, p. 3, 2007,

[7] J. A. Kahle, M. N. Day, H. P. Hofstee, C. R. Johns, T. R. Maeurer, and D. Shippy, “Introduction to the Cell Multiprocessor”, in IBM Journal Research and Development, vol. 49, pp. 589-604, 2005.

[8] S. Williams, J. Carter, L. Oliker, J. Shalf, and K. Yelick, “Lattice Boltzmann Simulation Optimization on Leading Multicore Platforms”, in International Parallel and Distributed Processing Symposium (IPDPS 2008), USA, pp. 1-14, 2008.

[9] M. Moadeli, and W. Vanderbauwhede, “A Performance Model of Multicast Communication in Wormhole-Routed Networks on-Chip”, in IEEE International Symposium on Parallel&Distributed Processing, Italy, pp. 1-8, 2009, ISBN: 978-1-4244-3751-1.

[10] F. Karim and A. Nguyen. “An Interconnect Architecture for Networking Systems on Chips”, in IEEE Micro, pp. 36-45, 2002.

[11] M. Moadeli, W. Vanderbauwhede, and A. Shahrabi, “Quarc: A Novel Network 0n-Chip Architecture”, in International Conference on Parallel and Distributed Systems, pp. 705-712, 2008.

[12] Super Micro Computer, Inc. Homepage, Blade servers, URL: <http://www.supermicro.com/servers/blade/module/SBI-7226T-T2.cfm>, 2010.

[13] STMicroelectronics. URL: <http://www.st.com >, 2010.

[14] J. Duato and S. Yalamanchili, “Interconnection Networks – An Engineering Approach”, Morgan Kaufman Publishers, Elsevier Science, 2003.

[15] N. Concer, S. Iamundo, and L. Bononi, “aEqualized: a Novel Routing Algorithm For The Spidergon Network On Chip”, in Design, Automation and Test in Europe, DATE 2009, IEEE CS Press, pp. 749-754, 2009.

[16] M. Moadeli, A. Shahrabi, W. Vanderbauwhede, and M. Ould-Khaoua, “An Analytical Performance Model for the Spidergon NoC”, in 21st International Conference on Advanced Networking and Applications (AINA'07), IEEE CS Press, pp. 1014 – 1021, 2007.

[17] J. Jaroš, “Evolutionary Design of Collective Communications on Wormhole Networks”, Ph.D. thesis, Brno, CZ, 2010.

TABLE III. ACHIEVED UPPER BOUNDS ON THE NUMBER OF COMMUNICATION STEPS τ_{CC} (WH, K-PORT, DIRECT NETWORKS), SLIM NODES.

CC	OAB	AAB	OAS	AAS
6-gon_1p	3	5	5	5
6-gon_2p	2	3	3	3
6-gon_3p	2	2	2	3
8-gon_1p	3	7	7	7
8-gon_2p	2	4	4	4
8-gon_3p	2	3	3	4
10-gon_1p	4	9	9	9
10-gon_2p	3	5	5	7
10-gon_3p	2	4/3	3	6
12-gon_1p	4	12/11	11	12
12-gon_2p	3	7/6	6	9
12-gon_3p	2	4	4	9
14-gon_1p	4	14/13	13	15
14-gon_2p	3	8/7	7	13
14-gon_3p	2	6/5	5	12
16-gon_1p	4	16/15	15	18
16-gon_2p	3	8	8	17
16-gon_3p	2	5	5	17

TABLE IV. ACHIEVED UPPER BOUNDS ON THE NUMBER OF COMMUNICATION STEPS τ_{CC} (WH, K-PORT, DIRECT NETWORKS), FAT NODES.

CC	OAB	AAB	OAS	AAS
2fat_6-gon_1p	4	12	11	12
2fat_6-gon_2p	3	6	6	10
2fat_6-gon_3p	3	5	4	10
2fat_8-gon_1p	4	16	15	17
2fat_8-gon_2p	3	9	8	16
2fat_8-gon_3p	2	6	5	16
2fat_10-gon_1p	5	20	19	25
2fat_10-gon_2p	3	12	10	25
2fat_10-gon_3p	3	11	7	25
2fat_12-gon_1p	5	24	23	37
2fat_12-gon_2p	3	16	12	36
2fat_12-gon_3p	3	12	8	36
2fat_14-gon_1p	5	29	27	50
2fat_14-gon_2p	4	20	14	50
2fat_14-gon_3p	3	16	9	49
2fat_16-gon_1p	6	27	31	66
2fat_16-gon_2p	4	28	16	66
2fat_16-gon_3p	3	20	11	66

A Hardware-in-the-Loop Testing Platform Based on a Common Off-The-Shelf Non-Real-Time Simulation PC

Daniel Ulmer*, Steffen Wittel†, Karsten Hünlich† and Wolfgang Rosenstiel‡
*IT-Designers GmbH, Esslingen, Germany
Email: daniel.ulmer@it-designers.de
†Distributed Systems Engineering GmbH, Esslingen, Germany
Email: {steffen.wittel,karsten.huenlich}@distributed-systems.de
‡University of Tübingen, Department of Computer Engineering, Tübingen, Germany
Email: rosenstiel@informatik.uni-tuebingen.de

***Abstract*—The rapidly growing amount of software in embedded real-time applications such as driver assistance functions in cars leads to an increasing workload in the field of software testing. An important issue is thereby the timing behavior of the software running on the target hardware. For testing this issue, real-time capable Hardware-in-the-Loop platforms are needed. These testing platforms are mostly custom-made, proprietary and in consequence expensive. Moreover, many software developers usually have to share few instances. This paper shows an approach for a real-time capable Hardware-in-the-Loop platform based on a common off-the-shelf PC running a non-real-time operating system. Thereby, the simulation software runs on the developer's desktop computer while an extended I/O interface ensures the real-time communication with the System Under Test even for complex timing requirements as shown in an example.**

***Keywords*-Testing; Hardware-in-the-Loop; Embedded Real-Time Systems; Temporal Behavior;**

I. Introduction

Software development for embedded real-time systems, in particular closed-loop control applications in the automotive industry running on Electronic Control Units (ECUs), requires a reliable testing of the timing behavior on the target hardware. Highly frequent hardware-software integration tests of the software module under development are required, especially if the software development is done in an agile or rapid prototyping manner. These tests are normally executed on a Hardware-in-the-Loop (HiL) testing platform.

The established HiL testing platforms are usually complex devices based on proprietary hardware and software, which makes the platforms very expensive. Often these testing platforms are based on standard PC hardware in combination with an Real-Time Operating System (RTOS) and therefore operated by separate tool chains. Since these testing platforms are very complex and hence expensive, they are usually shared by several developers and are located in separate laboratories instead of being close to the developers' desk, which inhibits the rapid prototyping development cycle.

The approach introduced in this paper uses an extended, real-time capable I/O interface denominated as Real Time Adapter (RTA) designed for the usage with a non-real-time desktop computer directly at the developers' desk. The PC is used to perform the simulation models and to define the expected timing behavior while the I/O interface is responsible for keeping and observing the timing towards the System Under Test (SUT). Unlike most commercial HiL testing platforms, this approach allows to specify an arbitrary timing behavior concerning the communication to the embedded SUT. Furthermore, the approach enables the engineer to use the same software tools for function development or unit testing as well as for testing on the target hardware.

ECUs for driver assistance functions are often connected via bus interfaces to their surrounding ECUs and can therefore be stimulated by supporting the corresponding bus interface. Even ECUs communicating via analogue or digital I/O ports with their environment are mostly capable of separating their application function from the I/O interfaces by stimulating the application functions via a common communication bus. Hence a HiL platform for functional testing on the target hardware is possible for this case.

Conducted experiments and the results obtained in an industrial setting addressing the tests of embedded systems connected via the industry standard Controller Area Network (CAN) [1] show that the combination of a real-time I/O interface and standard desktop hardware are as effective as established HiL testing platforms–but in a much more efficient way–enabling a much higher test frequency.

The following two sections of this paper give an insight into the testing of interconnected ECUs and the operating principles of the RTA as an intelligent I/O device. In Section 4, a comparison relating to timing issues is done between current HiL platforms and the introduced approach based on the RTA. Section 5 finally shows an example for a test setup used in the automotive industry.

II. Test of interconnected ECUs

Significant parts of vehicle functions, especially modern driver assistance functions, are realized with the help of software. Commonly, several ECUs and their respective software contribute to implement a vehicle function that can be experienced by the driver [2].

The distribution of software in different ECUs of the vehicle requires that the ECUs are able to communicate with each other. A common widely accepted approach for interconnecting the ECUs is by sending messages on a bus system such as CAN. In order to obtain a deterministic timing behavior the majority of the messages are sent in a cyclic manner with a pre-defined cycle time as shown in Figure 1. ECU1 periodically sends its calculation results to ECU2 and vice versa. Especially for closed-loop control vehicle functions–such as an Adaptive Cruise Control (ACC) [3]–it is important to meet the given timing requirements. The ECUs usually monitor the compliance with the pre-defined cycle strictly, because a violation can result in failure, which might be life-threatening to the passengers of the vehicle.

Since the CAN bus itself is not deterministic [4], the ECU is responsible for the correct communication timing. Additionally, the priority of a CAN message is depending on its message ID. The precision of the bus timing of a certain message is hence depending on the precision of the ECUs' RTOS and on the predefined message ID. Both, the ECU and the CAN bus contribute to a deviation of the intended cycle time that can be measured on the bus. If a message is supposed to be sent with a cycle time of 20 ms, the cycle time on the bus will be not exactly 20 ms. The ECUs will tolerate such an inaccuracy as long the deviation is below a specified limit.

However, modern driver assistance functions narrow progressively the tolerance band of the allowed timing faults while the CAN bus is populated by more and more ECUs with increasing bandwidth requirements that exacerbate the situation. Seen from a testing perspective, it is thus essential that the reaction on corrupted bus timing is tested. This implies that the testing device itself is able to meet the timing requirements in the first instance and moreover to manipulate it arbitrarily.

The first integration step in the development cycle, where the timing behavior of an ECU's CAN interface can be tested, is the execution of the developed software on the target hardware. A common approach is to do this on HiL testing platforms.

Further on, it is useful having the HiL testing platform close to the software-developers' desk, especially if the ECU software is developed in an agile manner with frequent integration steps that require frequent testing on the target hardware. Commonly, different software parts for driver assistance functions are coded and tested by several developers in parallel. If these software parts are integrated into one ECU, the developers have to share the available HiL platforms. Instead of having a HiL testing platform waiting for the developer, the developer often needs to wait for the HiL platform.

III. Real Time Adapter

The RTA [5] combines the functionality of a mobile data logger and an intelligent I/O device for CAN. Its core functions [6] are implemented in VHDL to increase the execution speed and run them as parallel as possible on the built-in FPGA. In the case of the mobile data logger the CAN messages from the SUT can be stored locally on the device, whereas in the case of the I/O device the messages are transferred to an external PC and vice versa via an Ethernet connection.

As illustrated in Figure 2, the PC can process the provided information and calculate the transmit time of the response based on high precision time stamps added by the RTA to each received CAN message. Hence, a variable processing time on the PC within the tolerance range does not matter. The RTA takes care about the correct sending points of the CAN messages as well as detects timing violations caused by messages with time stamps that cannot be transmitted in time. It decouples the non-real-time behavior of the PC from the precise real-time behavior towards the SUT, which even allows the performing of complex test cases with the exact timing at each test run. The timing of each message is thereby treated separately by the RTA and thus the transmit time can be simply manipulated during a test run. Especially, this characteristic is important for test cases that validate the correctness of the SUT communication timing.

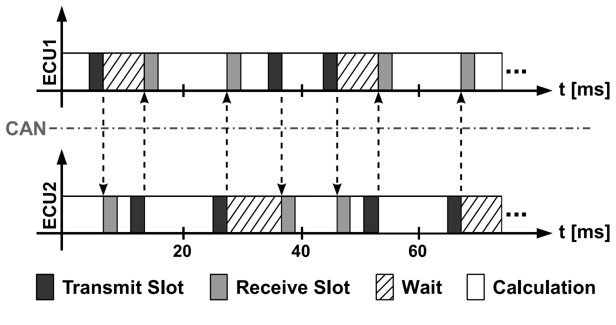


Figure 1. Cyclic communication of ECUs

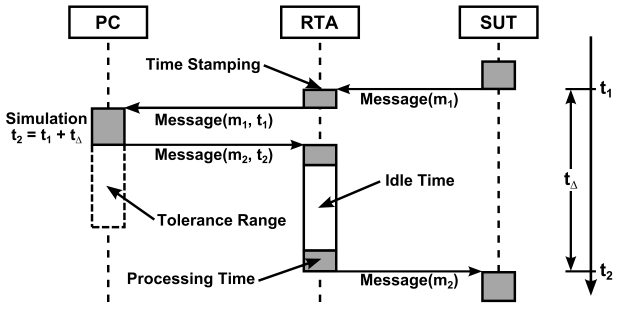


Figure 2. Sequence diagram of CAN RX/TX with an RTA

IV. HiL Platforms

In the following current HiL approaches are discussed with a closer look on their timing behavior. Additionally, the new approach that addresses the requirements for an agile usage as well as for the timing issues is introduced.

A. HiL Platform Based on an RTOS

Current HiL platforms, as they are introduced in [7], usually focus on ECU testing from a functional and non-functional perspective. This means that the testing platform covers the testing of the reaction to electrical errors as well as the test of the functions required by a driver assistance function. The approach of testing the whole test plan at only one testing platform makes this platform very complex from the hardware as well as from the software point of view. Although current solutions, as proposed by ETAS [8], are based on off-the-shelf computer hardware, they have to be expanded by several special software and hardware components needed to achieve the required functionality. One important software component is the Residual Bus Simulation (RBS), which is responsible for imitating the environment around the ECU seen from a communication point of view. If the SUT is connected via CAN buses to its surrounding components, the RBS needs to ensure the same communication behavior as established by the real environment of the SUT. For guaranteeing the real-time behavior of the CAN communication, an RTOS is used to implement the RBS for the CAN bus and the additionally required software components such as environment models. If the schedule of the RTOS is set up correctly, a precise execution of the desired CAN schedule is guaranteed within the tolerance of the RTOS.

Figure 3 shows the measured time between two CAN messages with the same message ID during a HiL test at a platform based on an RTOS [9][10][11]. According to the CAN schedule, the message is supposed to have a 20 ms cycle. The plot displays that this implementation achieves an average cycle time of almost 20 ms with a standard deviation of about 200 µs. Single outliers are reaching up to a period of 20.8 ms between two consecutive messages. In this example the measured timing still fulfils the SUT requirements.

B. HiL Platform for Functional Testing

For testing the functional behavior of different software modules running on the same ECU, it is helpful to have several testing platforms close to the developers' desks. Of course, for testing the ECUs reaction to electrical errors it is still necessary to use the complex platforms introduced before. For a quick test of a change in a hardware independent software module, HiL platforms based on a Common Off-The-Shelf (COTS) computer can be built that are connected via a CAN interface to the SUT. In this context using COTS components not only refers to hardware but additionally to software including a non-real-time Operating System (OS), typically Microsoft Windows. Additionally within the context of large companies, the IT support determines the use of virus scanners and other tools, if the PC interconnects with the corporate network. Using a standard computer means that it might also be a laptop. In this case it is easily possible to use the HiL setup within a test vehicle or while being at a field trial. Another advantage of using the standard desktop OS is that the already existing tool chain can be used to set up the HiL platform. Especially, the libraries of environment models for Model-in-the-Loop (MiL) and Software-in-the-Loop (SiL) simulations from earlier integration steps of the driver assistance function can be reused without the need of being ported to an RTOS environment.

Figure 4 shows the time between two consecutive CAN messages of the same message ID during a HiL test on such a platform without an RTOS. The plot displays that the implementation based on a COTS computer and a CAN interface achieves an average cycle time of 20 ms with

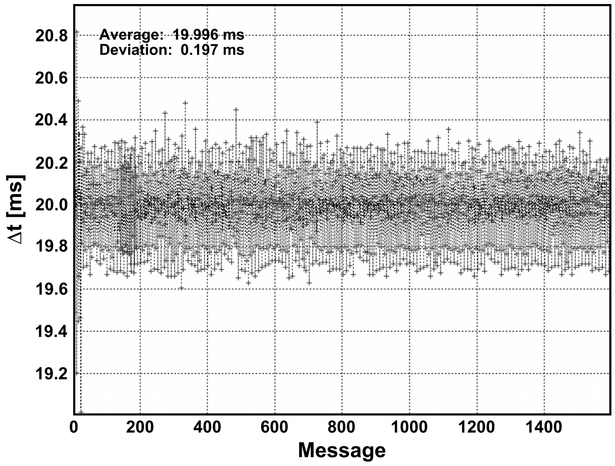


Figure 3. RTOS HiL – 20 ms cycle time message

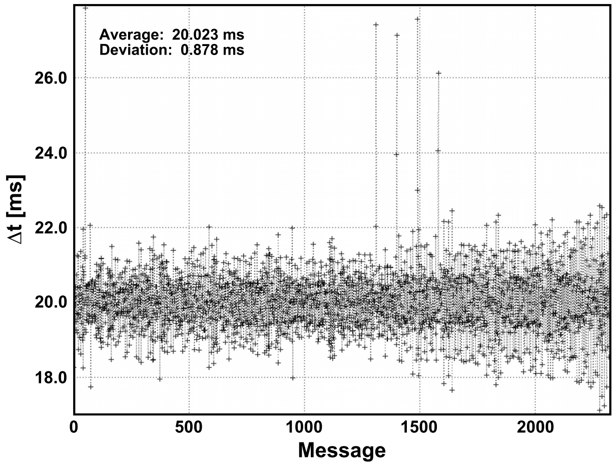


Figure 4. Non-real-time OS HiL – 20 ms cycle time message

a more than fourfold standard deviation of approximately 900 µs. Getting worse, in this case outliers of up to 8 ms can be seen. This approach only works, if the ECU tolerates such outliers.

A major drawback of this approach is that the environment models and the RBS have to be either implemented on the OS of the COTS computer or at least the RBS has to be shifted to the CAN interface. In the first case, the timing behavior of the RBS is depending on the timing behavior of the non-real-time OS. In the latter case, a separate tool chain is necessary to implement the RBS on the CAN interface. This leads to a fix communication schedule, which can be only manipulated at runtime if a complex handshake between the PC and the RBS is set up. If the implementation of the timing supervising software within the ECU is not too strict, the first approach works in practical use.

C. HiL Based on a COTS Computer and an RTA

Since the timing requirements of the ECUs tighten and the implemented driver assistance functions require more and more precise data at an accurate point in time, the timing behavior shown in Figure 4 is not acceptable anymore. Additionally, if the implemented function, e.g., for interacting vehicles [12], is not only depending on the data value but also on its arrival time, the targeted testing of the reaction on certain bus timing becomes necessary. A HiL platform, which solves the timing issues while leaving the RBS on the COTS computer (PC), is introduced in [5] and [6]. The approach leaves it up to the PC to define the intended sending time of a message. This time stamp is then handed over together with the payload to the RTA. While the computer is responsible for calculating timing and content, the RTA precisely plans, executes and supervises the desired timing. If for any reason the desired timing cannot be kept within a certain tolerance, the RTA informs the simulation software on the computer. It is then up to the simulation application to repeat the test case. Thereby, an upper limit prevents the HiL testing platform from repeating the same test case too often.

Timing violations usually originates from the non-real-time OS on the PC in combination with the time consuming or concurrent execution of programs during a test run, e.g., anti-virus scanners, mail software or automatic update clients. The test implementation itself and the resources consumption associated with it also affect the timing. Test cases, which need more processing time on the PC for one simulation step as the expected cycle time of the SUT, are not suitable to be executed on this platform.

Figure 5 shows the result of the introduced solution for a current ECU with driver assistance functions. The intended cycle time of 20 ms is kept with a standard deviation of 5 µs. Even the outliers, which occur in this case due to the occupied bus, are less than 40 µs.

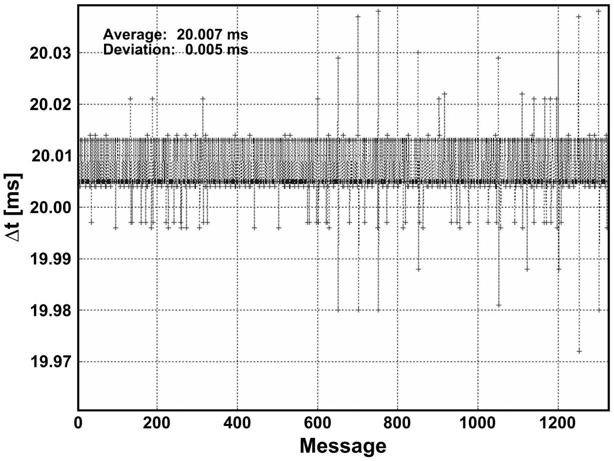


Figure 5. RTA HiL – 20 ms cycle message

The performance of the HiL testing platform primarily depends on the COTS hardware used to set up the platform, which determines the test case limitations in terms of timing. Practical experiences with a prototypical implementation show that approximately one of 1000 test cases has to be repeated. Moreover, measurements during the evaluation revealed an average pass through time of about 4 ms to receive a CAN message from the SUT and send the response back. The time also includes the calculation of a common test step within the simulation on the PC. In the example this means that the HiL testing platform has roughly 16 ms at a cycle time of 20 ms to compensate outliers occurred during the performing of a test case. Based on these obtained results the outliers are not an exclusion criterion for the use in a production environment, because they are detected and reported by the RTA.

V. Example

An example for a test setup used in the automotive industry is displayed in Figure 6, which comprises of a standard PC with an RTA as well as of the SUT itself consisting of two CPUs that are connected to the same clock oscillator. Thereby, the PC and the RTA are used to simulate the ECU's environment. For safety critical reasons, some applications within the ECU are tested on module level embedded into the final hardware. The *Communication CPU* has two tasks with 20 ms cycle time. On the one hand, it

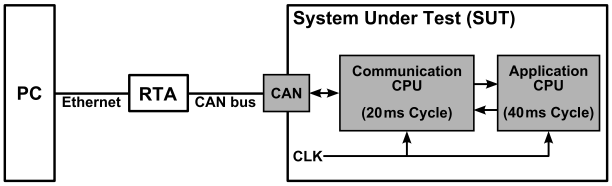


Figure 6. Example for a test setup

implements the bus communication that consists of receiving and transmitting messages and updating the internal signal database. On the other hand, this CPU is used to validate the results of critical functions running on the *Application CPU*. The *Application CPU* runs at 40 ms cycle time and is responsible for processing the implemented driver assistance functions.

One software module running on the *Application CPU* implements a safety critical requirement. In this example we assume that a sensor sends a signal denominated *Object Type*. This signal is specified to be zero for two CAN cycles and four for the following three CAN cycles, if the sensor is faulty. The safety critical requirement of the software module is to detect this situation and to prevent a driver assistance function from interfering. The correct implementation of the software module is to be tested on the target hardware and hence at a HiL platform. Since the clocks of the SUT and the HiL platform are independent, it cannot be guaranteed that the sequence is received correctly at the SUT's internal data interface. To achieve reproducible test results, it is necessary to synchronize the testing platform with the SUT. The synchronization mechanism is shown in Figure 7. Some *Application Results* are handed over from the *Application CPU* to the *Communication CPU*, which transmits the corresponding CAN message on the CAN bus. The RTA delivers this message together with a receive time stamp to the PC running the environment simulation. After calculating the simulation environment model, the result is handed over to the RTA for being sent 41 ms ahead in time. This ensures that the result is available for the *Application CPU* right before a new application cycle begins.

Listing 1 shows a pseudo code sample for an implementation on a PC based platform for functional testing. Since the sending time of the message is in this case depending on the scheduling of the *TransmitThread* of the non-real-time OS, it cannot be guaranteed that the sequence is sent as specified.

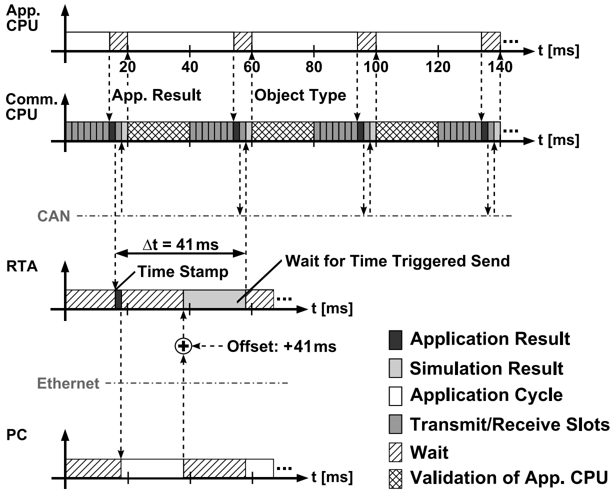


Figure 7. Synchronization of the testing platform with the SUT

```
WHILE(NOT quit)
BEGIN

  // Receive CAN Message
  Receive(in_message)

  // Calculate Environment Simulation Model
  out_message = CalcEnvModel(in_message)

  // Calculate Output Message Time Stamp
  time_stamp = in_message.time_stamp + 41

  // Transmit CAN Message using the Windows
  // Event Timer in a separate Thread
  TransmitThread(out_message, time_stamp)

  // Wait until next Cycle
  WaitForNextWindowsTimeEvent()

END
```

Listing 1. Standard PC synchronization mechanism

Listing 2 illustrates that in case of an RTA based HiL testing platform the precise sending of the message is done by the RTA and therefore independently of the OS timing deviations. In the worst case, a message is sent too late to the RTA and the test case is then being declared invalid and repeated.

```
WHILE(NOT quit)
BEGIN

  // Receive CAN Message
  RTA_ReceiveMessage(in_message)

  // Calculate Environment Simulation Model
  out_message = CalcEnvModel(in_message)

  // Calculate Output Message Time Stamp
  time_stamp = in_message.time_stamp + 41

  // Transmit CAN Message to RTA
  RTA_TTS_TransmitMessage(out_message, time_stamp)

  // Wait until next Cycle
  RTA_WaitForNextCycle()

END
```

Listing 2. RTA synchronization mechanism

Figure 8 shows the results achieved on the CAN bus with a bus load of 60 % and a cycle time of 20 ms for the CAN messages. The sequence of two cycles zero and three cycles four is precisely executed. In the project context we have implemented this testing challenge on the RTA based HiL platform since this platform is available at every developers' desk and the modification of the existing simulation code has been limited to adding a constant offset to the time stamp of an incoming message. We have decided against an implementation on an RTOS HiL since there is only one instance available, which can either be used for implementing new features or for running tests. Synchronizing the time

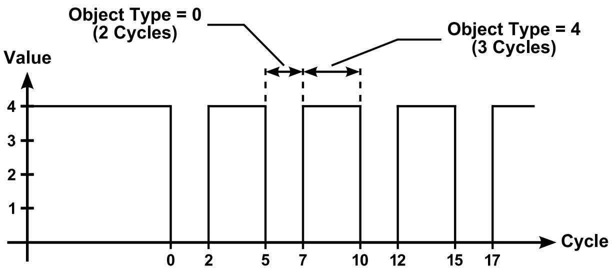


Figure 8. CAN trace for cyclic stimulation

slice based RTOS to the SUT would have meant to change the complete simulation kernel and therefore several days of implementation work.

VI. Conclusion and Future Work

The measurements demonstrate that it is possible to implement a HiL testing platform fulfilling the timing requirements of modern driver assistance functions and the requirements of an agile or rapid prototyping development process within the automotive industry. It has also been shown that current testing platforms address one of these aspects while the RTA approach addresses both. It has also been argued that the achieved timing on the CAN bus of the RTA based HiL platform is more precise than the timing of the RTOS HiL. It is left for future work to study the advantages of the RTA approach in terms of the definition and flexible manipulation of the timing behavior, e.g., for deterministic robustness tests of the function software. One aspect might be the modeling of a statistic temporal distribution where the parameters can be influenced by random testing or by evolutionary testing. Additionally, the RTA approach might be used as a cost efficient HiL setup for a continuous integration tool chain for embedded software development. Due to the usually large number of variants on the level of hardware-software integration, a high test volume must be considered here. However, each test can be executed at maximum in real-time for each variant. This means that for quick results many parallel HiL platforms are necessary. The price efficient HiL testing platform based on the RTA is a necessary step to implement this idea.

References

[1] ISO, *ISO 11898-1:2003: Road vehicles — Controller area network (CAN) — Part 1: Data link layer and physical signalling*. International Organization for Standardization, 1993.

[2] C. Marscholik and P. Subke, *Road vehicles - Diagnostic communication: Technology and Applications*. Hüthig, 2008.

[3] Daimler AG, “The challenge of accident prevention”, *Milestones in Vehicle Safety. The Vision of Accident-free Driving*, 2009.

[4] K. Etschberger, *Controller Area Network. Basics, Protocols, Chips and Applications.* IXXAT Automation, 2001.

[5] IT-Designers GmbH, “RZA”, Access Date: November 10, 2010. [Online]. Available: http://www.it-designers.de/RTA

[6] D. Ulmer, A. Theissler, and K. Hünlich, “PC-Based Measuring and Test System for High-Precision Recording and In-The-Loop-Simulation of Driver Assistance Functions”, in *Proceedings of the Embedded World Conference*, 2010.

[7] C. Marscholik and P. Subke, *Datenkommunikation im Automobil*. Hüthig, 2007.

[8] ETAS GmbH, “LABCAR System Components - ETAS Products”, Access Date: November 10, 2010. [Online]. Available: http://www.etas.com/en/products/labcar_system_components.php

[9] ETAS GmbH, “LABCAR-RTPC Real-Time Simulation Target for HiL Testing”, Access Date: November 10, 2010. [Online]. Available: http://www.etas.com/en/products/labcar_rtpc.php

[10] G. Wittler and J. Crepin, “Real-time and Performance Aspects of Hardware-in-the-Loop (HiL) Testing Systems”, *ATZonline*, 2007.

[11] J. Kiszka, “Xenomai: The RTOS Chameleon for Linux”, Real-Time Systems Group, Leibniz Universität Hannover, Tech. Rep., 2007.

[12] D. Ulmer and A. Theissler, “Application of the V-Model for the development of interacting vehicles and resulting requirements for an adequate testing platform”, in *Proceedings of the Software and Systems Quality Conferences*, 2009.

Novel Modulo 2^n+1 Subtractor and Multiplier

Dina Younes
Department of Microelectronics
Brno University of Technology
Brno, Czech Republic
e-mail: xyoune00@stud.feec.vutbr.cz

Pavel Steffan
Department of Microelectronics
Brno University of Technology
Brno, Czech Republic
e-mail: steffan@feec.vutbr.cz

***Abstract*— This paper introduces a novel design of modulo 2^n+1 subtractor in Residue Number System (RNS). A novel design of modulo 2^n+1 multiplier has also been introduced by utilizing the presented subtractor. The two designs are suitable for small values of n, as they depend on the normal (binary) representation of RNS numbers, instead of diminished-one representation which has difficulties in representing zero operands and results, and need to deal with them separately. The presented circuits were implemented and simulated using VHDL to prove the theoretical consideration.**

Keywords-Residue number system; modulo 2^n+1; subtractor; multiplier

I. INTRODUCTION

The residue number system (RNS) is a non-weighted integer number system [1], which decomposes binary large numbers to smaller residues. Obviously, there is no carry propagation problem between residues [2]. RNS offers the potential for high-speed and parallel arithmetic. The RNS based on the moduli set (2^n-1, 2^n, 2^n+1) is most frequently utilized to achieve a high-performance RNS application since the resulting RNS architecture performs fast residue arithmetic.

The 2^n+1 modulo arithmetic gains the most significant importance because modulo 2^n+1 channel is critical for whole system in terms of area and delay. It is considerably more difficult than 2^n-1 modulo, in the sense that it cannot be realized with the same speed or efficiency [3]. The representation of a number in modulo 2^n+1 arithmetic has to be (n+1) bit long instead of n-bit, so that the whole system will require blocks that are (n+1) bits long. Therefore 2^n+1 adders and multipliers became of interest to many researches.

In many publications about modulo 2^n+1 adders and multipliers [4][5][6], the diminished-one representation of residues has been used in order to solve the problem of (n+1) bit long operands and use only n-bit long operands. As a result, only n-bits are used in the computation units. But this representation has some difficulties in representing zero operands and results as it has to be treated separately. However, by using small values of n, the normal representation of residues can be used without causing a considerable effect on the overall delay of the system.

Small values of n will produce small dynamic ranges. Small dynamic ranges are usually less than 15 bits. The presented design is very attractive for many digital signal processing applications that use small dynamic ranges. An example of such application is the [0,255] range grayscale image data [7]. Or even RGB images with 8 bit format. For this application, a small moduli set {7,8,9} will be enough for the encoding. The dynamic range of this set is small (9 bits).

RNS multipliers are widely used in many DSP applications, such as image processing, FIR filters and Fourier transform.

In this paper, a novel design of modulo 2^n+1 subtractor is presented, and it has been used in designing modulo 2^n+1 multiplier. These designs are suitable for small values of n, which is favorable from a practical point of view because the overall speed of system will be increased [8].

The proposed circuits were implemented and simulated using VHDL.

The organization of this paper is as below:

In Section 2, an overview of RNS arithmetic is given. The design of the proposed subtractor is shown in Section 3. In Section 4, the proposed multiplier and a description of its circuit are presented. The results and a comparison with an existing 2^n+1 multiplier are stated in Section5. Finally, the conclusion is discussed in Section 6.

II. RNS OVERVIEW ARITHMETIC

The primary advantage of RNS is that addition, subtraction, and multiplication can be performed independently and in parallel on the various residues.

The residual number system (RNS) is defined by a set of numbers m_1, m_2, ..., m_n called the moduli. Where the great common divider (GCD) for m_i , m_j = 1. In this system, an integer X is represented by an ordered set of residues, {x_1, x_2, ..., x_n} where:

$$x_i = X \; mod \;\; m_i \tag{1}$$

Where X: $$0 \le X < \prod_{i=1}^{N} m_i$$

Arithmetic operations (addition, subtraction and multiplication) are performed totally parallel on those residues.

Assuming that A and B are two RNS numbers; the addition (subtraction) of these two numbers is given by:

$$A \pm B = \left\{ \left|a_1 \pm b_1\right|_{m_1}, \left|a_2 \pm b_2\right|_{m_2}, \ldots, \left|a_N \pm b_N\right|_{m_N} \right\} \quad (2)$$

Also the multiplication of *A* and *B* is given by:

$$A \times B = \left\{ \left|a_1 \times b_1\right|_{m_1}, \left|a_2 \times b_2\right|_{m_2}, \ldots, \left|a_N \times b_N\right|_{m_N} \right\} \quad (3)$$

The strongest points in this system are the independency, and carry free among the residues. In other words, each residue can be treated as a separated integer.

III. Proposed RNS Subtractor

Subtraction is an operation widely met in digital signal processing applications [9] [10] for operations such as mean error estimation, mean square error estimation and calculation of sum of absolute differences. Since modulo arithmetic is also frequently used in these types of applications, efficient modulo subtraction circuits are welcome. However, very little work [11] has been focused on designing modulo 2^n+1 subtractors.

To design modulo 2^n+1 subtractor, a binary subtractor, a binary adder and a multiplexer have been used, as shown in Fig. 1. This simple structure and small number of elements used in the circuit provide less delay and more efficiency to accomplish subtractive operation.

The operands used are (n+1) bit long, because they are residues resulted from binary to RNS conversion with respect to modulo 2^n+1. The output of the proposed subtractor is also (n+1) bit long.

For modulo m=2^n+1, the subtraction of two residues is defined as:

$$C = \left|A - B\right|_{2^n+1} = \begin{cases} \left|A - B\right|_{2^n+1} & \mathit{if} \;\; A - B \geq 0 \\ \left|A - B + \left(2^n + 1\right)\right|_{2^n+1} & \mathit{if} \;\; A - B < 0 \end{cases} \quad (4)$$

The second case is implemented by adding 1 to the result of subtraction [12].

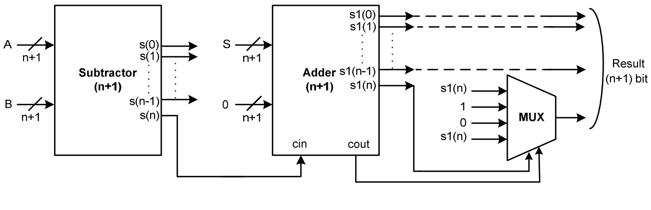


Figure 1. Proposed modulo 2^n+1 subtractor

However dealing with operands of (n+1) bit long is not same as dealing with just n-bit, because the MSB (2^n bit) will create some confusion, especially when *A<B*. To overcome that confusion, a multiplexer has been added to correct the final output of the proposed subtractor.

For example: n=4, m=2^n+1=17, A=0, B=1
$|A\text{-}B|_{2n+1} = |0\text{-}1|_{17} < 0$

```
 0:     0 0000
-1:     1 1111 –
        1 1111
             1 +
        0 0000
```

By adding ‘1’ to correct the result, we got “0 0000”, instead of the correct result “1 0000”. Therefore the multiplexer was added to overcome cases like this one. The carry out after the addition is ‘1’ and the MSB is ‘0’, so according to the multiplexer, the MSB will become ‘1’, and the final output will become “1 0000”.

IV. Proposed RNS Multiplier

Let $A=a_n \ldots a_0$ and $B=b_n \ldots b_0$ refer to two (n+1) bit modulo 2^n+1 operands, such that $0 \leq A,B \leq 2^n$. The multiplication of *A* and *B* is given by:

$$R = A \times B \quad (5)$$

Assuming that *A*, *B* are expressed using (n+1) bits:

$$R = \sum_{i=0}^{2n} r_i 2^i \quad (6)$$

Where r_i is the *i*th bit resulting from *AB*:

$$R = \left| \sum_{i=0}^{n-1} r_i 2^i + \left|2^n\right|_{2^n+1} \sum_{i=n}^{2n} r_i 2^{i-n} \right|_{2^n+1} \quad (7)$$

$$R = \left| X + 2^n Y \right|_{2^n+1} \quad (8)$$

An important aspect in RNS is:

$$\left|2^n\right|_{2^n+1} = \left|2^n + 1 - 1\right|_{2^n+1} = \left|-1\right|_{2^n+1} \quad (9)$$

Substituting equation (9) in (7):

$$R = \left| \sum_{i=0}^{n-1} r_i 2^i - \sum_{i=n}^{2n} r_i 2^{i-n} \right|_{2^n+1} \quad (10)$$

$$R = \left| X - Y \right|_{2^n+1} \quad (11)$$

The presented modulo 2^n+1 multiplier circuit is shown in Fig. 2. It consists of a binary multiplier (n+1)×(n+1), and a modulo 2^n+1 subtractor. The output of the multiplier is separated into two (n+1) bit operands, and fed into modulo 2^n+1 subtractor.

As noticed from equation (11), to multiply two modulo 2^n+1 numbers, a modulo 2^n+1 subtractor is needed.

The presented modulo 2^n+1 subtractor is used, but the operands of that subtractor are (n+1) bit long, therefore we just made the MSB of X = ‘0’; so both operands become (n+1) bit long, and can be applied to the subtractor to acquire the correct result.

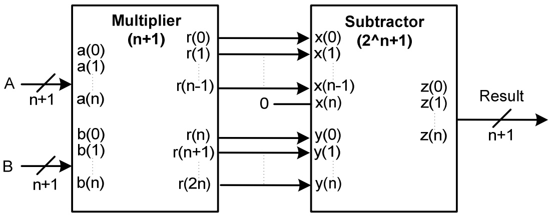


Figure 2. Proposed modulo 2^n+1 multiplier

V. Implementations and Results

The presented modulo 2^n+1 multiplier has been implemented and simulated using VHDL. A comparison of this multiplier with an existing one [13] has been done. Table 1 shows the delay comparison between the two designs. As can be noticed that for small values of n, the delay in the presented multiplier is considerably smaller than the one stated in [13], and that makes it more efficient for applications requiring small dynamic ranges, such as image processing of images with 8 bit format. But as the value of n increases, the delay also increases. We have noticed that at the value of n=12 the delay in the multiplier stated in [13] becomes less.

Table 1. The Delay Comparison between the Presented Multiplier and the Multiplier Stated in [13]

n	[13]	The presented multiplier	Delay reduced
3	20.2 ns	15.2 ns	24.75 %
4	25 ns	18.3 ns	26.8 %
8	31 ns	27.5 ns	11.3 %
10	35.6 ns	32.6 ns	8.4 %
11	36 ns	34.9 ns	3.1 %
12	36.2 ns	36.8 ns	-
14	40.6 ns	43.7 ns	-

VI. Conclusion

Novel simplified architectures of 2^n+1 modulo subtractor and multiplier have been presented and detailed in this paper. To realize modulo 2^n+1 subtractor, a binary subtractor, a binary adder and a multiplexer have been used. The proposed subtractor has been used to realize modulo 2^n+1 multiplier. RNS residues were presented using binary representation instead of diminished-one representation that has been used recently. This representation acquires additional components to solve the difficulties resulted in zero representing. The main advantages of the proposed circuits are design simplicity, reduced computation complexity and reduced delay in the system when using small values of n. These circuits can be effectively used in digital signal processing applications that require a small dynamic range, such as image processing of images with 8 bit format, where the range of image data is [0,255]. The effectiveness of these architectures has been proved by presenting a comparison with an existing 2^n+1 modulo multiplier. This comparison has been done using VHDL implementation and simulation of the design.

Acknowledgment

Funding for this work was provided by Czech Ministry of Education in the frame of Research Program MSM0021630503 MIKROSYN New Trends in Microelectronic Systems and Nanotechnologies and ENIAC JU Project No. 120001 Nanoelectronics for an Energy Efficient Electrical Car and ARTEMIS JU Project No. 100205 Process Oriented Electronic Control Units for Electric Vehicles Developed on a multi-system real-time embedded platform.

References

[1] M. A. Sonderstrand , “Residue Number System Arithmetic”. Modern Applications in Digital Signal Processing, New York: IEEE Press, 1986.

[2] Szabo and R. Tanaka, “Residue arithmetic and its applications to computer technology”. New York, McGraw-Hill, 1967.

[3] A. Omondi and B. Premkumar, “Residue Number Sustem Theory and Implementation”. Imperial College Press, 2007.

[4] H. T. Vergos, C. Efstathiou, and D. Nikolos, “Diminished-One Modulo 2^n+1 Adder Design,” *IEEE Transactions on Computers*, vol. 51, no. 12, pp 1389-1399, December 2002.

[5] H.T. Vergos, and D. Bakalis, “On the Use of Diminished-1 Adders for Weighted Modulo 2^n+1 Arithmetic Components,” In *Proc. Euromicro Conference on Digital System Design*, pp. 752-759, 2008.

[6] C. Efstathiou, H. T. Vergos, G. Dimitrakopoulos and D. Nikolos, “Efficient diminished-1 modulo 2^n+1 multipliers,” *IEEE Transactions on Computers*, vol. 54, no. 4, pp. 491–496, April 2005.

[7] Wei Wang, M.N.S. Swamy, and M.O. Ahmad, “RNS Application for Digital Image Processing”, In *System-on-Chip for Real-Time Applications, 2004.Proceedings. 4th IEEE International Workshop on*, pp. 77-80, July 2004.

[8] B. Parhami, “Computer arithmetic: algorithms and hardware designs”, Oxford, 2002.

[9] P. G. Fernandez, and A. Lloris, “RNS-based implementation of 8x8 point 2D-DCT over field-programmable devices,” *Electronics Letters*, vol. 39, no. 1, pp. 21-23, January 2003.

[10] P.M. Matutino, and L. Sousa, “An RNS based Specific Processor for Computing the Minimum Sum-of-Absolute-Differences,” In *Proc. Euromicro Conference on Digital System Design*, pp. 768-775, 2008.

[11] S. Timarchi, K. Navi, and M. Hosseinzade, “New Design of RNS Subtractor for modulo 2^n+1,” In *Proc. Int. Conference on Information and Communication Technologies*, pp 2803-2808, 2006.

[12] A. Hiasat, “New Memoryless, Mod *(2^n - 1)* Residue Multiplier”, *Electronics Letters,* vol. 28, no. 3, pp. 314-315, January 1992.

[13] R. Zimmermann, “Efficient VLSI implementation of modulo (2^n±1) addition and multiplication”, In *Computer Arithmetic, 1999. Proceedings. 14th IEEE Symposium on*, pp. 158-167, 1999

Real-time Component Labelling with Centre of Gravity Calculation on FPGA

Abdul Waheed Malik, Benny Thörnberg
Department of Information Technology and Media
Mid Sweden University,
Sundsvall, Sweden
{waheed.malik | benny.thornberg}@miun.se

Xin Cheng, Najeem Lawal
Department of information Technology and Media
Mid Sweden University,
Sundsvall, Sweden
{xin.cheng | najeem.lawal}@miun.se

Abstract—**In this paper, we present a hardware unit for real time component labelling with Centre of Gravity calculation. The main targeted application area is light spots, used as references for robotic navigation. Centre of Gravity calculation can be done in parallel with a single pass component labelling unit without first having to resolve merged labels. We present a hardware architecture suitable for implementation of this Centre of Gravity unit on Field Programmable Gate Arrays. As a result, we get high frame speed, low in power and low in latency design. The device utilization and estimated power dissipation are reported for the Xilinx Virtex II pro device simulated at 86, Video Graphics Adaptor (VGA), sized frames per second. Maximum speed is 410 frames per second at 126 MHz clock.**

Keywords- Centre of Gravity (COG); Component Labelling; Position Measurement

I. INTRODUCTION

Object detection and measurement of an object's position are important aspects for many image processing problems and machine vision systems. In medical image processing, we can see it in marker recognition and leukocyte tracking [1]. We can see the same requirement for automatic target recognition delineation and for light spots used as references for robot navigation [2]. In all those applications, it is really important to measure the object's positions correctly. When using light spots for robotic navigation, it is also necessary for the processing of video frames to be fast. High frame speed and low latency becomes important performance measures. Field Programmable Gate Arrays (FPGAs) are the preferred computational platform to reach this high performance. FPGAs offer massive parallelism, on-chip memories and arithmetic units and are therefore found to be most suitable for front end video processing [5]. This has motivated us to develop a FPGA based hardware unit for component labelling and COG calculation.

This is the most aggressive implementation of component labelling with feature calculation with minimum latency. It is shown that calculation of COG can be done without first resolving the complex chains of labels. The resolving of labels is done after the whole frame is labelled, avoiding any dependency on horizontal synchronization as described in [6], [10], thus maximizing the frame speed. Only one pass labelling is used as compared to classical two pass labelling, and no need to store image as described in [9]. Thus, the latency for our implementation is much lower than the previous implementations.

The developed architecture is suitable for smart cameras with a built in computational platform and having a low bandwidth output communication channel [3].

The smart camera only sends the processed and refined information, rather than sending large amount of video data. Machine vision algorithms are often divided into the following steps [4]. Video is acquired from the image sensor at Image acquisition. Image objects are extracted from the pre-processed video data at Segmentation, as shown in Figure 1A. During labelling, pixels belonging to the same image component are assigned a unique label. At Feature extraction an image component is described, for example in terms of region features such as area, ellipse, square or circle parameters. Components can also be described in terms of gray value features such as mean gray value or position. This feature information can then be used for Classification of image components. Information about recognized objects in the camera's observation area can be transmitted to the camera output using typically a very low bandwidth.

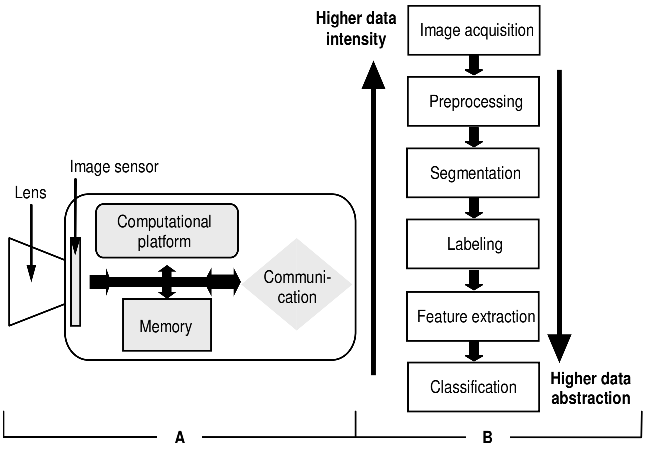


Figure 1. A) Smart camera. B) Fundamental steps of machine vision

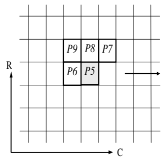


Figure 2. Neighbourhood for eight connectivity labelling.

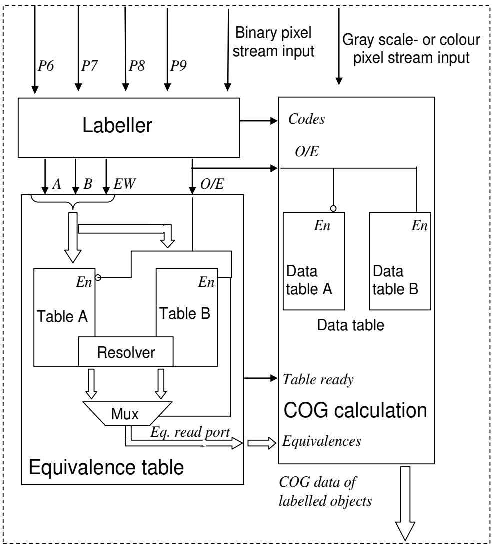


Figure 3. Kernel for labelling and COG computation

In this work, we focus on the COG calculation along with the image component labelling. Eight connectivity labelling process has been used, and labels are resolved after every frame. As shown in Figure 2, P5 is the current pixel for labelling, and it is labelled depending on labels P6 to P9. If the labeller finds two different labels in the neighborhood, it assigns the smaller label to the current pixel and marks that those two labels are equivalent and should be merged.

The equivalent labelled pair (A, B) is sent to the Equivalence table for merging. Label merging is targeted to either Table A or B, depending on odd or even frames (O/E) [7]. Resolving of linked lists of labels is thus frame interleaved with the labelling process. The architecture is shown in Figure 3. We assign a label to the current binary pixel depending on the previously labelled pixels from P6 to P9. At the same time as we are assigning a label to the current pixel, we are also accumulating pixel data in Data table A or B for the COG calculation without first having to resolve linked lists of labels. This method has previously been analyzed for COG in [7]. We have found published results on high speed, low latency hardware component labeller [6], but no results on how such a labeller can be efficiently combined with calculation on image component features such as COG.

The COG calculation method will be discussed in Section II. We discuss hardware architecture in Section III, while memory requirement is discussed in Section IV. Performance and device utilization are discussed in Section VI.

II COG Calculation

The COG for an image component *O* in an image is calculated as

$$(r_o,c_o)=\frac{\sum_{r_i,c_i\in O} r_i I(r_i,c_i)}{\sum_{r_i,c_i\in O} I(r_i,c_i)},\frac{\sum_{r_i,c_i\in O} c_i I(r_i,c_i)}{\sum_{r_i,c_i\in O} I(r_i,c_i)} \qquad (1)$$

In the Equation 1, (r_o,c_o) is the mass centre of the object. $I(r_i,c_i)$ is intensity of pixels and r_i, c_i are the row and column count respectively. Let us assume that an object region is divided into two sub regions, *S* and *T,* as a result of the first pass of labelling. *S* and *T* belonging to the single object $S \cup T$ are resolved at the end of first pass and in parallel with labelling of the next frame. The numerator and denominator according to Equation 2 are accumulated in to *Data table A* or *B,* at the same time as the first labelling pass. At the end of the first labelling pass, when the codes are resolved in equivalence table, the data stored in *Data table A* or *B* will be merged into numerator and denominator for the region $S \cup T$. This data merging for regions *S* and *T* is illustrated in Equation 2 for row the dimension. Thus we add the numerator data from different codes of same image component, and the same is true for the denominator. The subsequent step will be to perform the final division to conclude the COG computation. Since, COG is computed in parallel with the first pass of the image component labelling, there is no need for a second labelling pass [6][7].

$$r_o{,}=\frac{\sum_{r_i,c_i\in S} r_i I(r_i,c_i)+\sum_{r_i,c_i\in T} r_i I(r_i,c_i)}{\sum_{r_i,c_i\in S} I(r_i,c_i)+\sum_{r_i,c_i\in T} I(r_i,c_i)}=\frac{N}{D} \qquad (2)$$

It can be seen that we perform the multiply and accumulate (MAC) operation along with the labelling process depending on different codes. Before applying the COG algorithm to a video frame, objects of interest must first be separated from the background at an image segmentation step [4]. This image segmentation is at its simplest form a threshold applied globally on the grey levels of the image.

III. PROPOSED ARCHITECTURE

In this section, we will describe the hardware architecture for calculating the COG from labelled pixel data. The *COG calculation* shown in Figure 3 is further divided into two main modules: *Sequencer* and *Multiply & Accumulate (MAC) unit.* As shown in Figure 4, The *Sequencer* is controlling all the data merging and computation and is also responsible for sending computed COG data to an arbitrary communication controller. The *MAC unit* accumulates the nominators and denominators and performs serial divisions for final COG output. This accumulation of pixel data in the *MAC unit* is done in parallel with the labelling process. When the equivalence table is resolved, the *Table ready* signal becomes active and the sequencer starts merging data from regions belonging to same image objects. The sequencer is able to do this data merging based on the resolved equivalences read from the equivalence table, as shown in Figures 3 and 4. After merging, the numerators and denominators are ready for the division. At completion of the division, a *Compute done* signal is sent back to *Sequencer,* which then enables *Feature Strobe* to send the COG value to the communication device. In our experiment, we used an asynchronous serial port. The sequencer scans the equivalence table for all image objects until all objects are computed and transmitted. Two data tables, *A* and *B*, are used for numerators and denominators, as shown in Figure 5.

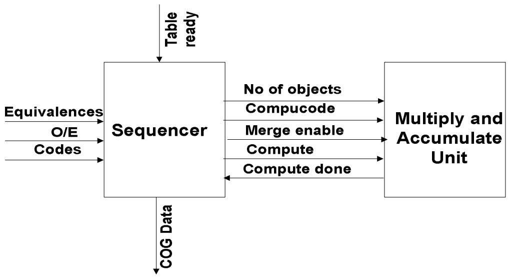


Figure 4. Architecture for COG computation.

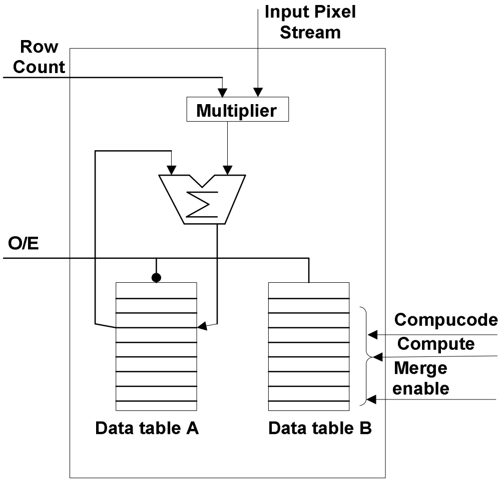


Figure 5. MAC unit for row centre computation.

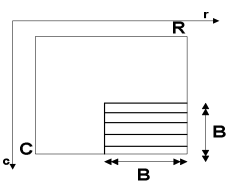


Figure 6. Worst case scenario for storage requirement.

One table accumulates the values after MAC operations and at the same time the other table is used for merging the data used for COG calculation. It will be shown in the next section, the memory storage requirement for the data tables are dependent on the maximum allowable object size, the maximum number of objects, and maximum pixel intensity.

IV. STORAGE REQUIREMENT

In this section, we will analyse the memory storage requirement for the COG calculation. According to Section II and Equation 2, numerators N and denominators D are accumulated in memory storage. The worst case scenario will be when a large image component of size BxB pixels is present at the right down most corner of the video frame, as shown in Figure 6. The frame size is R, C number of rows and columns. The maximum pixel intensity Imax is assumed to be a power of two, and we assume that the worst case object has maximum intensity for all its pixels. First, we calculate the maximum integer value I_{Num} for numerator row centre assuming that the row dimension is the largest. From Equation 2 we can conclude that,

$$I_{Num} \le \sum_{r=R-B+1}^{R} \sum_{c=C-B+1}^{C} rI_{\max} = BI_{\max} \sum_{r=R-B+1}^{R} r \qquad (3)$$

,

$$\sum_{r=R-B+1}^{R} r = (R-B+1+R-B+2+\ldots+R-B+B) \qquad (4)$$

$$\sum_{r=R-B+1}^{R} r = B(R-B) + \sum_{r=1}^{B} r \qquad (5)$$

$$\sum_{r=R-B+1}^{R} r = B(R-B) + \frac{B\cdot(B+1)}{2} \qquad (6)$$

Substituting Equation (6) into log2 of Equation (3) gives,

$$S_{Num} = \log_2\left(B^2 I_{\max}\left(R + \frac{1-B}{2}\right)\right) \qquad (7)$$

S_{Num} is thus the number of bits required to store one single numerator. For our experiments, we have used the following values for R, C, B and Imax: R=640, C=480, B=170, Imax=255. For these values, $S_{Num} \le 32$. This means that for the maximum image component of size 170x170, we need 32 bits for accumulating the numerator in one single memory cell. For the maximum allowable number of labels, L=1024, we then conclude that two block RAMs with word length of sixteen bits and depth of memory equal to L are needed. As we use interleaving for memory access, a total of four block RAMs are required. The above expression for row centre is also valid for column centre, so four block RAMs also required for column centre. Equation 9 shows the storage requirement for the denominators.

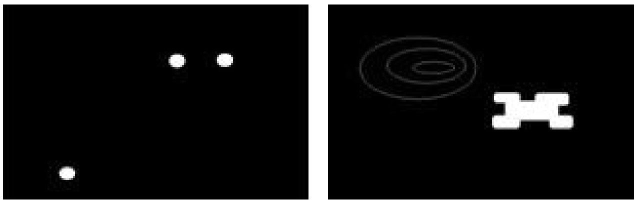


Figure 7. Input stimuli

TABLE I. DEVICE UTILIZATION AND POWER CALCULATION

Slices	RAMB16	Slice flip-flop	4 input LUT
3876 (28%)	17 (12%)	936 (3%)	7589 (27%)

a)

Dynamic power(mW)	Clock	Signals	Logic	IOs	Mult
	30	0.5	0.2	0.6	0.7
Dynamic Power	32.0 mW (at 27Mhz clock)				
Quiescent Power	103 mW				
Total Power	135 mW (at 27MHz clock)				
Frequency	Max 126 MHz				
Latency	313584 clock cycles 11.6 ms (at 27 MHz clock and 86 fps)				

b)

$$I_{Den} \le \sum_{r=R-B+1}^{R} \sum_{c=C-B+1}^{C} I_{\max} = BI_{\max} \sum_{r=R-B+1}^{R} \quad (8)$$

$$S_{Den} = \log_2\left(B^2 I_{\max}\right) \quad (9)$$

For the values of B and Imax chosen above, S_{Den}=23 so four block rams of 1024x16 is required for frame interleaved storage of L number of denominators. These denominators are the same for both row and column dimensions. Four additional block RAMs are used in the labelling process for maintaining equivalence tables A and B in Figure 3. One block ram is allocated for the line buffer used to maintain the neighbourhood shown in Figure 2. We now summarize the number of required block RAMs used for storage of numerators in both row and column dimensions (8), denominators (4), equivalence tables (4) and line buffer (1). This means that a total of (17) block RAMs are required for the combined labeller and COG calculator presented in this paper.

V. FUNCTIONAL VERIFICATION

The proposed hardware architecture for the calculation of COG was captured in VHDL. The model was simulated at register transfer level using two different input stimuli as shown in Figure 7. A frame size of R=640 and C=480 pixels was used for the experiment. Simulation output showed the correct number of objects and correct COG values for all objects shown in Figure 7.

VI. PERFORMANCE AND DEVICE UTILIZATION

The COG computation along with labelling as explained in section III was captured in VHDL and synthesized for implementation on the Xilinx VirtexII Pro device having speed grade 6. Post route and placement simulations were performed and design files were analyzed by the Xpower tool included in the Xilinx ISE Foundation toolset [8]. The pixel clock frequency was set to 27 MHz and at a frame speed of 86 fps. The input stimuli are shown in Figure 7. The power consumption, maximum clock frequency, latency and device utilization are results as reported by the toolset after synthesis and simulation, as shown in Table 1. We define latency as the time from the first pixel in a frame arriving at the input of the hardware unit until COG data of the first image object in the same frame starts to transmit on the output.

VII. DISCUSSION

The work presented in this paper shows the most efficient parallelization of the first pass of labelling along with COG calculation.

The performance of the labeller along with COG calculation is shown in Table 1a and 1b. One block RAM is used for the delay of previously assigned labels. Two triple port memories are used for the equivalence table, two read ports and one write port. The synthesis tool duplicates the ram in order to implement single write and dual read port memory. Twelve block RAMs are used for storing the numerators and denominators used for COG calculation. The static power dissipation is dominant, and we can see that only 28% of the available slices are active. The Maximum clock frequency reported by the toolset is 126 MHz. This clock frequency corresponds to 410 frames per second for a video format of 640 by 480 pixels, assuming no synchronisation overhead. This is a relevant assumption for a FPGA based smart camera having a high speed CMOS sensor connected directly to it. The latency is only 11.6 ms for the simulation at 86 fps. This latency is almost exactly the time of one frame with addition of the clock cycles needed for the serial divisors to conclude COG computation for the first object.

From Section IV, it is obvious that the memory storage requirement depends on the maximum allowable image component size BxB, as well as maximum number of labels L. These parameters must be set at system synthesis time and with a margin with respect to the expected video input. More efficient use of the block RAMs will thus require a hardware centric dynamic memory management.

VIII. CONCLUSION

In this paper, we presented a hardware architecture for computation of connected component labelling along with

COG calculation. This implementation is suitable for embedded machine vision systems and smart camera applications having high demands on frame speed, power and latency. We have reason to believe that this work can be extended with computation of additional image object features such as area, bounding box or ellipse parameters.

REFERENCES

[1] H. C. Van Assen, H. A. Vrooman, M. Egmont-Petersen, J. G. Bosch, G. Koning, E. L. Van Der Linden, B. Goedhart, and J. H. C. Reiber, "*Automated calibration in vascular X-ray images using the accurate localization of catheter marker bands*", *invest. Radiol*, vol. 35, no. 4, pp 219-226, April 2000.

[2] X. Cheng, B.Thörnberg, A.W. Malik and N. Lawal, “Hardware centric machine vision for high precision center of gravity calculation”, *Proc. of world academy of science, engineering and technology*, Vol. 64, pp. 736-743, Rome, Italy, 2010.

[3] W.WOLF, B.Ozer and T. Lv, “*Smart cameras as embedded systems*”, *IEE computer*, vol.35, No. 9, pp. 48-53, Sept, 2002.

[4] C. Steger, M. Ulrich and C. Wiedemann, Machine vision algorithems and applications, Wiley-VCH 2008.

[5] M. Wnuk, “*Remarks on hardware implementation of image processing algorithms*”. *Journal of applied mathematics and computer science*. Vol. 18, No. 1, pp105--110 (2008).

[6] C.T. Johnston and D.G. Bailey, “*FPGA implementation of a single pass connected component algorithm*”, *Proc. Of 4th IEEE symp. On electronic design test & applications*, pp 228-231, Hong Kong, China 2008.

[7] B. Thörnberg and N. Lawal, “*Real-time component labelling and feature extraction on FPGA*”, *Proc. of International Symposium on Signals, Circuits and Systems*, pp1-4, Iasi, Romania 2009.

[8] www.xilinx.com, last accessed, 14 November 2010.

[9] D.K. Kim, D.R. Lee, T.C. Pham, T.T. Nguyen and J.W. Jeon, “*Real- time component labeling and boundary tracing system based on FPGA*”, *Proc.2007 IEEE Int. Conf. on Robotics and Biomimetics* (ROBIO '07), pp 189-194, 15-18 Dec. 2007 , Sanya, China

[10] D.G. Bailey and C.T. Johnston, “*Connected component analysis of streamed images*”, *Proc. 2008 Int. Conf. on Field Programmable Logic and Applications* (FPL), pp 679-682, 8-10 Sept. 2008, Heidelberg, Germany.

Improved Performance of a Microstrip Antenna Array

Using a tree structure patch fed by electromagnetic coupling

Tomader Mazri
The signals, systems and components laboratory
Faculty of Science and Technology
Fez, Morocco
Tomader20@gmail.com

Fatima Riouch
The microwave laboratory
National Institute of Posts and Telecommunications
Rabat, Morocco
riouch@inpt.ac.ma

Najiba El Amrani El Idrissi
The signals, systems and components laboratory
Faculty of Science and Technology
Fez, Morocco
elamrani.naj@gmail.com

***Abstract*—The aim of this investigation is to improve the performance of a microstrip antenna array using a tree structure patch supplied by electromagnetic coupling. This array is considered as a first step to design an adaptive microstrip antenna for UMTS use. The patch distribution structure used in this experiment allowed a great improvement of gain, directivity as well as the adaptation level of the studied array. The following work was done between the signals, systems and components laboratory of FST-Fez and the microwave laboratory of the National Institute of Posts and Telecommunications.**

***Keywords*—microstrip antenna array; electromagnetic coupling; adaptive microstrip antenna; tree patch distribution structure.**

I. INTRODUCTION

The UMTS (Universal Mobile Telecommunications System) is the cellular standard for mobile telecommunication systems of the third generation [1]. It has been adopted worldwide by 1998 but its service has been delayed due to the implementation costs.

Its special feature is the simultaneous transmission of voice and data with higher rates than those permitted by previous generations.

The development of these systems requires technological advances in electronic components, computer software, coding techniques and antennas.

Indeed, the antenna is one of the key points of wireless network since it represents the last link in the chain that allows emission, transmission and reception of the signal and therefore the information contained in it [1].

The ultimate goal of this work is designing an adaptive microstrip antenna for base stations of UMTS telecommunication networks, to improve the cover.

To reach this purpose, the parameters of our antenna (resonant-frequency, geometry and bandwidth) will be considered for an UMTS application.

The circuit will be made with FR4, a commonly used material for the manufacture of printed circuit with the following characteristics (thickness: 1.6 mm ε_r: 4.5 and tgδ: 0.02).

II. FEEDING BY ELECTROMAGNETIC COUPLING

In Section 2, we will introduce the electromagnetic coupling between radiating elements in a printed antenna array, and we will discuss the different types of this coupling.

A. Introduction to electromagnetic coupling between radiating elements

The electromagnetic interferences between radiating elements in a printed antenna array, is expressed by the modification of the surface currents distribution. This phenomenon, called coupling, depends on the antenna type and the distance between its elements. The coupling between two printed periodical antennas has a great importance in the design of antennas arrays, because it may cause a change in the radiation pattern.

The current flowing in each antenna induces currents in the all other antennas, whatever they are supplied or not.

The mutual coupling is due to the simultaneous effects of radiation in free space and the propagation of surface waves. This is an important criterion which should be considered while calculating array characteristics.

The theoretical calculation of mutual coupling depends on the antenna type and the distance between its elements. Jedlicka and Carver have studied experimentally the effect of coupling between patch antennas for circular and rectangular geometries [2]. Different methods have been presented to calculate the coupling coefficient between microstrip antennas.

They were proposed by various authors such as Sindora, Pénard, Pozar, et al. [3].

B. *Coupling in the E plan and the H plan*

The radiation patterns are usually represented in two orthogonal planes "E plane and H plane", in relation to the principal direction:

- E Plan: location of space points where the radiated electric field is contained in this plan.

- H Plan: location of space points where the radiated magnetic field is contained in this plan.

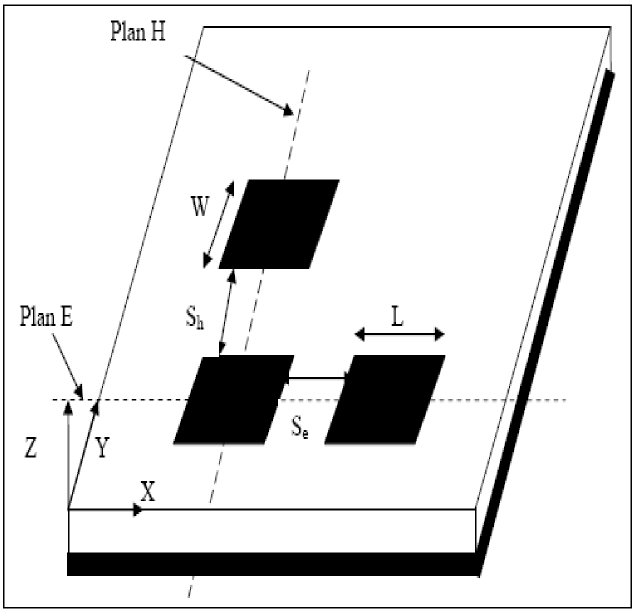


Figure 1. Electromagnetic coupling between patches antennas in the E plane and H.

Furthermore, two coupling types are distinguished:

- Horizontal coupling or coupling in the E plan: This means coupling between two elements in the same substrate, along the x direction with a Se coupling separation. All the W widths of the patches (in the Y direction) have the same size.

- Vertical coupling or coupling in the H plan, along the y direction with a S_h coupling separation. The L lengths of patches (in the direction of x) are identical.

III. Tree Structure and Improvement of the Array Performances

The aim of this work is to improve the performances of a microstrip antenna array of a circular shape by choosing a tree patch distribution structure. To illustrate the improvement through the comparison of findings, we will present first the basic array results.

The simulation will be performed using the ADS tool “Advanced Design System”. The substrate used for this simulation is FR4 for a resonant frequency of 2GHz.

A. *The studied array*

Our studied array *“Fig. 2”* consists of eight circular patches supplied by microstrip lines and adapted using coplanar notches [4].

The connecting lines should be sized to be adapted to 50 Ohms at the input of the array. Wilkinson dividers were integrated in order to obtain an impedance of 100 Ohm at the entrance of the patches [5].

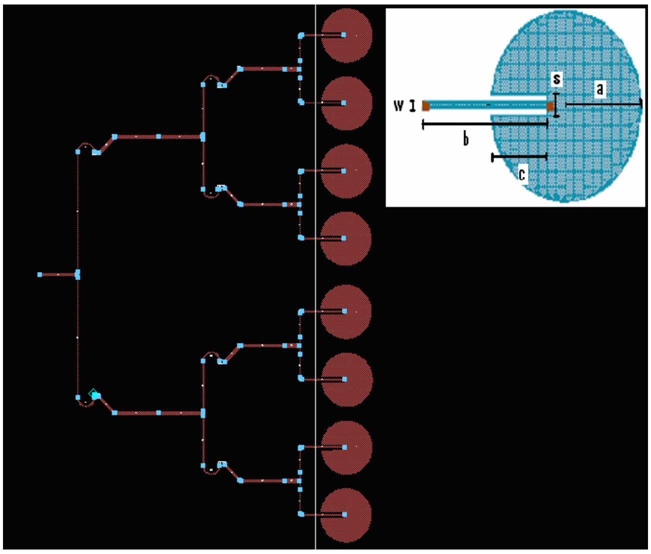


Figure 2. Microstrip array antenna with eight elements circular.

The values of the different parts of the patch and the connecting lines are:

a = 20.70mm.
b = 34.00mm.
c = 15.40mm.
w = 1.25mm.
s = 4.40mm.

The adaptation quality of an antenna is defined by giving either its characteristic impedance (usually 50 ohms) or its reflection coefficient; Figure 3 shows its value for our array of eight patches at 2 GHz.

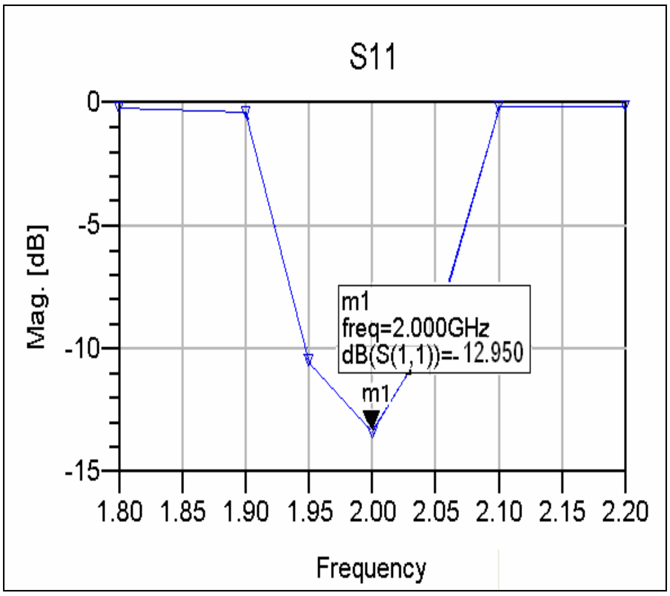
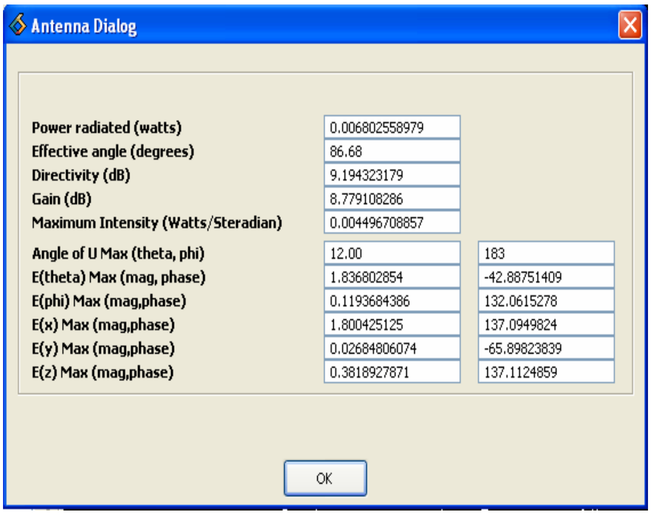


Figure 3. Reflection coefficient of the microstrip array antenna with eight circular elements (GHz).

The simulation results show an adaptation coefficient of -12 dB. Concerning the array characteristics, we can see 9dB directivity, a gain of 8dB and an effective angle of 86 degrees

"Fig. 4". This array allows a bandwidth of 90 MHz. It also presents an electromagnetic coupling in the H plane.



Power radiated (watts)	0.006802558979	
Effective angle (degrees)	86.68	
Directivity (dB)	9.194323179	
Gain (dB)	8.779108286	
Maximum Intensity (Watts/Steradian)	0.004496708857	
Angle of U Max (theta, phi)	12.00	183
E(theta) Max (mag, phase)	1.836802854	-42.88751409
E(phi) Max (mag,phase)	0.1193684386	132.0615278
E(x) Max (mag,phase)	1.800425125	137.0949824
E(y) Max (mag,phase)	0.02684806074	-65.89823839
E(z) Max (mag,phase)	0.3818927871	137.1124859

Figure 4. Characteristics of the microstrip antenna array with eight elements circular.

B. *The array with a tree structure*

We have been inspired for this structure by YAGUI-UDA antenna. The eight patches supplied by microstrip lines represent the radiator element, and the patches supplied by electromagnetic coupling represent the director element of the antenna.

In this simulation our array has been developed by introducing circular patches arranged in a tree structure and supplied by electromagnetic coupling *"Fig. 5"*. This array presents an electromagnetic coupling in the E and H planes.

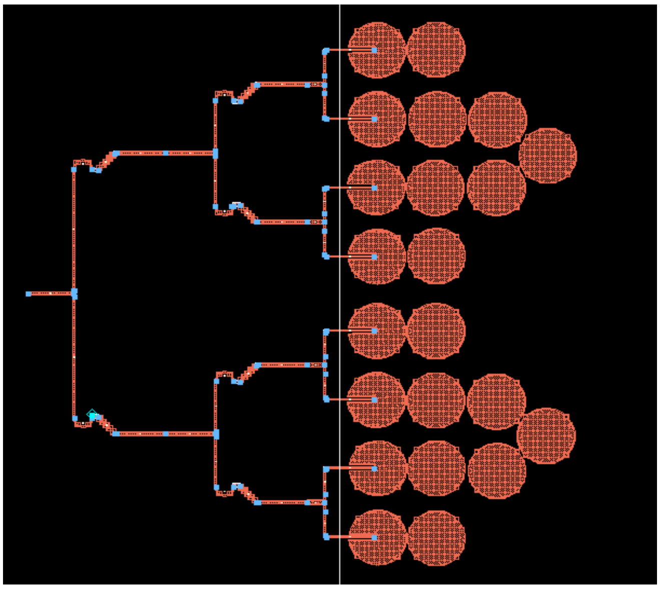


Figure 5. Microstrip antenna arrays with a tree structure.

For this structure, the simulation results give an adaptation coefficient of -15 dB "Fig. 6", and the array is characterized by a directivity of 12 dB, a gain of 11.4dB and an effective angle of 46 degrees "Fig. 7". This array allows a bandwidth of 140 MHz.

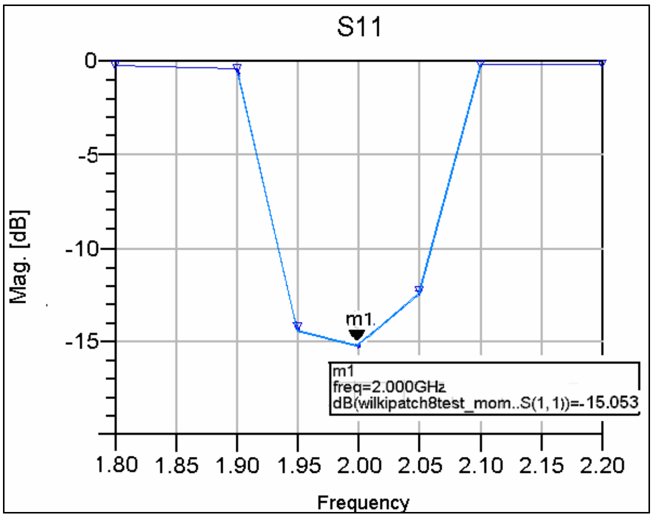
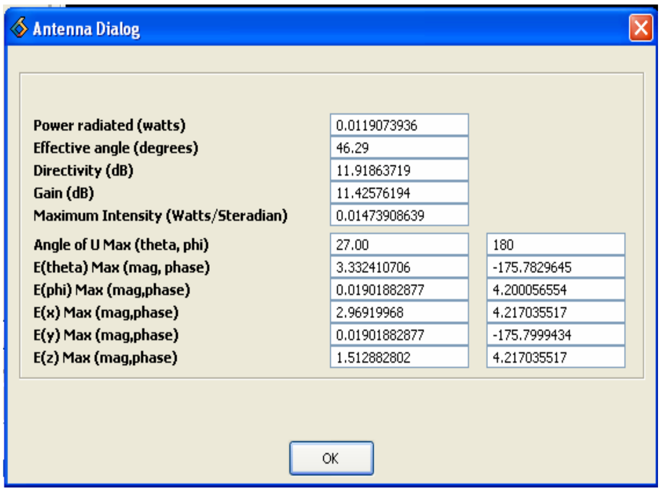


Figure 6. Reflection coefficient of microstrip antenna arrays with tree structure.



Power radiated (watts)	0.0119073936	
Effective angle (degrees)	46.29	
Directivity (dB)	11.91863719	
Gain (dB)	11.42576194	
Maximum Intensity (Watts/Steradian)	0.01473908639	
Angle of U Max (theta, phi)	27.00	180
E(theta) Max (mag, phase)	3.332410706	-175.7829645
E(phi) Max (mag,phase)	0.01901882877	4.200056554
E(x) Max (mag,phase)	2.96919968	4.217035517
E(y) Max (mag,phase)	0.01901882877	-175.7999434
E(z) Max (mag,phase)	1.512882802	4.217035517

Figure 7. Characteristics of the microstrip antenna array with a tree structure.

The radiation pattern is given by the Figure 8.

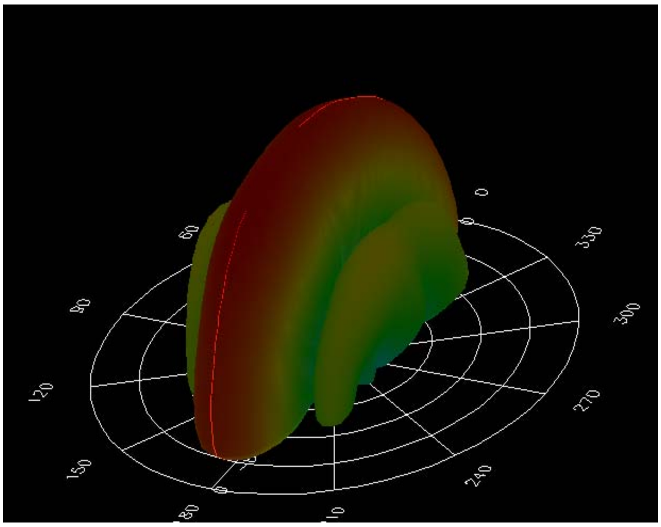


Figure 8. Radiation pattern of the antenna array with a tree structure.

Our network consists of a linear distribution of patches where the appearance of side lobes that have the same level. If these lobes hamper the application, the solution is simply to apply a law of amplitude weighting to reduce their levels. We also note that we have a directive main lobe and this is consistent with the results given by the computer (Figure 7).

IV. CONCLUSION

The experiments performed until now have allowed an improvement of performances of our microstrip array antenna by using a tree structure patch supplied by an electromagnetic coupling. The simulation results show an adaptation coefficient of -15 dB, and the array is characterized by a directivity of 12dB, a gain of 11.4dB and an effective angle of 46 degree. This network will form the basic network of our application; more precisely it will be the pointing network of users. Indeed it will be developed in order to have the opportunity to change its angle depending on demand. So as a perspective, we will integrate the intelligent function of our network and improving the bandwidth to meet the need for a UMTS application.

REFERENCES

[1] S. Bencheikh, H.Griguer, A. Hazaoud, A. Najid, and P.Muhlethaler ' studies parametric and experimental of the Bi-bandpatch antennas with filter for the basic stations needs', TELECOM and JFMMA 2005, pp. 165-169. Rabat, Morocco 23 to 25 March 2005.

[2] L. Chouti 'Contribution to the study of rectangular printed antennas dual band and multi band considering the coupling effect 'thesis argued in 2009. University Mentouri-Constantine, Faculty of Science Department of Electronic Engineering.

[3] P. R. Haddad and D. M. Posard "Anomalous mutual coupling between microstrip antennas", IEEE Transactions on Antennas and Propagations, Vol. 42, N°. 11, pp. 1545-1549, November 1994.

[4] T. Mazri, N. El Amrani, F. Riouch, M. Drissi , and E. Marzorf 'Adaptation Improvement of a circular microstrip antenna by using notches' TELECOM and JFMMA 2009, pp. 65-66. Agadir, Morocco 12 to 14 March 2009.

[5] T. Mazri, N. El Amrani, and F. Riouch 'Simulation of a microstrip antenna for adaptive array use UMTS'OPTIQUE 2010, pp. 122-125. Tangier, Morocco 21 to 23 April 2010.

Access Control for Coordinated Multi-antenna Cellular Architecture with Scheduling

Xiaodong Xu, Dan Hu, Xiaofeng Tao, Zhijie Hao
Wireless Technology Innovation Institute (WTI), Key Lab of Universal Wireless Comm., Ministry of Education
Beijing University of Posts and Telecommunications (BUPT)
Beijing, China
E-mail: xuxiaodong@bupt.edu.cn, hudan@mail.wtilabs.cn, taoxf@bupt.edu.cn, haozhijie@mail.wtilabs.cn

***Abstract*—While current research focuses on Enhanced 3rd Generation and IMT-Advanced mobile systems, many advanced techniques are investigated by world-wide research institutes and standard organization, such as Multi-Input Multi-Output, Coordinated Multi-Point and coordinated multi-antenna cellular network architecture. Based on these novel techniques, the access control strategies also need to be developed. Based on Maximum Utility Principle Access Control method, improved access control algorithm with combination of scheduling for coordinated multi-antenna cellular architectures is proposed in this paper. Two algorithms are brought out with the Proportional Fairness and Maximum C/I utility function respectively. By application in Coordinated Multi-Point based Group Cell architecture, performance evaluation and analyses verify the merits of two proposed algorithms in improving system throughput, user fairness and efficiency of system resources usage.**

Keywords-Access Control; coordinated multi-antenna; scheduling; Maximum Utility Principle Access Control; Group Cell

I. INTRODUCTION

The objective of the Enhanced 3rd Generation (E3G), 4G has been anticipated to provide users after the year 2010 with the data rate up to 100Mbps or 1Gbps in mobility environments [1-2]. Numerous research plans and projects towards E3G and 4G have been initiated in Europe, East Asia and North America, etc. Many international standardization organizations, such as 3GPP Long Term Evolution (LTE) [1] and LTE-Advanced [3], have initialized the research and standardization of E3G systems. Moreover, International Telecommunication Union (ITU) has also launched 4G standardization work, named IMT-Advanced [4].

With the research and development for E3G and 4G systems, a lot of advanced physical layer technologies show their merits to be applied in next generation mobile telecommunication systems. Among these techniques, the multi-antenna techniques and multi-carrier techniques, such as Multi-Input Multi-Output (MIMO) and Orthogonal Frequency Division Multiplex (OFDM), show their merits in improving system capacity and coverage. MIMO and OFDM techniques have been standardized in 3GPP LTE system as key techniques of E3G physical layer. Moreover, Coordinated Multi-Point joint transmission was proposed in 3GPP LTE-Advanced standard work as a key technique to mitigate Inter Cell Interference and further improve the cell-edge performance [5-6]. In this approach, if both data and channel of all users could be shared in real time, adjacent base stations could act as a single and distributed antenna array and hence, data to a user is simultaneously transmitted from multiple base stations to improve the received signal quality. Notice that Coordinated Multi-Point (CoMP) techniques have been already implemented in Group Cell architecture [7-8] with coordinated multi-antennas as early as 2001, which has been implemented in China Beyond 3G (B3G) Future Technologies for Universal Radio Environment (FuTURE) TDD systems in B3G trial network with OFDM, and MIMO techniques, etc.

Accordingly with the evolution of physical layer techniques, the Media Access Control (MAC) and Radio Resource Management (RRM) techniques are all facing the requirements for evolution. Furthermore, in order to apply CoMP joint transmission effectively, traditional RRM strategies for cooperation among coordinated cells also need to be evolved either.

The access control methods used in 2G/3G systems [9-10] cannot accommodate the features of coordinated multi-antenna cellular architecture, especially for users served by coordinated multiple antennas. There are a lot of challenges for coordinated multi-antenna access control strategy, such as the optimization of choosing multi-antennas to form the coordinated transmission set for access users, the definition of admission threshold for access users and the principal of admission and rejection etc.

In [11], Maximum Utility Principle Access Control (MUPAC) method for coordinated multi-antenna cellular architecture with application in Group Cell architecture was brought out as an example. MUPAC method can maximum the usage of limited system resources with guaranteeing access users' QoS requirements based on defined utility functions [11]. Furthermore, through MUPAC method, the interference increasing caused by access users can also be mitigated maximally and the accessing success probability, accessed user number can also be improved.

With MUPAC method, when the system is under heavy-load situation, MUPAC can fully show its merits in improve resource usage efficiency. However if the system is relatively light loaded and the capacity is enough for more users, MUPAC method may not fully use system capability to serve users with its best, because MUPAC cares more on the minimum QoS requirements for resource utility rather not on users better performance. So, scheduling can be used in combination with MUPAC method to solve this problem and improve user service experience after access success ratio improvement.

This paper presents two improvement algorithms based on MUPAC with scheduling utility function. One is Throughput Targeted MUPAC (TT-MUPAC) and another is Throughput and Fairness Targeted MUPAC (TFT-MUPAC) algorithm.

In Section II, MUPAC method is briefly introduced with an application example in coordinated multi-antenna Group Cell architecture. In Section III, TT-MUPAC and TFT-MUPAC are described in details combining with scheduling in access control. The performance evaluation and simulation results are stated in Section IV. Finally, there comes the conclusion.

II. Maximum Utility Principle Access Control Strategy in Group Cell Architecture

In order to further improve performance for the cell edge users, CoMP will be applied in LTE-Advanced. CoMP implies dynamic coordination among multiple geographically separated transmission points, which involves two schemes with coordinated scheduling and joint processing/transmission. By aggregating the joint processing of multiple cells, CoMP technology can increase the throughputs on the cell-edge. A typical system model for coordinated multi-antenna cellular architecture is Group Cell, which is described in Figure 1.

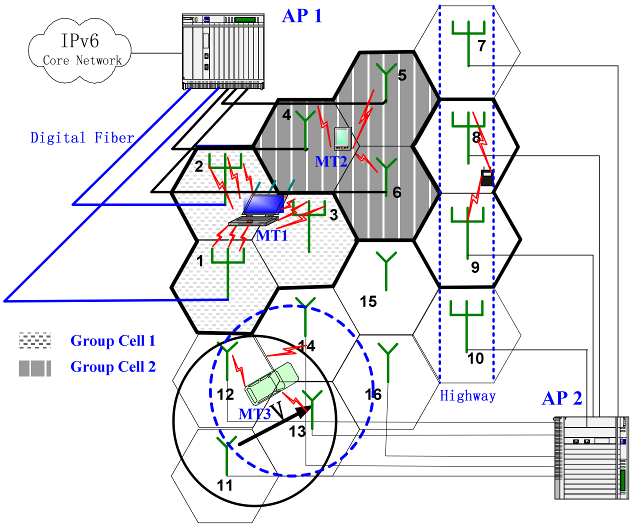


Figure 1. Group Cell Architecture with CoMP

In Group Cell architecture, users in the system are served by more than one antenna (Group Cell) included in Access Points (AP). The access control method in the Group Cell needs to solve the problem of how to choose multiple antennas to form the serving Group Cell and allocate appropriate resources to users. The size of Group Cell can be adjustable for users by their QoS requirements. Therefore, by adding antenna with maximum utility to user's current serving Group Cell step by step to fulfill the users' QoS requirements can solve this problem. This solution can maximize the usage of limited system resources with guaranteeing access users' QoS requirements. Furthermore, the interference caused by new users can also be mitigated maximally and the accessing success probability can also be improved.

The steps of adding antennas with maximum utility can be accomplished based on Dijkstra's Shortest Path Algorithm [12] in Graph Theory. Based on Dijkstra's Shortest Path Algorithm, when there are new users initiate their access attempts in Group Cell architecture, the shortest path in the Dijkstra's Algorithm can be replaced by the minimal cost of accessing process. The cost of accessing process includes the interference to other users and occupying system resources (antennas, channels and other resources). Furthermore, the cost can be represented by the utility functions, including the gains for the access user and deterioration to other users. Therefore, the seeking for shortest path in Dijkstra's Algorithm can be transferred to seeking the antennas or resources with maximum utility. The Maximum Utility Principle can improve the system capacity and load ability. By the Dijkstra's Shortest Path Algorithm and the Maximum Utility Principle, the user accessing in multi-antenna distributed Group Cell can be effectively accomplished.

The utility function of MUPAC has two aspects, including the gain of new antenna added in current serving Group Cell and the deterioration for other users existed in the system.

The utility function is shown as (1).

$$\begin{aligned} U(i,...,j,k) &= \zeta_{Ck}[G_C(i,...,j,k) - I_C(i,...,j,k)] \\ &+ \beta(1-\zeta_{Ck}) \max_{M \neq C} \left\{ \zeta_{Mi} \cdot ... \zeta_{Mj} \cdot \zeta_{Mk} [G_M(i,...,j,k) - I_M(i,...,j,k)] \right\} \end{aligned}, \quad (1)$$

where $U(i,...,j,k)$ denotes the utility of adding antenna k to current serving Group Cell formed by antennas $i,...,j$. C and M denote the resources and C is the current resource used by the serving Group Cell. ζ_{Ck} is an indicator function, which indicates the occupying information of resource C in antenna k.

$$\zeta_{Ck} = \begin{cases} 0, \textit{Resource C occupied in AE k} \\ 1, \textit{Resource C available in AE k} \end{cases}, \quad (2)$$

where $G_C(i,...,j,k)$ denotes the gain achieved by adding antenna k to current Group Cell with resource C. $I_C(i,...,j,k)$ denotes the interference to other users by adding antenna k to current serving Group Cell with C. β is a constant between 0 and 1 to introducing the penalty for replacing current resource C with different resource (resource C') for the new serving Group Cell. β can be set according to the current system load condition. The choice of C' to replace C can also be achieved by Maximum Utility Principle with the utility function, which is:

$$C' = \arg\max_{M \neq C} \left\{ \zeta_{Mi} \cdot ... \zeta_{Mj} \cdot \zeta_{Mk} [G_M(i,...,j,k) - I_M(i,...,j,k)] \right\}. \quad (3)$$

Considering the actual mobile systems, the gain and interference in utility function are usually represented by SINR. Therefore, (1) can be revised to:

$$
\begin{aligned}
&U(i,\ldots,j,k)\\
&=\zeta_{Ck}[\frac{\lg_{k,i}}{\sum\limits_{n\neq i,j\ldots,k}(1-\zeta_{Cn})\lg_{n,i}}-\sum_{n\neq i,\ldots,j,k}(1-\zeta_{Cn})\frac{\lg_{n,n}}{\lg_{k,n}}]\\
&+\beta(1-\zeta_{Ck})\arg\max_{M\neq C}\left\{\begin{matrix}\zeta_{Mi}\zeta_{Mj}\cdots\zeta_{Mk}\\ [\frac{\lg_{k,i}}{\sum\limits_{n\neq i,\ldots,j,k}(1-\zeta_{Mn})\lg_{n,i}}-\sum\limits_{n\neq i,\ldots,j,k}(1-\zeta_{Mn})\frac{\lg_{n,n}}{\lg_{k,n}}]\end{matrix}\right\}
\end{aligned}
\quad . (4)
$$

And (4) can also be revised to:

$$
U(k)=\arg\max_{C}\left\{\zeta_{Ck}[\frac{\lg_{k,k}}{\sum\limits_{n\neq k}(1-\zeta_{Cn})\lg_{n,k}}-\sum_{n\neq k}(1-\zeta_{Cn})\frac{\lg_{n,n}}{\lg_{k,n}}]\right\},\quad (5)
$$

where $\lg_{n,k}$ denotes the path gain between antenna n to the access user who is currently served by antenna k. The power for each antenna in (4) and (5) are equally allocated.

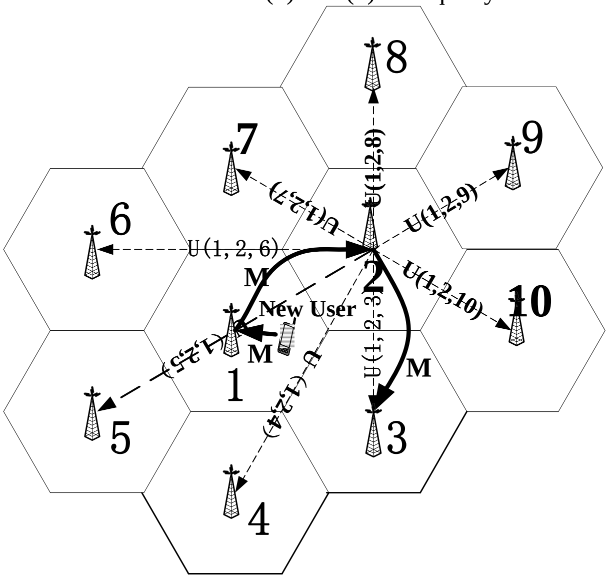


Figure 2. MUPAC process in Group Cell

The application example of MUPAC in Group Cell architecture is shown in Figure 2. The detailed implementation steps of MUPAC are shown as follows.

1) Access user initiates access attempt.

2) AP obtains the users' receiving pilot strength of each antenna.

3) Based on the information in step 2), AP calculates the utility of each antenna and available resource by utility function and chooses the first antenna and resource with the Maximum Utility Principle to form the serving Group Cell. If all the antennas detected by access user have no resource available, the access user will be transferred to the accessing waiting list. In Figure 2, the access user select antenna 1 as the first serving antenna with resource unit by Maximum Utility Principle.

4) AP obtains the users' receiving SINR of serving Group and compares it with user's QoS requirement. If current serving Group can provide adequate QoS to the user, the access process accomplishes successfully. Vice verse, the access user need more antennas added to the current serving Group. In Figure 2, the access user needs more antennas to get its desired QoS. So, antenna 2 is chosen to be added in current serving Group Cell.

5) AP obtains the users' receiving SINR of antennas excluding current serving Group and chooses the antenna with maximum utility to add it to the serving Group Cell. This step needs to guarantee the new antenna and current serving Group Cell to use the same resource. The utility function includes the penalty of resource changing. Then, goes to step 4). In Figure 2, antenna 3 is added and the serving Group Cell of 1, 2 and 3 has enough quality to serve the user with resource unit M.

III. Maximum Utility Principle Access Control with Scheduling

When we are choosing the algorithms for access control, we always care about the quality of services, as well as the efficiency of resource which is associated with the system capacity. Ensuring the QoS of access users' communications, MUPAC method gives the least sources to users to reach a minimum acceptable QoS. Considering the variable mobile communication environments and multi-user diversity, also the service experience of users, it will be helpful to implement scheduling into the process of access control for coordinated multi-antenna cellular architecture.

A. Throughput Targeted-MUPAC

In order to enable better use of the resources and reaching higher system throughput, we should consider using scheduling in access control to adapt to different environments and make full use of the resources. When the system load is light, MUPAC is not good enough, especially in the condition of dealing with data services. If we make full use of system resources and increase the system throughput, it would be beneficial to either the users or the system. TT-MUPAC brings out a good consideration on this point.

TT-MUPAC strategy gives different resources to different users in access control which depends on the system conditions. If the system is heavy-loaded with many services required, it gives the user the least resource to reach the required QoS. On the other hand, if the system is relatively light-loaded and there are many resources available, the access users will get most resource to improve system throughput.

In TT-MUPAC strategy, MAX C/I scheduling is employed. The key point of combination of MAX C/I and MUPAC is to give some users more resource to get multi-user diversity in the system. In this way, we can improve the system throughput obviously.

B. *Throughput and Fairness Targeted-MUPAC*

TT-MUPAC brings some advantages on system throughput, but when it comes to user fairness, the performance is decreased. Throughput and fairness are both important in access control strategy. So we should make some improvements on MUPAC and TT-MUPAC methods to reach a better performance on throughput and fairness. TFT-MUPAC method is proposed to achieve a balance between fairness and system throughput.

In the TFT-MUPAC strategy, the utility function of TFT-MUPAC should consider both system throughput and user fairness. In order to include the consideration of fairness into the access control strategy, we add a fairness factor into the utility function to present the improvement.

$$U'(i,...,j,k) = F \bullet G(i,...,j,k) = \left(\frac{R_{average}}{R_{generated}}\right)^{\gamma} \bullet \zeta_{Ck}[G_C(i,...,j,k) - I_C(i,...,j,k)] + \beta(1-\zeta_{Ck}) \max_{M \neq C} \left\{ \zeta_{Mi} \cdot ... \zeta_{Mj} \cdot \zeta_{Mk}[G_M(i,...,j,k) - I_M(i,...,j,k)] \right\} \quad , (6)$$

F denotes the fairness of the service quality, which is,

$$F = \left(\frac{R_{average}}{R_{generated}}\right)^{\gamma} \quad . \qquad (7)$$

$R_{generated}$ is the service quality user can get with the target antennas when he is accessed. $R_{average}$ is the average service quality of the users already in the system. γ is the factor of fairness and can be adjust with the actual situation. When γ is getting bigger, the fairness will be better. In the simulation of this paper, γ is set as 1.

The flow chart of TFT-MUPAC is show in the Figure 3.

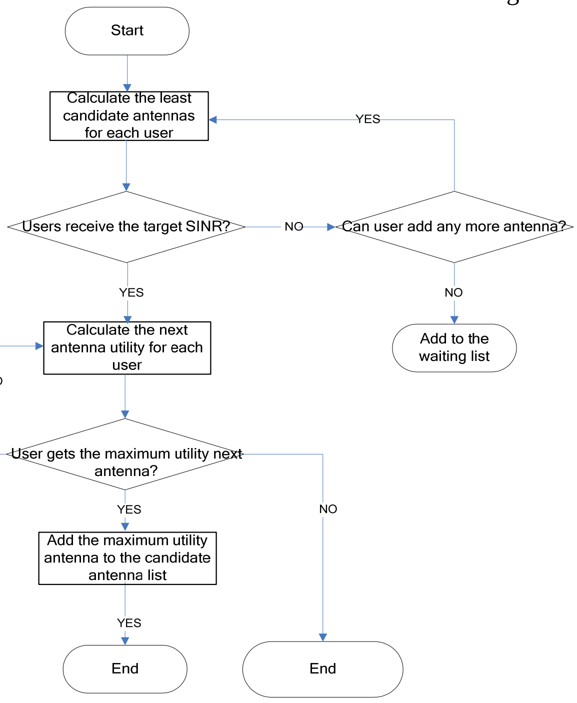

Start
Calculate the least candidate antennas for each user
YES
Users receive the target SINR?
NO
Can user add any more antenna?
YES
NO
Calculate the next antenna utility for each user
Add to the waiting list
NO
User gets the maximum utility next antenna?
YES
NO
Add the maximum utility antenna to the candidate antenna list
YES
End
End

Figure 3. TFT-MUPAC method Flow Chart

In the TFT-MUPAC strategy, antennas are allocated to receive a fairer QoS. At the same time the system throughput is also considered. Proportional fairness scheduling method is employed. In this way, system carries out a good performance on both system throughput and fairness.

IV. Performance Evaluation

For the performance evaluation and analyses, MUPAC method is taken for performance comparing based on coordinated multi-antenna Group Cell architecture. Maximum Utility Principle Access Control chooses antennas and allocates resources according to the Maximum Utility Principle. The Group Cell size of Maximum Utility Principle Access Control method is limited up to 4. TT-MUPAC and TFT-MUPAC employs scheduling with MAX C/I and Proportional Fairness algorithms. System-level simulation is adopted to evaluate these three access control methods by comparing the successfully accessed user numbers with different system load (total access user number generated), system throughput and fairness. The power allocation for these three algorithms is the same as fixed power allocation scheme. The simulation parameters and setting are shown in Table I.

TABLE I. Simulation Parameters and Setting

Parameters	Setting
Traditional inter-site distance	$500\sqrt{3}$ m
Group Cell inter-antenna distance	500m
Carrier Frequency	5.3GHz
Path gain model	25log10(d)+35.8 [13]
Shadow fading deviation	5dB
Total bandwidth	20MHz
Effective bandwidth	17.27MHz
Number of useful sub-carriers	884
Sub-carrier spacing	19.5KHz

The simulation results are shown in Figure 4 to Figure 8.

Figure 4 shows the system throughput of MUPAC and TT-MUPAC. TT-MUPAC has obvious throughput advantage over MUPAC scheme. The reason for this throughput gain mainly comes from multi-user diversity with MAX C/I scheduling. MUPAC only guarantees the minimum requirements of access users' QoS for maximum resource efficiency. By TT-MUPAC, scheduling is to improve the throughput with light load.

Figure 5 shows the access success rate of MUPAC and TT-MUPAC. TT-MUPAC is better than MUPAC, because TT-MUPAC use more resources for few users to get more throughputs. The relatively low efficiency of resource utility makes access users having less available resources and lows down the access success rate.

Figure 6 shows the fairness of access users based on MUPAC and TT-MUPAC by SINR variance. From the simulation results, TT-MUPAC has worse fairness than MUPAC. This is the nature of MAX C/I scheduling method.

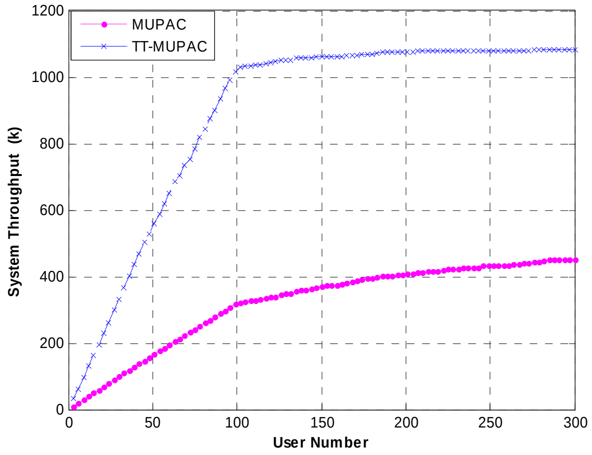


Figure 4. System Throughput of MUPAC vs. TT-MUPAC

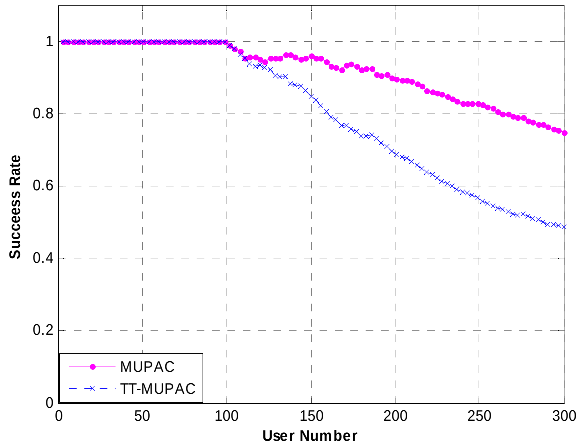


Figure 5. Access Succeed Rate of MUPAC vs. TT-MUPAC

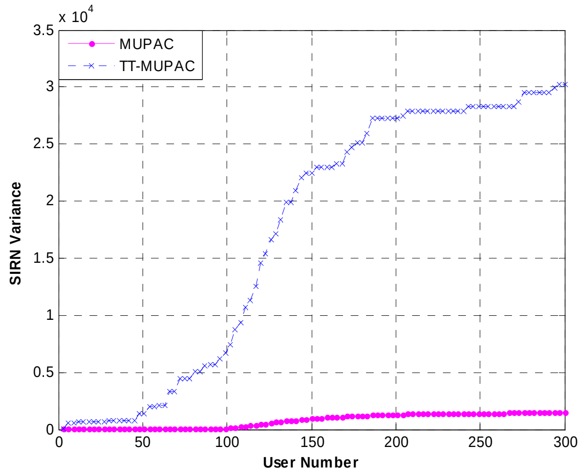


Figure 6. System Fairness of MUPAC vs. TT-MUPAC

Figure 7 and Figure 8 show the performance of MUPAC, TT-MUPAC and TFT-MUPAC, including system throughput and user fairness. Figure 7 shows the throughput performance of MUPAC, TT-MUPAC and TFT-MUPAC. TT-MUPAC has the best performance and TFT-MUPAC has the worst performance with the features of scheduling methods. Figure 8 shows the user fairness of these three methods. TFT-MUPAC has better fairness performance than MUPAC and TT-MUPAC.

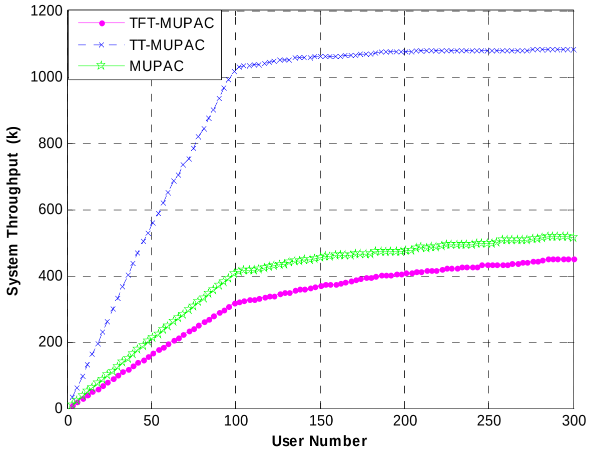


Figure 7. System Throughput of MUPAC, TT-MUPAC and TFT-MUPAC

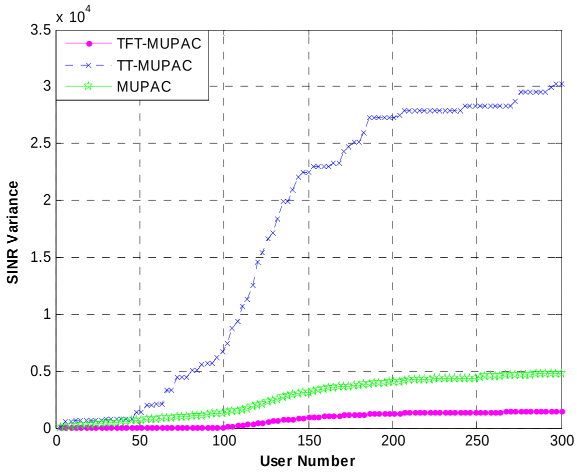


Figure 8. User Fairness of MUPAC, TT-MUPAC and TFT-MUPAC

V. CONCLUSIONS

Maximum Utility Principle Access Control method was proposed for coordinated multi-antenna cellular architecture by Dijkstra's Shortest Path Algorithm and utility function with Maximum Utility Principle for step by step multi-antenna choosing for access users.

Based on MUPAC, this paper proposed two improvements for MUPAC with scheduling algorithms. With combination of scheduling and access control strategy, Throughput Targeted-MUPAC and Throughput and Fairness

Targeted-MUPAC can get better performance of system throughput and user fairness respectively with appropriate resource utility efficiency to accommodating different situation of system load and access users. Taken coordinated multi-antenna cellular architecture - Group Cell as application, TT-MUPAC and TFT-MUPAC algorithm are described in details with the utility function, revised maximum utility principle and flow chart of accessing process. Performance evaluation and analyses verify the merits of TT-MUPAC and TFT-MUPAC algorithms in improving system throughput, accessing success rate and user fairness.

VI. ACKNOWLEDGMENT

This work is supported by Project 61001116 of National Natural Science Foundation of China, National 863 Project 2009AA011506 and State Emphasis Special Project 2009ZX03003-011-02

REFERENCES

[1] 3GPP. TR25.913, Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN), 2009.

[2] ITU-R 229-1/8, Draft text for ANNEX 4 of the Circular Letter on an invitation propose candidate radio interface tech. for IMT-Advanced.

[3] 3GPP R1-082024, A discussion on some technology components for LTE-Advanced," Ericsson, 3GPP TSGRAN WG1 #53, 2008.

[4] C. Wijting, K. Doppler, K. KallioJarvi, T. Svensson, M. Sternad, G. Auer and N. Johansson. Key technologies for IMT-advanced mobile communication systems, IEEE Wireless Communications, vol. 16, no. 3, pp. 76-85, 2009.

[5] 3GPP R1-082024, A discussion on some technology components for LTE-Advanced, Ericsson. 3GPP TSGRAN WG1 #53, Kansas City, MO, USA, 2008.

[6] 3GPP TR 36.814, Further Advancements for E-UTRA Physical Layer Aspects (Release 9), 2010

[7] P. Zhang, X. Tao, J. Zhang, Y. Wang, L. Li, and Y. Wang. The Visions from FuTURE Beyond 3G TDD, IEEE Communications Magazine, vol. 43, no.1, pp. 38-44, Jan. 2005.

[8] X. Tao, Z. Dai, and C. Tang. Generalized cellular network infrastructure and handover mode-group cell and group handover. Acta Electronic Sinica, vol. 32, no. 12A, pp. 114-117, 2004.

[9] R. M. Rao, C. Comaniciu, T. V. Lakshman, and H. V. Poor. An overview of CAC principles in DS-CDMA networks – call admission control in wireless multimedia networks, IEEE Signal Processing Magazine, vol. 21, no.5, pp. 51-58, 2004.

[10] N. Bambos, S. C. Chen, and G. J. Pottie. Channel access algorithms with active link protection for wireless communication networks with power control, IEEE/ACM Trans. Networking, vol. 8, no.5, pp. 583-597, 2000.

[11] X. Xu, C. Wu, X. Tao, Y. Wang, and P. Zhang. Maximum Utility Principle Access Control for Beyond 3G Mobile System, Wireless Communications and Mobile Computing, Journal of Wiley, vol. 7, no. 8, pp. 951-959, 2007.

[12] E. W. Dijkstra. A note on two problems in connection with graphs. Numerische Mathematik, pp:269-271, 1959.

[13] X. Zhao, J. Kivinen, P. Vainikainen, and K. Skog. Propagation characteristics for wideband outdoor mobile communications at 5.3 GHz, IEEE Journal on Selected Areas in Communications, vol. 20, no.3, pp. 507-514, 2002.

Using Grounded Theory as a Supportive Technique for System Requirements Analysis

Mohanad Halaweh
College of Information Technology
University of Dubai
Dubai, UAE
mhalaweh@ud.ac.ae

***Abstract*—Requirements analysis is a key phase in information systems development. During this phase, system analysts use different techniques and methods to determine and structure the systems requirements. In this paper, the author rationalises the use of grounded theory as a technique for requirements analysis. It aims to establish theoretically that applying grounded theory procedures and techniques will strengthen and add value to the analysis phase.**

Keywords-grounded theory; IS devdelopment; requirements analysis; requirements engineering

I. INTRODUCTION

Requirements analysis (RA) is a key phase in information systems (IS) development. During this phase, system analysts use different techniques and methods to determine and structure the systems requirements. In this paper, the author rationalises the use of grounded theory (GT) as a technique for requirements analysis. It aims to establish theoretically that applying grounded theory procedures and techniques will strengthen and add value to the analysis phase.

The next section provides an overview of the grounded theory method, and the third section briefly explains the requirement analysis phase of IS development. The fourth section presents literature review. The fifth section rationalises the use of grounded theory for requirements analysis and explains how it adds value and strengthens the analysis stage, and the final section presents the conclusion.

II. OVERVIEW OF GROUNDED THEORY

Grounded theory has been intensively used in IS and software engineering research [1][2][3][4][5][6]. It is a “qualitative research method that uses a systematic set of procedures to develop an inductively derived grounded theory about a phenomenon” [7] (p.24). It was originally developed by Glaser and Strauss in 1967 [8].

Although different schools of thought concerning grounded theory have arisen from the subsequent disagreement between the originators themselves, the author does not discuss those, as they are beyond the research scope. Furthermore, the aim is to show how grounded theory can be applied in requirements analysis by utilising the concepts proposed by Strauss and Corbin’s approach, and avoid the current debate concerning the theory itself. This section presents the essential concepts, techniques and procedures of grounded theory that will be used in requirements analysis by following Strauss and Corbin’s approach [7].

- *Theoretical Sampling*: Sampling in grounded theory is based on concepts shown to have theoretical relevance to the developing theory. It relates to the sampling of new data based on the analysis of that initially collected from the initial interviews, where the concepts that emerge constantly guide the researcher as to the nature of future data, their sources and the issues to be discussed in subsequent interviews in order to develop the categories. The initial questions for the fieldwork are based on concepts derived from literature (i.e. data gathered previously), which provides the researcher with a starting point and a focus; later, the sampling becomes more in-depth. Strauss and Corbin [7] explain that the sampling should focus on sampling incidents and not persons – in other words, collecting data about what informants do in terms of action/interaction, condition and consequence of the action. The researcher continues this process until the theoretical base is *saturated*, where no new data and ideas emerge regarding the developed concepts and categories.
- *Coding* is the key process in grounded theory [7]. It begins in the early stages after the first interviews for data collection. This process comprises three coding steps:

1. **Open coding** is “the process of breaking down, examining, comparing, conceptualizing and categorizing data” [7] (p.61) by which concepts

and their proprieties and dimensions are identified from data transcribed by the researchers. This can be achieved either line by line or by focusing on main ideas in sentences or paragraphs. Each code represents a word or sentence containing a meaningful idea, and a group of codes (two or more) forms a concept. A concept is an abstract representation of an event, object or action. In open coding, events, objects and actions are compared with others in terms of similarities and differences in order to give them, when similar, the same name. The name or label that is assigned for a category should be selected logically and usually represents the data and is related to it. A reading of the literature gives the researcher an initial set of concepts that can be used. However, researchers should not be constrained by these concepts; rather, they should focus on the words and phrases used by the participants themselves. It is in this way that names are assigned to categories [7].

2. **Axial coding** is the process of reassembling data broken down through open coding. Essentially, it is the process of relating categories to subcategories. Categories are higher in level and more abstract than concepts, and are generated by a constant comparison of the similarities and differences between such concepts. This is done by using what is called the 'paradigm model', which enables the researcher to think systematically about the data and relate them to each other. This model addresses the relationships between the categories by considering the following aspects: *causal conditions*, *phenomenon*, *context*, *intervening conditions*, *action/interaction* and *consequences.*

3. **Selective coding** is the process of integrating and refining the theory. The first step in integration is identifying the central or core category that represents the main theme of the research/ phenomena. It must appear repeatedly in the data. The central category acts as a master that pulls the other categories together to form an explanatory "whole picture" by using the paradigm model. In this step, the categories are refined at a high level of abstraction. The integration is not dissimilar to axial coding except that it is done at a higher, more abstract level of analysis, and the subcategories are linked to the core category.

- Through the coding process, two analytical techniques are used. The first is *constant comparative analysis*, which is a continuous process of identifying conceptual categories and their properties emerging from data by a consistent comparison of that data. The researcher needs to be *sensitive*, which means being able to identify what data is significant and to assign it a meaning. This sensitivity comes from experience, especially if the researcher is familiar with the subject under investigation. The literature review is another source of *theoretical sensitivity*, as are the expressions of the interviewees themselves, in particular, when they repeat the same phrases and concepts. The other technique is the *asking of questions*. Once the researcher names the concept (event, idea, action and incident), he or she asks questions such as what an object is and what it represents.
- *Conceptualisation and abstraction*: Grounded theory aims to develop theories and concepts that can be generalised and applied to other situations. The generalisability of the grounded theory is partly achieved through a process of abstraction by moving from a detailed description to a higher level of abstraction; the more abstract the concepts, the greater the theory applicability [7].

III. REQUIREMENT ANALYSIS

Many methodologies are used for IS development. Two major methodologies have been used for system development: structured analysis and design and object-oriented analysis and design (OOAD). Generally, regardless of which methodology is used, the core phases for system development are analysis, design, implementation and testing. The purpose of requirements analysis is to understand the business problem and the customer (i.e. organisational) needs of the proposed system. The New York State Project Management Guidebook [9] pointed out:

> "The primary goal of [requirements analysis] is to create a detailed functional specification defining the full set of system capabilities to be implemented, along with accompanying data and process models illustrating the information to be managed and the processes to be supported by the new system".

Requirements are descriptions and specifications about the functions (what the system should do) and proprieties of the system. In fact, accuracy and completeness of the requirements affect the quality of the final developed system. A systematic process for requirements analysis is also known as requirements engineering (RE).

Requirements analysis involves two main activities that are achieved by the analyst: requirements determination/ elicitation and requirements structuring. Different techniques used for requirements determination include questionnaires, interviews, observation, documents and reports, and other modern techniques such as joint application development (JAD) and prototyping.

Analysts also use different models to structure and represent the requirements such as data flow diagram (DFD), and entity relationship diagram (ERD). In the case of OOAD, the analyst uses object/class diagrams, use case diagrams, and other models. Various techniques and approaches were proposed for requirement analysis such as goal-oriented/ goal-driven requirements analysis, scenario-based requirements analysis, inquiry based requirement analysis, and ontology based requirements analysis. In this paper, we propose a supportive approach for requirements analysis using GT.

V. Related work

Many research in IS development and the software engineering field has used the grounded theory method, as there is a widely held belief that it is a reliable method by which to elicit systems and user requirements [1][2][3][4][5][6]. Galal-Edeen [10] indicated that a requirement engineer who produces a statement of system requirements is, in reality, engaged in generating "grounded theories." Grounded theory was originally developed and used in social sciences and was later adopted by other fields such as information systems and software engineering. One issue emerges from this inheritance to other fields: Can the grounded theory method be applied in requirements engineering by a systems analyst (SA) or a psychologist researcher (for example) to analyze the requirements, supposing that he/she knows the business problem and questions?

To answer this question, Carvalho et al. [11] conducted empirical research in software engineering to generate a process model using the grounded theory method. The same gathered data were analysed by two researchers. The first researcher is a psychologist with a limited background in software engineering, but with knowledge of qualitative research methods and experience in the use of grounded theory. The second researcher is a software engineer, with a solid background in software engineering and experience in process modelling. The resulting model produced by the psychologist, however, significantly differed from that produced by an experienced process engineer using the same data.

One of the main differences in the models emphasizes that modelers should not rely solely on qualitative methods to analyze process data, but rather on their experience of the research area and the technical aspects that appear in the gathered data. The psychologist was more likely to miss artifacts and activities. The notion here is that even when using qualitative research methods adopted from the social sciences, the SA should have theoretical sensitivity of the research problem in order to produce practical and relevant results. Chakraborty and Dehlinger [12] state that there is a lack of systematic procedures within requirements engineering that enable the bridging between qualitative data and the final description of the system. In addition, the focus has been on the representation of the system by UML models as an example. This leads to reduced traceability between source data (i.e., the requirements) and the final proposed models. Therefore, they proposed using grounded theory in requirements engineering to alleviate this deficiency. They provided a demonstration of how the grounded theory method can be used to interpret the requirements for an enterprise system by applying the grounded theory coding process (open, axial and selective coding) on an illustrative example (university support system). Although the illustration was useful, the authors did not highlight how elements of grounded theory (such as theoretical sampling, theoretical sensitivity, data saturation, and constant comparative analysis) can be operationalised and applied to requirement analysis, and what is the added value of its application in this context as an alternative or supportive technique to the current requirements analysis methods and techniques. The current research takes further steps to technically reveal how the concepts of GT support the requirements analysis process by providing a methodology.

VI. Requirement Analysis using Grounded theory

Figure 1 illustrates how grounded theory elements can be used as a technique for requirement analysis. As shown in this figure, the SA starts with a perception that there is a business problem or receipt of a request for proposal to modify or maintain the current system. The analyst starts without any pre-assumed functions or components of the required system. In fact, this is essential, as many information systems fail because system analysts and developers assume that the requested system is similar to ones that were already developed by them and that they know the requirements. However, by using GT, the analysts can listen to users and remain open to accepting new and unique requirements. This is the characteristic of GT that guides an SA to start without any predefined requirements, as each system has a certain specialty. Then, the analyst interacts with the users to find out what they would like in the new system. The users are selected for their relevance to the business problem by applying theoretical sampling. Sampling is theoretical based, which is helpful for identifying involved users who will interact and use the system. Identifying the right users assists the analyst in identifying the right systems requirements. This also supports the concept of a user-centered design, in that the analyst does not force his or her predefined functions/features/requirements. Requirements are collected principally from interviews, but possibly also from documents, observations and reports. The concept of prototyping for requirements gathering conforms to the concept of theoretical sampling. Analysts gather the initial requirements from the first user and the gathered requirements guide him or her to discuss them with the second user, and third user, and so on. Perhaps after that, the analyst will return to the first user to solicit feedback regarding his or her systems' needs, as it is an iterative process.

In fieldwork, the interplay between data collection and analysis is processed simultaneously by identifying the requirements emerging from the first interviews, so that they become more specified as time progresses, since the SA validates them with the next users. At the same time, theoretical sensitivity and sampling, and constant comparison between requirements (functions, processes, objects, and attributes are compared with others in terms of similarities and differences in order to group similar ones together, assign a name to them, and eliminate repeated ones) are taken into account, finally resulting in the data becoming saturated. That is the point at which no new requirements emerge. Repeating the same data during data collection advises the analyst that this requirement (function/attribute, process) is a priority for the system. In addition, this information indicates to the analyst that the data collected from the users is saturated.

A systematic process of coding begins once the requirements have been gathered. The analysts continue to apply a constant comparative by comparing concepts that have common attributes and combining them to generate a category. This category can be a class in OOAD or a super entity type in ERD. As much as the analyst conceptualizes at a higher level, he or she can generate superclasses. The outcomes from each coding step are shown in Figure 1: codes and concepts, categories and relationships between them, and categories and associated subcategories. These ultimately form the informal model. The corresponding outcomes in RA are shown in Figure 1. In open coding, the outcomes could be a list of functions, processes, entities, objects, attributes and classes. The outcome from axial coding is the association between classes (e.g., "is a") or relationship between entities. The outcome from the selective coding is a refinement of the classes and entities found in open coding to a higher level, which includes super entities, superclasses, and related subclasses, and the generalization and specialization relationships between them. The resulting categories and relationships (equivalent outcomes in RA such as classes and super entities) may not end up being fully saturated. Consequently, a second round of data collection and analysis is initiated, which leads to the developments of a new version of the model.

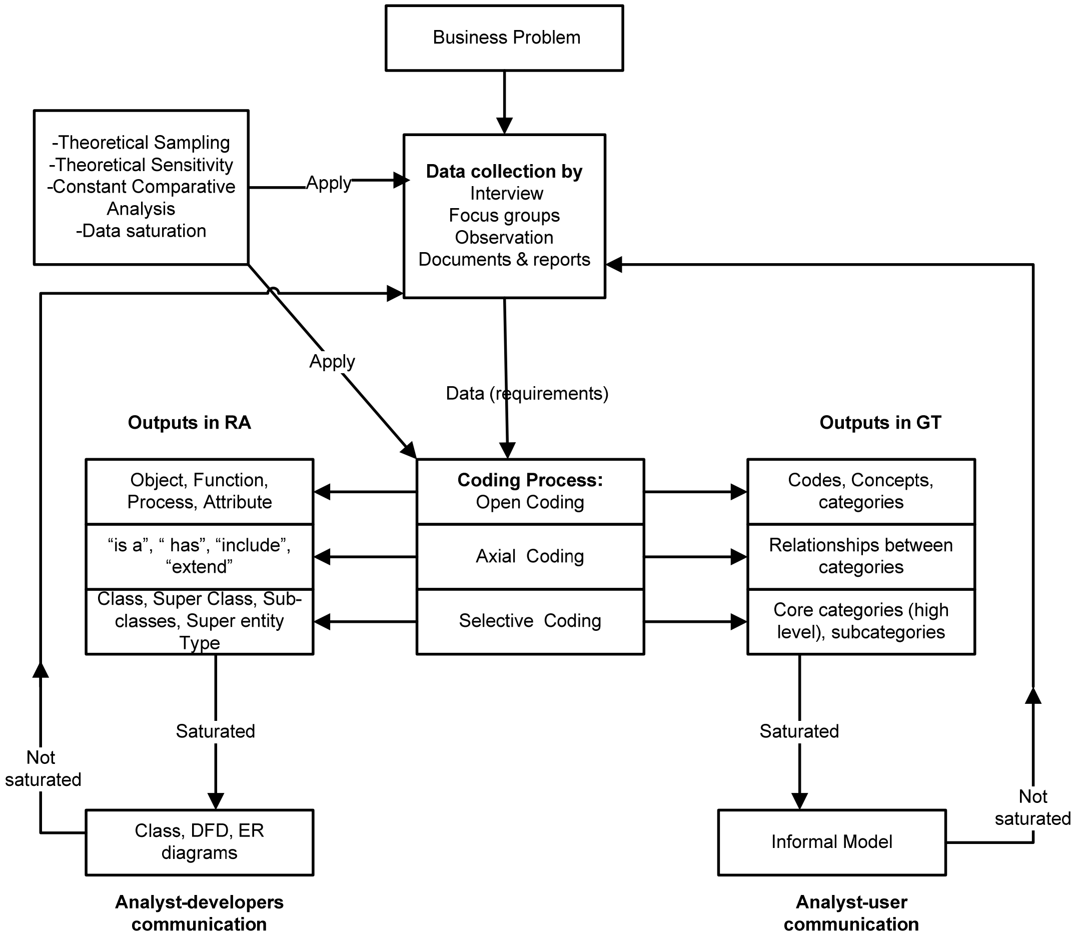

Business Problem
-Theoretical Sampling
-Theoretical Sensitivity
-Constant Comparative Analysis
-Data saturation
Apply
Data collection by
Interview
Focus groups
Observation
Documents & reports
Apply
Data (requirements)
Outputs in RA
Outputs in GT
Object, Function, Process, Attribute
"is a", " has", "include", "extend"
Class, Super Class, Sub-classes, Super entity Type
Coding Process: Open Coding
Axial Coding
Selective Coding
Codes, Concepts, categories
Relationships between categories
Core categories (high level), subcategories
Saturated
Saturated
Not saturated
Not saturated
Class, DFD, ER diagrams
Informal Model
Analyst-developers communication
Analyst-user communication

Figure 1. Using grounded theory concepts in requirement analysis

In qualitative research, in particular, grounded theory, the researcher is part of the research problem and is not independent. Hence, in this case, the analyst is part of the process and participates if something is missing from the user. Consequently, the role of the analyst is to complete the system requirements, as users may not always provide all of the requirements or may not focus on non-functional requirements such as performance and security, thus requiring interference from the analyst. However, this interference should come at the final stages after the users reveal all of their needs.

The resulting model from grounded theory is informal; this means that no standard notation or rules exist for drawing this model, as is the case in the ERD and DFD model (see an example of informal models in IS research-applied GT: [4]; [5] [13]). Informal models are used throughout all communication between the SA and the end user, which are based on simple language and representation understood by the end user. On the other hand, the equivalent model in RA such as UML models (e.g., class diagram) is easily created from the informal models and used in communication between the analyst and developers. Table 1 shows the outcomes from the grounded theory and the equivalent elements in OOAD (e.g., class/object, use case diagrams), ERD, and DFD.

Strengths of Using Grounded Theory as a Technique for Requirements Analysis:

- Analysts will find it easy to convert the resulting model of GT into a data model (ERD), process model (DFD) and other models. The GT model works as an intermediary medium to facilitate moving from a large amount of detailed data to standard analysis models.
- Models resulting from GT can be used as a communication tool between the analyst (development team) and end users. The real world represented by the informal model is closer to the end user, and they like visualisation. At the same time, it is not a formal analysis model (DFD, ERD, and class diagram), which may require some effort from the end user to understand its notation and rules.
- Following the GT procedures will assist in gathering complete requirements, and building a system based on user requests, which ultimately satisfies the users' needs. GT supports the concept of a user-centered design, as the requirements are user-based driven, and no predefined requirements are forced.
- GT will guide the analyst based on theoretical sampling to identify the relevant users who will interact and use the system, and who will explain the system requirements.

TABLE I. OUTCOME FROM GROUNDED THEORY AND THE EQUIVALENT ELEMENTS IN OOAD, ERD, AND DFD

Grounded Theory	**OOAD**	**ERD**	**DFD**
Codes (event, action, object,) concept	Object, use case, method	Entity/Entity type	Process, data store , data flow
Group of concepts (category)	Class	Entity type	
Group categories upper/general category	Supper Class	Super entity type	
Relationships between categories and sub-categories *(Consequences, causal conditions,Action/interaction, intervening conditions*)	"is a" " has" "include" "extend"	Verbs represents the association between entities	
Context (properties)	Attribute	Attribute	
Conceptualisation	Specialisation/ Generalisation	Specialisation/ Generalisation	DFD decomposing into sub-process
Data	Requirements	Requirements	Requirements
Theory/informal model	Object/class model	Data model	Process model

- Applying the conceptualisation technique by moving from the descriptive details into more abstract concepts assists in defining the system data and functions. This is also helpful in the case of using OOAD to specify the super and sub classes in class diagrams, and the objects in object diagrams that represent (instances) detailed data and a high abstract concept (category) that represents the class. The linkage between the categories and its subcategories specify the inherent relationship between the parent and child classes (inheriting classes).
- Data saturation will assist the analyst in deciding when to stop gathering requirements or direct him to identify new sources of data if there is repetition in the data.
- The core category(s) assists the analyst in specifying the functional requirements. The core category represents data that is repeated many times, which refers to the main system needs. It also represents an agreement about the indispensable functions, without which the system would be incomplete.

- All traditional techniques of data gathering are combined into one method, as GT employs a group of techniques: observation, interviews, focus groups and documentation of current systems. In addition, any gathered text is considered input to the theory/model.
- GT will assist the analyst in identifying the non-technical aspects associated with developing the system, such as user resistance change, and political and power issues emerging as a result of introducing the system within the organisation. The reason is that the nature of GT is used to understand the organisational and social phenomenon. This may not be considered by analysts who do not apply GT as their focus, rather focusing only on the technical systems requirements. Analysts can advise the decision makers and management about any potential problems associated with introducing the system. This may also help to specify an appropriate system installation and training policy. In addition, this can guide the development team to design a system that can overcome some organisational problem.

VII. CONCLUSION AND FUTURE WORKS

This research provides a concept for using GT as a supportive technique for requirements analysis. Although the author has presented logical justification for this method, this conceptual proposal must be validated by a real example. This would be interesting work for future research to apply the methodology illustrated in Figure1. Practically, this paper proposes applying GT as an effective approach for requirements analysis by showing the strengths of its application.

REFERENCES

[1] T. Linden and J. Cybulski, "Application of Grounded Theory to Exploring Multimedia Design Practices", 7th Pacific Asia Conference on Information Systems, 10-13 July, 2003, Adelaide, South Australia.

[2] M. Sorrentino and F. Virili, "Web Services System Development: a Grounded Theory Study", 18th Bled eConference eIntegration in Action, 6-8 June, 2005, Bled, Slovenia.

[3] Bo H. Hansen and K. Kautz, "Grounded Theory Applied – Studying Information Systems Development Methodologies in Practice", Proceedings of the 38th Hawaii International Conference on System Sciences, 3-6 January, 2005, Big Island, HI, US.

[4] G. Coleman and R. O'Connor, "Using grounded theory to understand software process improvement: A study of Irish software product companies", Information and Software Technology, vol. 49, 2007, pp. 654–667.

[5] S. Georgieva1, and G. Allan, "Best Practices in Project Management Through a Grounded Theory Lens", Electronic Journal of Business Research Methods, vol. 6, no. 1, 2008, pp. 43-52.

[6] S. Seidel and J. Recker, "Using Grounded Theory for Studying Business Process Management Phenomena", 17th European Conference on Information Systems, 2009.

[7] A. Strauss and J. Corbin, Basics of Qualitative Research: Grounded Theory Procedures and Techniques". SAGE Publication, London, 1990.

[8] B. Glaser and A. Strauss, The discovery of Grounded Theory". Chicago: Aldine, 1967.

[9] The New York State Project Management Guidebook, Release 2, 2003. The New York State Office for Technology**.** Retrieved 27 July, 2010, from www.cio.ny.gov/pmmp/guidebook2/SystemReq.pdf

[10] G. H. Galal-Edeen, "Information Systems Requirements Engineering: An Interpretive Approach", The Egyptian Informatics Journal, vol. 6, no. 2, 2005, pp. 154-174.

[11] L. Carvalho, L. Scott, and R. Jeffery "An exploratory study into the use of qualitative research methods in descriptive process modeling". Information and Software Technology, vol. 47, 2005, pp. 113–127.

[12] S. Chakraborty and J. Dehlinger "Applying the Grounded Theory Method to Derive Enterprise System Requirements", 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing, 27-29 May 2009, Daegu, Korea, pp. 333-338.

[13] G. Allan, "A critique of using grounded theory as a research method". Electronic Journal of Business Research Methods, vol. 2, no.1, pp. 1-10, 2003.

Mobile Agent for Orchestrating Web Services

Charif Mahmoudi
LACL, Paris 12 university
Laboratory of Algorithmic, Complexity and Logic
Computer Science Department
Paris 12 University, France
cm@ramses.fr

Fabrice Mourlin
LACL, Paris 12 university
Laboratory of Algorithmic, Complexity and Logic
Computer Science Department
Paris 12 University, France
fabrice.mourlin@wanadoo.fr

***Abstract*— Mobile agent concept can be considered as a mediator between concepts. We use this principle for management of Web Services at runtime. A set of services are placed on distinct computers, and our business process need to coordinate all the services. Mobile agents allow us to navigate on the computers to access to their local services. The role of mobile agent is not only a trigger of service, but also a transformer and a memory. A transformer because, it applies transformation onto input data and output data to adapt business interfaces. Secondly, it is also a memory because; it manages the state of the business process during its scheduling. This has an essential impact for error handling. Diagnostics are created directly by observation of location of mobile agent but also its progression into mission realization. Depending on the runtime context, computation can be restarted when resource become available. We highlight our results on poll application for training evaluation.**

Keywords- Mobile agent; web service, computation

I. INTRODUCTION

In distributed system, resources are placed on computers and some of them are shared between software. Also, a resource can be available for an application and not for another. A key concept is adaptability. First of all, we consider resource as data exposed on network or anything which is accessible through a distributed protocol. Now the problem is how to manage a computation, if access to part of input data is not possible. Runtime context can become unstable if a computation service is unavailable. Reasons are multiple: data can be locked by another application or load performance does not allow executing another local service. In that case, this local anomaly can involve a global perturbation. It is essential to solve this problem locally, near the origin of problem itself.

Our objective is to build a solution for adapting a business process in case of anomaly at runtime. It means that a strategy has to be deployed for finding another resource for instance, or for waiting its availability. Some works already exist, which use replication of resources [1]. The idea is to manage a pool of resources like data sources and a priority list. When a resource is missing, its successor is selected. Global knowledge of configuration is possible only with toy project. Moreover, a clone resource is not always a solution; when a web service is not available, a solution could be to save messages into a message queue and to replay its content later. Also, it appears that decisions have to be done depending on local information.

This decision power should have the property to move close to the location where the problem occurs. The action to invoke depends on the local properties of the wished resource. Mobile agent can become a solution to export decision power near to required resources. Into next section, we introduce mobile agent concept, then the use of web service as a distributed exposure of a software part.

II. MOBILE AGENT AND MOBILE HOST

In our working context, mobile agent is considered as a piece of executable code which has the ability to move from one computer to another one. Its business action depends on the location where it is. For instance, it can wish to access to data set for preparing a computation. Or, a mobile agent can access to an authorization service. In both cases, mobile agents arrive on a computer where a specific service is available. It means that every computers which accept to belong to a computation, have to be first identified. Then, mobile agent can migrate to these nodes, if local services are useful for their mission.

Thereby, it appears that two software components are essential into our architecture: mobile agent and agent host. A deployment of a distributed application over a network means at least one agent host per computer and a local registry where all remote services are published. The role of agent host is richer than it appears, because it receives mobile agents and negotiates their arrival.

Negotiation step is also a key feature because two sets of constraints have to be resolve when a mobile agent arrives. On one hand, there is a set of requirements due to what the mobile agent need to use locally (to read a dataset, to extract pattern from result set, etc.). To sum up, a constraint is a couple based on a resource and an action. For instance a constraint might be a file of real data and a read action. On the other hand, there is another set of permissions, managed by agent host. Each local resource has its own constraints, for instance a local file is only readable by a particular kind of mobile agent or a specific origin or code base. For each incoming agent, agent host launches a negotiator component and its conclusion is eventually an agent authorization for continuing its local activity.

Furthermore, when mobile agent is accepted by a host, its activities are under control of a supervisor, whether it transgresses the according permissions. For instance, mobile agent can try to add anything at the end of a file, or it can decide to access new local resource. Because, these permissions cannot be implied from the first ones, it should have asked these permissions before. That privileged action won't be evaluated and an exception occurred. We have already presented these works in previous conferences [2], [3] about two domains: monitoring applications and numerical computations. Mobile agents are often used as remote probes, which collect anomaly about a specific protocol, filter data and extract priority data and send them to server. For computation case, mobile agents manage all parts of a whole computation; in case of problem, we have the ability to replay since an event of the execution. They play role of distributed transaction manager. This feature is always important for the use of Web Services.

III. ADVANTAGE OF WEB SERVICE FOR COMPUTING

A. *Web Service*

Distributed service is a very simple mechanism for doing functional thing that has been created since the middle of 1970s with old technologies like Corba for instance [4]. More concretely, it is a new way to do something for which there is a need such a mathematical computation or data extraction. The difference this time is that it is based on open standards upon which the entire industry agrees: one of them is XML language.

We consider web service as a programmatic tool that allows developers to promote local resource onto a distributed protocol used by other pieces of software. As an example, matrix reduction (Choleski program on Figure 1) can be considered as a local program written in C language and a web service is built as an ideal client of that program. In that case, web service is just application of Proxy (CholeskiWS component on Figure 1) and Adapter design pattern [5]. But, web service is not only a wrapping of business code; it uses also standard kind of stream for requesting and eventually for the response (based on SOAP language). Moreover, XML language has rich properties which allow validation, transformation, enrichment and so on. Also dynamic statements can be realized at run time by the use of web service broker (see Figure 1). That one is a remote façade to for all requests about Web Services, which are managed locally (through WebService interface, see Figure 1).

Web Services standard also includes a technology for metadata information about the methods and parameters for a remote method called Web Services Description Language (WSDL) [6]. This has impact on problem recovery, when timeout is achieved or when negotiation fails. Then, an alternative has to be found (a clone of web service or a backup of current request).

Web Services are based on existing standard protocols like HTTP or FTP. This means that Web Services are stateless in nature. Statelessness is a good feature for scalability and is one of the reasons that HTTP is such a scalable protocol. However, this feature limits some of the types of code you can write as a web service. Generally, web service methods perform all the required work in a single invocation. It is not unusual to create web service methods that internally call many other methods to perform a task. For example, if you have transactional method calls, they could all execute within the same web service method call.

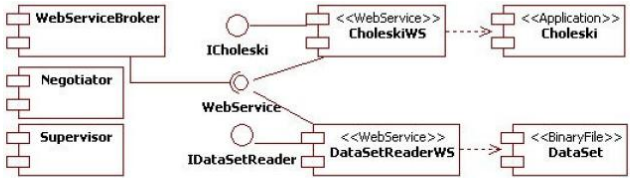


Figure 1. Components on an agent host computer (MAH)

But today, the difficulty is not to call dynamically a web service but to organize a whole algorithm based on a set of Web Services (CholeskiWS and DataSetReaderWS for instance). This means definition of business process, for instance for linear equation resolution, it starts by data extraction step, then a filters are applied, next computation can be done and finally results are consolidated by the use of previous computation. This whole algorithm should be interpreted as a main entity. Next section is about this step composition.

B. *Web Service Composition*

The composition of Web Services is more complex than a composite structure of objects. A working context has to be managed and the sequence of web service calls is not the only useful operator. Several languages were defined for the composition of web service such as WSFL/XLang [7] or BPEL [8]. Other definitions come from specification languages, such as BPMN [9] or polyadic pi calculus [8]. In a more pragmatic approach, a notation based on XML language allows easier management especially at run time. Business Process Elementary Language (BPEL) seems to have more dedicated frameworks and tools such as ODE, Orchestration Director Engine [10]. Enterprise Service Bus (ESB) are also BPEL consumers, they interpret a BPEL stream as a business process. These tools are based on a design pattern called VETRO standing for Validation, Enrichment, Transformation, Routing, and Operation. The BPEL process uses various services provided by other components. The BPEL process itself also provides a service to other components. Its main drawback is the routing step, which costs time especially when problems occur (such a failure). Programmers should develop their own strategy to detect lack of service and to recover it. This is the kernel of our current work.

In our computing context, we have several components with different size and technologies and we need to adapt them into a whole application. Also, web service is a right answer to that problem; this means interoperability of several programming languages, standardization of data exchange via XML language, and also lazy resource management. Moreover, it is natural to anticipate that the compositions are performed dynamically by a large number of end-users.

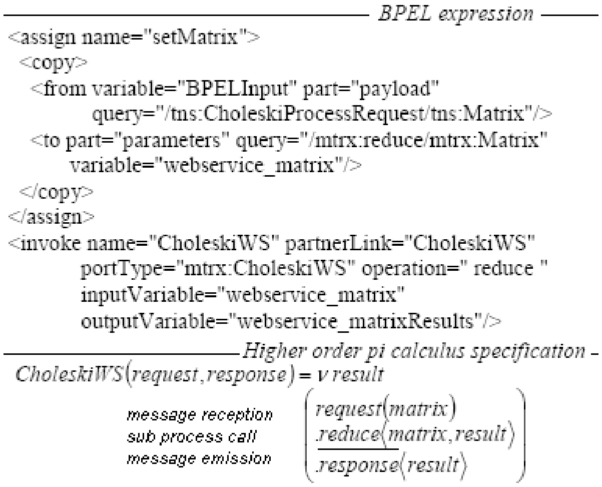
However, the current process technology based on central process engines implies the adoption of BPEL for this purpose. In fact, in BPEL, processes and composite services are synonymous. A business process is an activity, which has a hierarchical structure. It consists of a collection of sub-activities, either all to be executed in some prescribed order, such as in sequence or in parallel. When such a process is interpreted by a central engine, it involves a large set of XML messages over network. This traffic can be perturbed by data overloading on several web service engines.

When we work on the effects of a business process over the other, again the volume of messages increases always to a key limit and system test are necessary. Because a central architecture is a too strong constraint, we decided to assign the treatment of a business process to mobile agent. Its role is to interpret BPEL definition and to adapt its behavior to the current context. The main idea behind this adaptation is to move from agent server to the computer where the web service broker is. The first impact is the reduction of message number and changes of security control.

C. Mobile agents as web service pilot

Our mobile agent philosophy on BPEL is that it is a language for describing how to implement a collection of message-based communication capabilities in terms of state manipulation and messages exchanged with external services.

The core of mobile agent uses an integration layer implementation to receive and deliver messages to external parts and to get access to resources such as input dataset or web service. Also, we developed mobile agent as a BPEL engine with specific challenges such as the state management of executing process and the distribution of parallel statements. Our approach is based on our previous works about higher order pi calculus [12]. This formal specification language has its own operational semantics. We defined each operator of BPEL by the use of pi calculus language [11] (see Figure 2).



BPEL expression

```
<assign name="setMatrix">
  <copy>
    <from variable="BPELInput" part="payload"
            query="/tns:CholeskiProcessRequest/tns:Matrix"/>
    <to part="parameters" query="/mtrx:reduce/mtrx:Matrix"
         variable="webservice_matrix"/>
  </copy>
</assign>
<invoke name="CholeskiWS" partnerLink="CholeskiWS"
          portType="mtrx:CholeskiWS" operation=" reduce "
          inputVariable="webservice_matrix"
          outputVariable="webservice_matrixResults"/>
```

Higher order pi calculus specification

$$CholeskiWS(request, response) = \nu\, result \begin{pmatrix} request(matrix) \\ .\overline{reduce}\langle matrix, result\rangle \\ .\overline{response}\langle result\rangle \end{pmatrix}$$

Figure 2. Transformation from BPEL language pi calculus

We have already implemented a pi calculus interpreter from a reference operational semantics [11]. Now, we have implemented mobile agent behavior from these formal rules as an application of Interpreter pattern [5]. An agent becomes a BPEL script evaluator. From BPEL expression (see Figure 2: a piece of a specification of Choleski process is displayed), we used an operational semantics of BPEL language based on pi calculus. Then, mobile agent can interpret script as its mission. During this evaluation, intermediate states are saved. This will be powerful information for future comparison or equivalence detection.

In next section, we present the structure of our application and its deployment over network. The location of service forces specific placement and impose roadmaps for mobile agents.

IV. Architecture

We adopted a two level specification [12]. The upper one defines software architecture and how components are defined with stereotype. This description is useful for understanding technical frameworks. The lower level of specification describes how our components are placed onto the node of network and it gives a map of available services.

A. Software architecture

1) Overview

Main part of our platform is called mobile agent server. It first, receives requests from client and delegates given activity to a mobile agent. At the beginning, a mobile agent is enough for a treatment, but it could need help if the business process is too complex. This means that it is not a simple sequence of web service calls.

BPEL is based on Web Services in the sense that each of the business process involved is assumed to be implemented as a Web service. Also, as mentioned before, all elements of a computation have to be equipped for their future use. Also, when a request is received about a given business process, we built a new business space, called BS (see Figure 4). This distributed structure is useful at runtime, because it contains all parts of the execution context and it manages some aspects such as transaction management, state backup and part of safety control.

We have already defined BPEL scripts for some of our business process. They are all about statistics; and more precisely two kinds of statistics descriptive and inferential. They allow us to create graphs and charts, averages, dispersion of data, probability and its distributions etc on large data sets. Our experience about BPEL language allows us to define two different types of business processes: executable processes and abstract processes. Our models are based on actual process of statistics domain. Also, we declared only executable process which can be executed by an orchestration engine at least.

2) BPEL mobile engine

To define business processes, BPEL describes a variety of elements: the actors in a business transaction, the messages that need to be transmitted, the type of Web Services that are required and the kinds of Web Services connections that are required for operations. All these data pilot the behavior of our agent which can find relation between the BPEL process definition and the WSDL which defines the interface details of the Web Services.

For all of the local action our implementation of BPEL language is quite similar to official reference. But when it's about invoke statement or data collection; we replace a remote call with an agent migration and a local call.

Of course, our agent uses same standard languages (SOAP, WSDL and UDDI), but we reduce message volume and we consider as pre treatments some aspects which were evaluated before during a call. This takes into account a part of security controls, event tracking, etc. As example, the interpretation of invoke element is one of the main changes.

This XML block (see Figure 2) contains all the details to realize a service call. Also, the agent divides this operation into two steps: migration and call. If the url connection is not valid and exception is raised and the business space is notified on this problem, then a message is logged into anomaly report. This will be the trace that a business process failed and a robot could be collected them and a recover strategy could be implemnted.

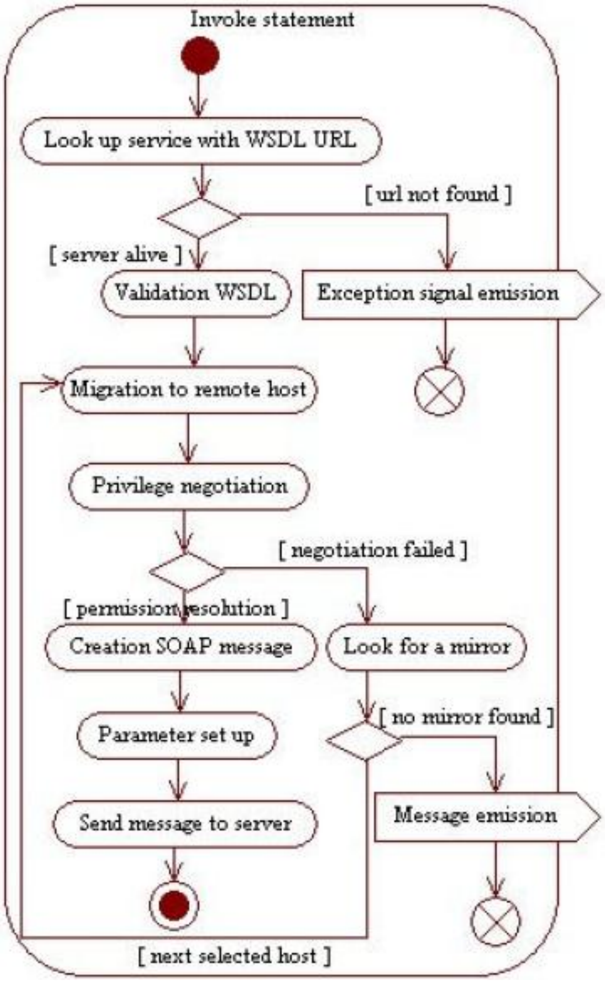


Figure 3. State chart for Invoke command interpretation

When the url is valid, it asks remote web server for receiving WSDL stream (see Figure 3). Then, after its reception, he uses the service tag (from WSDL stream) to find out its destination. It notifies its business space about its future migration and look up the acceptance service of its next destination. If this technical service is available, it will move from its current host to the host where the wished service is published.

For this migration, mobile agent needs a technical service, called acceptance service, which has to be defined on next host and published into local register. An acceptance service allows knowing if a mobile agent has enough rights to perform all the actions onto the current host. Our objective is to prevent anomaly before raising an exception, when mobile agent does not have enough permission to access to local resources. Controller exception can not be totally suppressed because, mobile agent can access to a local resource which changes its mission (for instance, a file contains a list of email addresses and they will be used to notify that the mission is ended). The negotiation between agent host and mobile agents looks like a frontier where all incoming agent are placed into a queue. This is a partial filter against access violation and misleading actions.

3) Negotiation at entrance

When an agent host receives a mobile agent, its role is to check the action list, the mobile agent wants to realize. And then it validates with the permissions, it has for this agent. If these permissions are not sufficient for the evaluation of the agent mission, the negotiation fails (see Figure 3) and mobile agent has to find another agent host where the service is declared.

For the permission assignment, mobile agent has to have certificates and details about its origin and its owner. When agent host accepts this new worker, it is launching in parallel a supervisor component to observe what mobile agent will really do during its mission. This is particularly crucial when mobile agent sends requests about importation of other agents or when it uses value of local resource.

Main events are recorded into local register: agent importation and exportation, but also negotiation success and failure, local resource access and extra data about current business process. This is used to trace behavior and also to look for a better execution time. Blockings occurs when resources are not shared, then object locks are declared. These enable multiple agents to independently work on shared data without interfering with each other. The mutual exclusion refers to the mutually exclusive execution of monitor regions by multiple mobile agents. At any one time, only one agent can be executing a monitor region of a particular monitor. If two mobile agents are not working with any common data or resource, they usually can't interfere with each other and needn't execute in a mutually exclusive way.

Agent host can apply an order among all current mobile agents. A higher priority agent that is never blocked will interfere with any lower priority agent, even if none of the agents share data. The higher priority agent will monopolize the CPU at the expense of the lower priority agents. Lower priority agents will never get any CPU time. In such a case, a monitor that protects no data may be used by agent host to orchestrate these agents to ensure all agents get some CPU time. The monitor strategy is based on quota distribution which depends on past analysis of traces of event. So, a host can privilege current agents compared to the new ones. Similarly, a host can fail negotiation because it guesses; there have already been enough agents. Idem, in case the negotiation fails, mobile agent is looking for a mirror site (see Figure 3).

B. Material architecture

Material architecture is to model the physical aspect of an object-oriented software system. It models the run-time configuration in a static view and visualizes the distribution of components in an application. In most cases, it involves modeling the material configurations together with the software components that lived on. In our working context, this involves the display of main services which are deployed on computer of the business space. The current work doesn't take into account new materials during execution.

1) Basic services

In a distributed system, our basic schema of services contains two kinds of lookup services on each used computer. Technical services are registered into Jini lookup service [13], while Web Services and business services are published into UDDI registers [14]. Both kinds of registers need front web server and sql database service as persistent layer. Of course, more technical services are present for transaction management and security control, but the size of the document does not allow us to detail more.

2) New services

The Figure 4 depicts three kinds of node: agent server (MAS), agent host computer (MAH) and UDDI register node (MAR). Of course a concrete computation net has more than one node per category, but this figure allows us to explain concrete scheduling of a business process from client request.

The agent server is a facade which receives all business requests. Depending on business process, a business space is created and equipped to receive the whole scenario. This means a transaction can be open if it is required by the business process. Then, evaluation of BPEL script can start with a first lookup into UDDI registers. By the end of its mission, mobile agent will come back to agent server and it will notify the business space about its stop.

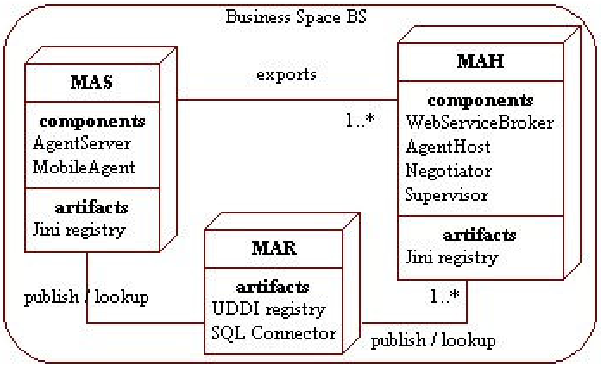


Figure 4: Deployment diagram of a mobile agent application

V. RESULTS

Our experiments show new interesting functionalities, main result is the ability to replay a business service following the same strategy as the first time. Performance studies need to replay a process or to choose between several strategies. Our second result is about anomaly tracking and more particularly how to resume from blocking states.

Enterprise service bus help to debug by the use of save SOAP messages. This information is not complete enough to understand the reasons of a problem. Often, it is concretion of a sequence of events. Also, because agent host saves locally trace of events, it becomes easy to build prefix event sequence from a given event. We developed specific mobile agent for event collection. They allow developer and architect to build a map of distributed events.

Finally, our recovery process was of great value in case of anomaly. Mobile agent can decide to look for a clone of given service and then continue the evaluation of BPEL script. This new service depends on the agent host where mobile agent is.

VI. CONCLUSION

To sum up, we consider our experiments as a validation of new ideas about mobility of actions. We knew that adaptability is a key feature into distributed system on wide network. We establish that mobility of code by the use of agent is a right way for finding alternative when a problem occurs. It becomes crucial to manage set of agents in order to improve negotiation step and supervision.

REFERENCES

[1] Jamali, N. and Xinghui Zhao. 1993. “Self-Adaptive and Self-Organizing Systems”. *SASO '07. First International Conference on*, 9-11 July 2007, 311 - 314. DOI= 10.1109/SASO.2007. pp. 49.

[2] Cyril Dumont and Fabrice Mourlin: Space Based Architecture for Numerical Solving. CIMCA/IAWTIC/ISE 2008: pp. 309-314, 2008.

[3] Mâamoun Bernichi and Fabrice Mourlin: Software management based on mobile agents. ICSNC 2007: 64-74.

[4] Robert Orfali, Dan Harkey and Jeri Edwards, CORBA Fundamentals and Programming, edited by Jon Siegel, publishing by John Wiley & sons, NY, ISBN 0471-12148-7.

[5] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides, "Design Patterns: Elements of Reusable Object-Oriented Software", ISBN 978-0201633610 , ISBN 0-201-63361-2, Addison Wesley Professional (Nov 10, 1994).

[6] Benslimane, Djamal; Schahram Dustdar, and Amit Sheth (2008). "Services Mashups: The New Generation of Web Applications". IEEE Internet Computing, vol. 12, no. 5. Institute of Electrical and Electronics Engineers. pp. 13–15..

[7] Frank Leymann, Web Services Flow Language (WSFL 1.0), IBM, May 2001. Web Services Flow Language (WSFL 1.0)"

Frank Leymann] URI:http://www.ibm.com/software/ solutions/webservices /pdf/WSFL.pdf.

[8] Yuli Vasiliev, "SOA and WS-BPEL: Composing Service-Oriented Architecture Solutions with PHP and Open-Source ActiveBPEL", Packt Publishing (September 10, 2007), pp. 316, ISBN-10: 184719270X, ISBN-13: 978-1847192707.

[9] Thomas Allweyer, "BPMN 2.0 Introduction to the Standard for Business Process Modeling", ISBN 978-3-8391-4985-0, Paperback, pp. 156.

[10] Petri Nets and Other Models of Concurrency – ICATPN 2007, Lecture Notes in Computer Science, 2007, Volume 4546/2007, 263-280, DOI: 10.1007/978-3-540-73094-1_17

[11] R. Milner, J. Parrow, and D. Walker (1992). "A calculus of mobile processes". Information and Computation 100 (100): 1--40. doi:10.1016/0890-5401(92)90008-4.

[12] Mâamoun Bernichi and Fabrice Mourlin, "Two Level Specification for Mobile Agent Application," icons, pp.54-59, 2010 Fifth International Conference on Systems, 2010

[13] Scott Oaks and Henry Wong, "Jini in a Nutshell", Publisher: O'Reilly March 2000 pp. 413 Print ISBN:978-1-56592-759-9, ISBN 10: 1-56592-759-1

[1] Tyler Jewell, David Chappell, "Java Web Services", March 2002, 0-596-00269-6, 276 pages, O'Reilly & Associate

Architecture Patterns for a Ubiquitous Identity Management System

Anders Fongen
Norwegian Defence Research Establishment
Norway
anders.fongen@ffi.no

***Abstract*—The design of an Identity Management System (IdM) must strike a balance between protocol overhead, software footprint and security level in order to operate successfully under the resource constraints found in mobile and wireless systems. But, what is good for a constrained system is also good for everyone else, in the sense that reduced overhead benefits all business application processing. This paper contributes to the discussion of IdM construction by suggesting patterns that preserves existing investments, maintains adaptability, scalability and modularity of the IdM. It also provides a discussion where security level is balanced with other non-functional requirements, which is seen less often in security research. A prototypical IdM systems built upon the proposed principles is also presented to some detail.**

Keywords*-*Identity management

I. Introduction

Identity Management (IdM) refers to the arrangement of manual procedures and software components which are needed to identify and control the use of computing resources. IdM also supports the privacy and integrity of data.

IdM involves tasks like key and certificate generation, role and attribute management, authentication operations, access control and auditing. Together, the IdM comprises a large set of distributed software components and a number of networking protocols. Besides, the components of an IdM will interface to business components and its management procedures will interface with procedures involving matter of law, human resources and business ethics [1].

Consequently, it is crucial to the successful deployment of an IdM that certain design principles are observed. The purpose of this paper is to present a set of design guidelines which serves as *design patterns* for the construction of an IdM.

Identity Management should maintain the following set of design patterns:

- Use existing Public Key Infrastructure (PKI)
- Federate domains for guest access
- Roles matter, not identity
- Domains are autonomous
- Avoid belt-and-suspenders protocols
- Trust has a lifetime
- Limit the unconditional trust

By following this list of rules the identity management system requires less connectivity and bandwidth, and consequently is better fit for a mobile and wireless computing environment. The identity management system becomes applicable to a wider range of environments, thus the use of the word ”ubiquitous” in the title.

The remainder of the paper is organized as follows: Section II explores briefly the design patterns just listed. Section III presents a short survey of existing IdM architectures. Sections IV and V present the Gismo IdM system and its protocols, followed by a section with some conclusive remarks.

II. Candidate Design Patterns

A. Use Existing PKI

In most organizations, there are formal procedures related to employee and inventory information. Quality of that information is crucial in order to detect fraud and theft. Some organization have also implemented a Public Key Infrastructure (PKI) (or are planning to do so) for the purpose of public key management. A PKI in operation will be the result of a long planning process, complicated software deployment and configuration, and the development of several new managerial interfaces between the HR and IT departments. An operational PKI represents a significant investment that should be built upon when an IdM is being developed.

B. Federate Domains for Guest Access

Back then, there was the idea of a PKI which could operate on a very large scale, e.g., for every citizen of a nation, and serve a large number of applications. Today, a national PKI is believed to provide keys only for limited communication between citizens and public sector. Other PKIs will provide keys for banks, other for Internet shopping and again, others for professional communication.

IdMs have the potential to bridge the gaps between different *domains* of key administration, meaning that they can manage trust relations between domains in an articulated manner. Domain federations allow subjects to bring their credentials across domains for controlled access and trust.

C. Roles Matter, not Identity

The rule in “traditional” user management in standalone computers has been never to grant privileges directly to subjects. Subjects should be assigned to *groups*, and groups given access rights. *Role-based* or *attribute-based* access

ISBN:978-1-61208-114-4

control [2] is built on this idea, which is several decades old and well proven.

This separation makes lots of sense in a distributed environment. It means that only the IdM service needs to maintain actual identities, whereas the providers of business services maintain the mapping between access rights and *roles* or *attributes*.

In a domain federation, this separation is crucial. Although some IdM systems for domain federations provide mapping between user names on different systems (hopefully for legacy reasons only), the only scalable approach is to allow the users to be represented by a set of roles/attributes.

D. Domains are Autonomous

All domains of identity management wish to be autonomous. They establish identification procedures based on their own business and security policies, according to national legislation and the ethics of their profession. They will determine what services will be made available to residents and guests of the domain. They decide by themselves the access rights that are associated with subject attributes. *Domain federations should not impose federated authorities*.

Another matter of domain autonomy is role (or attribute) *privacy*. The attributes associated with a subject may be of sensitive nature, since they may reveal information about the subject's authority. Consequently, the domain must be in control of how attributes are exposed inside and outside the domain [3].

E. Avoid belt-and-suspenders protocols

The network cost associated with the operation of a PKI is substantial, and inhibits this operation in parts of the network where the bandwidth is narrow or the connectivity is episodic [4]. Networks with such conditions include wireless mobile networks (MANET) and military tactical networks. Wireless networks are more exposed to intrusion attacks than a wired network. Ironically, the parts of the network that really need the protection that a PKI could offer, are thus the parts least suited to use it!

Consequently, the networking protocols (and the security policies they result from) must ensure that the network capacity requirements does not exceed the expected performance of the technology in place. This may require a closer inspection of the risk estimate, and some belt-and-suspenders security requirements may have to be relieved.

F. Trust has a lifetime

This pattern is firmly related to the previous paragraph. It is a matter of reducing the network traffic through a "trust has a lifetime" decision. For example, a validated public key is believed to be valid for some duration, and will not need to be revalidated in this period. This principle is well established through the distribution interval of certificate revocation lists (CRLs).

This principle reduces the number of necessary operations from both the client and the server to the security services. They do not longer need to receive credentials and validation information for each business operations, since this information can be cached and re-used for a while.

G. Limit the unconditional trust

The last design pattern is related to the number of *trust anchors*. A trust anchor is a subject whose signature is unconditionally trusted. All trust relationships are derived from a trust anchor through a chain of signatures. The security of the entire system collapses if a trust anchor gets compromised. Therefore, the number of trust anchors should be low for the sake of system security and robustness [5].

III. Existing IdM Architectures

The proposed design is related to the SAML 2.0 architecture for federated identity management [6] and the WS-Security [7] and WS-Trust standards [8], but this model aims to provide better answers to the challenges of mobile and tactical environments.

Based on a survey of existing models for federated identity management like Liberty Alliance [9], Shibboleth [10], and OpenID [11], it is an observation that they are *not* well suited for low-bandwidth, mobile or disadvantaged networks for the following reasons:

- They require much connectivity, in the sense that every new connection with a service involves operations on the identity provision servers.
- They require a coordinated replication of user registries, so that an excessive amount of work is needed to maintain user information in a highly dynamic network.

The same survey also indicates that these approaches to identity federation create rather strong coupling between the security domains; they either require mapping between local user identities, or mapping between local and federated identities. Both approaches could be replaced by an RBAC (role based access control) [2] arrangement that removes the need for replicated user identities in order to weaken the coupling between the domains.

Please observe that the term "federated" in this paper refers to federation of servers from different communities with different security requirements. The term "federation" as used in the related literature may refer to a group of servers in the same domain, in which case coordination is a much simpler problem.

IV. The Gismo Architecture

Following the guidelines given in Section II, a IdM prototype was built for the purpose of experimentation. The prototype has been implemented in Java for operation in a Web Services environment. The protocol data units have accordingly been coded in XML syntax, to the extent possible using suitable XML standards (SAML, WS-security, WS-addressing, etc.).

The functional components of the Gismo[1] IdM and their relations are shown in Figure 1.

A. The Domain

In the context of this project, the term “Domain” means a population of services and subjects with the following set of properties:

- Members (services and subjects) belong to one domain only
- All members of a domain share the same *Certificate Authority* (CA) and *trust anchor*.

B. Community of Interest

Inside a domain, there are one or more *Communities of Interest* (COI). For each COI, there is one *Identity Provider* (IdP). Members of a COI are services and subjects, which can be members of several COIs (inside the same domain). Two subjects can have authenticated communication (client-server or message exchange) if they are members of the same COI, or members of two COIs with a *trust relationship*.

C. The Identity Statement

The Identity Statement (IS) is similar to a public key certificate in the sense that it attests a binding between a public key and the identity information of the “owner” of the private key. In addition, the IS contains a set of roles/attributes associated with the represented identity.

The identity statements are issued and signed by the identity provider, and are therefore valid only inside the COI served by that IdP.

There is *no revocation checking* associated with identity statements. An IS is therefore meant to be short-lived, i.e., expire after a duration comparable to the issue interval of certificate revocation lists.

D. The Identity Provider

The Identity Provider (IdP) is a CA-like service which issues identity statements for members of the COI. Upon requests from subjects, their IS are issued and returned to the clients for use in different authentication procedures.

Another important task for an IdP is to provide identity statements for *guests*. If a subject sends an IS issued by an IdP with which there exists a trust relationship, a *guest IS* is issued. The guest IS contains the same information as the original IS, except that attributes may have been added or removed. It also bears a new signature.

E. The Authentication Protocol

Several authentication protocols have been proposed under the Gismo IdM project, with the goal to reduce the number of protocol round trips and to explore the relation between network cost and risk.

Protection against replay attack in authentication protocols is quite costly, since it requires the service to remember previous requests (identified by e.g., nonces) for the maximum allowed clock skew period, *also during a crash* (i.e., across “incarnations”). This is a hard problem, since lightweight service platforms (like embedded systems) may not be able to offer the transactional stable storage which is needed to implement this mechanism.

Under the conditions that the service is stateless, i.e., a request is not altering the state of the system (i.e., a lookup service), replay protection is not needed, provided that only the intended client can read the reply. The authentication protocol may under such circumstances simply encrypt the reply with the public key of the client to achieve this effect.

Another matter is the number of protocol round trip. During a separate authentication phase, client and service can mutually authenticate themselves before the actual service call is made. A more effective approach would be to piggyback the client authentication in the service request, and the service authentication in the response, as shown in Figure 3. This reduces the number of round trips, but the risk remains that a mere request to a fraudulent service may compromise sensitive information. This is (in the author’s opinion) a far-fetched risk: An attacker who is able to stage such an advanced attack would benefit more from simple eavesdrop than a “hit and run” attack (a fraudulent service which is not able to authenticate itself would trigger an intrusion alarm and subsequent hunt for the intruder).

Under other conditions, e.g., a protected and authenticated conversation, a more traditional approach would still be the best choice where mutual authentication and session key exchange takes place before the information flow starts.

F. Cross-COI Operation

An important property of an IdM architecture is the ability to offer services to members of a different organization in a well controlled manner. This property is an important part of the Gismo IdM and is based on *guest IS* to indicate the approval of a guest identity, and the *cross-COI IS* to indicate the trust relationship between to COIs. Together with an RBAC/ABAC based access control framework, guest may be given access under a fine-grained policy.

Trust relationships between two COIs are expressed by a pair of IS where they attest each other’s public keys and identities. These *cross-COI ISes* links the signature on an IS from a remote COI to the trust anchor (the CA) of the local COI, and conveys the delegation of trust from the local IdP to the remote IdP.

[1] “Gismo” is the acronym for the Norwegian expression “Fundamental IT security for mobile operations”

G. Proof of validity

Members of a COI trust the CA of the domain, i.e., the CA is their *trust anchor*. They also need to trust the IdP, since the identity statements bear its signature. The IdP may be declared as a trust anchor, too, but there are good reasons (mentioned in Sect. II-G) why the number of trust anchors should be kept to a minimum

The trust in the IdP could be derived from the CA through a PKI-style *validation* of the IdP's certificate, which is not a desirable solution for reasons of network economy and architectural coupling.

Rather, it is a preferred solution that the IdP is the only central service that the members know about, and that the IdP itself can provide a "proof of validity" for its key and certificate. Given this proof, any member can conclude that the key of the IdP is authentic and not revoked at the moment.

The proof of validity (POV) may have several forms, depending on whether the CA is the direct or indirect issuer of the IdP's certificate. It should contain all certificates from (and including) the IdP's certificate and up to (not including) the trust anchor (normally the root CA). It should also provide proof that none of the certificates on this list is revoked at the moment.

The proof of non-revocation cannot be a revocation list, since it is not possible to provide positive information in it, only negative. What is needed is a positive revocation status (meaning not revoked), which can be the output of a *validation server*, e.g., one that is based on the SCVP or OCSP protocols. These responses must be signed with a key that is attested by the trust anchor through a signature chain.

The CA could issue an SCVP response on a regular basis which the IdP could hand out on demand, but that would require a custom built CA and a violation to the rule in Sect. II-A. Standard PKI services must be used, which would likely be the signed and timestamped output from certificate status providers (using OCSP) if available. If the trust anchor refuses to issue revocation status in any other form than through CRLs then one is out of luck and needs to declare the IdP as the trust anchor for the members of the COI.

H. Attribute Protection

Subject attributes in an IS (elsewhere also called *roles*) are name/value pairs which can describe any aspect of the subject. It can be used to store the subject's native language in order to improve the user interface of a service etc., or describe the subject's authorizations for access control support.

Attributes may contain sensitive information which should be adequately protected. The ultimate protection is for the IdP to issue an IS for the purpose of one particular service, encrypted with the public key of this service. On the other hand, this arrangement makes the IS non-cacheable and requires frequent connection to the IdP, effectively making it into a single point of failure.

The Gismo IdM approach is taking a middle road. An IS issued for use in a COI should be cacheable and be used for all services and conversations withing the COI until the IS expires. When an IdP receives an IS from a guest who is requesting a guest IS, only attributes marked for export are copied into the guest IS, the other are removed. Since there exists a trust relationship between these two IdPs it is reasonable to trust a "foreign" IdP to do this honestly and correctly. It is also reasonable to allow services and subjects in the same COI to share attribute knowledge, since the COI membership of shared goals and shared responsibility also implies a level of trust (and since they might obtain this information anyway through listening on the shared data links).

V. Protocol and Data Structure Details

At this point the design principles and the main functional components of the Gismo IdM have been explained, and the paper will commence with a description of the data structures and protocols in greater detail.

A. The Identity Statement

As previously described in Section IV-C, the authentication mechanisms relies heavily on the data structures called *Identity Statement* (IS). Formally, the identity statement of principal x signed by the IdP of COI a is denoted $(Id_x)_a$ and has this structure:

$(Id_x)_a = Name_x + PublicKey_x + Attributes_x + Timestamp + Serialnumber + Signature_a$

$Attributes_x$ denotes a set of name-value pairs which describes the roles etc. of the subject. It may be used for access control purposes. $Signature_a$ indicates that the entire statement is signed by the IdP of CIO a. The IdP of COI a will from now on be denoted IdP_a.

In the proposed system, the identity statement is formatted according the the SAML 2.0 syntax requirements, which means that it is coded in XML. The SAML assertion is used in a so-called "Holder of Key" mode, which means that the authentication process requires a demonstration of the private key corresponding to the public key bound in the identity statement.

B. Identity Statement Issuance

The discussion in Section IV-H identified the need to protect subject attributes outside the Community of Interest (COI), which means that only members of a COI should not be allowed to ask the Identity Provider (IdP) for an IS regarding a COI member.

There is no easy way to distinguish a member from a non-member (without a costly authentication phase). The design choice has therefore been to issue an IS only to the subject itself. This requires a straightforward SSL-based

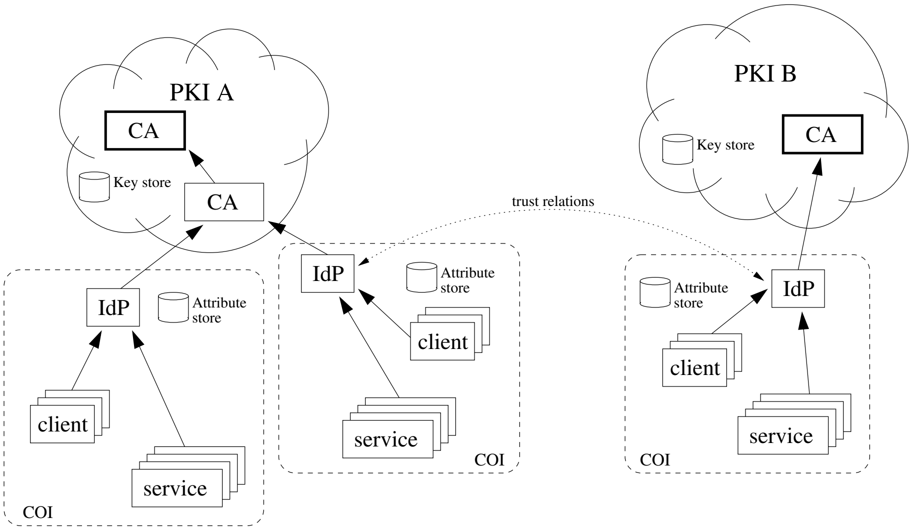


Figure 1. The functional components of a federated IdM. Observe that the IdP serves one single COI, and the trust relations are formed between COIs, not domains. Key management is handled by the PKI whereas the attribute management is done by the IdPs on the COI level

client authentication based on the client key pair related to the IS. An alternative approach would be to encrypt the IS with the subject's public key, but that is less flexible towards future policy changes.

A part of the service semantics is that the subject's key pair is *validated* before the IS is issued. If the key pair is generated by a PKI (as suggested in Section II-A) the IdP should use the available PKI-based validation mechanisms for this purpose, and deny the IS request if the key is invalidated or revoked.

C. Issuance of Guest Identity Statement

The IdP is responsible for the issuance of guest identity statements as explained in Section IV-D. Presented with $(Id_x)_a$, the IdP_b (IdP of COI b) can issue the identity statement $(Id_x)_b$ provided that there exists a trust relationship between COI b and a expressed by an identity statement issued by IdP_b with IdP_a as the *subject*. This is called a cross-COI IS and expressed as $(Id_a)_b$. With the guest IS $(IdP_x)_b$, the subject x which is a member of COI a, can authenticate itself to members (e.g., services) of COI b.

For two-way authentication in a guest COI, e.g., for the client from COI a to trust the signed response from a member of COI b, the reverse cross-COI IS is needed, termed $(Id_b)_a$, to link the signature key to the client's trust anchor. Therefore, $(Id_b)_a$ is included in the response of the guest IS issuance. $(Id_b)_a$ is issued to IdP_b by IdP_a (as a normal IS issue) and stored by IdP_b for the purpose of guest IS issue.

Figure 2 illustrates the guest IS issuing protocol as a two stage process involving two IdPs. Key validation takes place only in the first stage. The required proof of validity (Section IV-G) is assumed to have been issued at an earlier occasion.

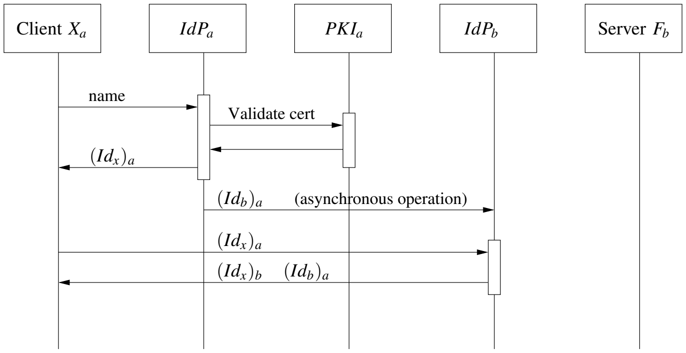


Figure 2. The identity statement issuing protocol. The IdP of COI A, termed IdP_a, issues a "native" identity statement to the client, which is given to IdP_b which in turn issues a guest identity statement. The term PKI_a denotes a set of certificate validation services in domain a.

TABLE I
ABBREVIATIONS USED IN THE FIGURES

Client X_a	Client X of COI a
IdP_a	Identity provider of COI a
PKI_a	Validation services in domain a
Server F_b	Server F in COI b
$(Id_x)_a$	Identity statement for identity x, issued by IdP_a
$(msg)S_x$	Message msg signed with private key of x
$(msg)E_x$	Message msg encrypted with public key of x

D. The Authentication Protocol

Section IV-E provides a discussion on the effectiveness of authentication protocols. The Gismo IdM offers a range of authentication protocols with different properties, two of which are presented in this paper. Figure 3 shows a protocol suited for a server with the necessary resources to implement replay protection. The data elements needed for mutual authentication (signature, timestamp, nonce, servername) are

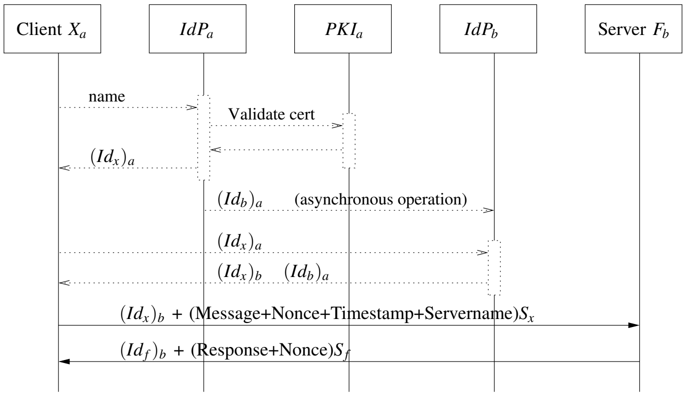


Figure 3. The authentication protocol for the stateful service. Both the request and response are signed with the sender's private key as a part of authentication process. A timestamp, a nonce and the server's name is included for replay protection.

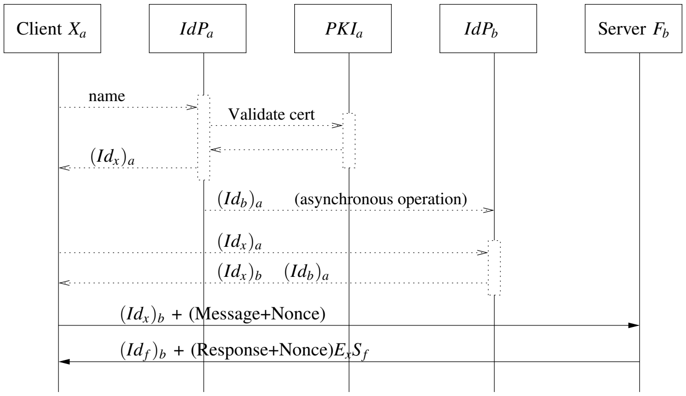


Figure 4. The authentication protocol for the stateless service. There is no replay attack protection since they are not considered as threats, but the response need to be protected for reasons of response replay and information compromise.

piggybacked on the request and response messages in order to save a protocol round trip. The remaining security risk, which results from this choice is marginal, as pointed out in Section IV-E.

Figure 4 illustrates the much simpler authentication to a stateless service. All requests are processed since they do not alter the system state (other than consume resources), but the authentication requirements are enforced through the encryption of the response. The response is signed for the purpose of server authentication, and includes a nonce for protection against response replay. The nonce is not remembered across invocations and introduces no state space in the sever.

E. Notes on Implementation

The Gismo IdM is implemented as a proof of concept in Java and targeted for web services use. It employs relevant WS standards (SAML, WS-Security, WS-Addressing, etc.) and implements the authentication protocols as *WS Message Handlers*, including the client interface to the IdP services.

VI. Summary and Conclusions

In the course of this paper, a slightly different approach to the construction of IdM has been made, where a balance between security level and other non-functional requirements have been sought. A prototypical system built upon the proposed principles has been presented in some details.

Ongoing efforts on the Gismo IdM includes integration with Role Based Access Control, combining IdM with arrangements for object type enforcement and principles of least privilege. Future plans include porting the software to the Android platform for study of its performance in mobile computing environments.

References

[1] N. Delessy, E. B. Fernandez, and M. M. Larrondo-Petrie, "A pattern language for identity management," in *Proceedings of the International Multi-Conference on Computing in the Global Information Technology*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 31–31.

[2] R. Sandhu, D. Ferraiolo, and R. Kuhn, "The NIST model for role-based access control: towards a unified standard," in *RBAC '00: Proceedings of the fifth ACM workshop on Role-based access control*. New York, NY, USA: ACM, 2000, pp. 47–63.

[3] A. Bhargav-Spantzel, A. C. Squicciarini, and E. Bertino, "Establishing and protecting digital identity in federation systems," *J. Comput. Secur.*, vol. 14, pp. 269–300, May 2006.

[4] A. Fongen, "Scalability analysis of selected certificate validation scenarios," in *IEEE MILCOM*, San Diego, CA, USA, Nov. 2010, pp. 1–7.

[5] C. Wallace and G. Beier, "Practical and secure trust anchor management and usage," in *Proceedings of the 9th Symposium on Identity and Trust on the Internet*, ser. IDTRUST '10. ACM, 2010, pp. 97–107.

[6] N. Ragouzis, J. Hughes, R. Philpott, E. Maler, P. Madsen, and T. Scavo, *Security Assertion Markup Language (SAML) V2.0 Technical Overview*, OASIS Committee Draft, March 2008.

[7] K. Lawrence and C. Kaler, *Web Services Security: SOAP Message Security 1.1*, OASIS Standard Specification, 2004.

[8] ——, *WS-Trust 1.4*, OASIS Standard, 2009. [Online]. Available: http://docs.oasis-open.org/ws-sx/ws-trust/v1.4/os/ws-trust-1.4-spec-os.pdf [retrieved November 9, 2010]

[9] "The Libery Alliance." [Online]. Available: http://www.projectliberty.org/ [retrieved November 9, 2010]

[10] "Shibboleth." [Online]. Available: http://shibboleth.internet2.edu/ [retrieved November 9, 2010]

[11] "OpenID." [Online]. Available: http://openid.net/ [retrieved November 9, 2010]

Instantaneous Autonomous Aerial Reconnaissance for Civil Applications

A UAV based approach to support security and rescue forces

Florian Segor, Axel Bürkle, Matthias Kollmann, Rainer Schönbein
IAS - Interoperabilität und Assistenzsysteme
Fraunhofer IOSB
Karlsruhe, Germany
{florian.segor, axel.buerkle, matthias.kollmann, rainer.schoenbein}@iosb.fraunhofer.de

***Abstract* - The Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB) deals with the interoperability of stationary and mobile sensors and the development of assistance systems, which optimize and simplify the operation of such systems. In particular one of the focuses is research on swarms with airborne miniature drones and their applications. The photo flight presented in this paper is one of the applications developed to bring the advantages of a swarm into a realistic scenario. With the aim to support rescue or security forces in action, the photo flight generates an immediate up-to-date situation picture by using an autonomous swarm of miniature drones.**

Keywords - aerial situation image, unmanned aerial vehicles, swarm, search and rescue

I. INTRODUCTION

This paper presents our most recent work on a software module called “photo flight”, which was developed as part of the ground control station AMFIS [1]. AMFIS is a component-based modular construction kit currently under development as a research prototype. It already has served as the basis for developing specific products in the military and homeland security market. Applications have been demonstrated in exercises for the EU (PASR[1] program), German Armed Forces, and the defense industry. The surveillance system AMFIS is an adaptable modular system for managing mobile as well as stationary sensors. The main task of this ground control station is to work as an ergonomic user interface and a data integration hub between multiple sensors mounted on light UAVs (unmanned aerial vehicles) or UGVs (unmanned ground vehicles), stationary platforms (network cameras), ad hoc networked sensors, and a superordinated control center.

The photo flight is a special feature of the flight route planning in AMFIS that allows the user to generate a highly up-to-date aerial picture of a predefined area in a short time. The software module itself is designed to work as independent standalone software as well as a part of the complex control system AMFIS.

After a short survey of related work an overview of the application scenarios is presented, followed by a description of the airborne platform in section IV. Section V introduces the used algorithms followed by the description of the post processing, conclusions and future work.

II. RELATED WORK

As far as we know the photo flight is a quite unique project. However, there are some projects with a similar scope.

At the “Universität der Bundeswehr” in Munich Dr. P. Reidelstuerz is developing a UAV for precision farming [2]. It is used to analyze agricultural areas from the air to find the regions that need further manuring to optimize the growth of the crop. A commercial of the shelf fixed wing model is equipped with an autopilot and either a near infrared or a high quality camera. With this technique the biomass development and the intensity of the photosynthesis of the plants can be monitored.

The AirShield project (Airborne Remote Sensing for Hazard Inspection by Network Enabled Lightweight Drones) [3][4], which is part of the national security research program funded by the German Federal Ministry of Education and Research (BMBF), focuses on the development of an autonomous swarm of micro UAVs to support emergency units and improve the information basis in case of huge disasters. The aim is to detect potentially leaking CBRNE contaminants in their spatial extent and to carry out danger analysis with the help of these data without endangering human life. The swarm is supported by a highly flexible communication system, which allows communication between the swarm members and between the swarm and the ground station.

The precision farming project as well as the AirShield project are very promising and showed first results. However, the application aim of both projects differs from ours although we plan to extend the photo flight to scenarios similar to the ones of AirShield (see Extensions and Further Work).

III. APPLICATION SCENARIOS

The security feeling of our society has significantly changed during the past years. Besides the risks arising from natural disasters, there are dangers in connection with criminal or terroristic activities, traffic accidents or accidents in industrial environments. Especially in the civil domain in case of big incidents there is a need for a better data basis to

[1] Preparatory Action for Security Research

support the rescue forces in decision making. The search for buried people after building collapses or the clarification of fires at big factories or chemical plants are possible scenarios addressed by our system.

Many of these events have very similar characteristics. They cannot be foreseen in their temporal and local occurrence so that situational in situ security or supervision systems are not present. The data basis on which decisions can be made is rather thin and therefore the present situation is very unclear to the rescue forces at the beginning of a mission. Exactly in such situations it is extremely important to understand the context as fast as possible to initiate the suitable measures specifically and efficiently.

An up-to-date aerial image can be a valuable additional piece of information to support the briefing and decision making process of the applied forces. However, helicopters or supervision airplanes that can supply this information are very expensive or even unavailable. Up-to-date high-resolution pictures from an earth observation satellite would provide the best solution in most cases. But under normal circumstances these systems will not be available. Nevertheless, it would usually take too long till a satellite reaches the desired position to provide this information. A small, transportable and above all fast and easily deployable system that is able to produce similar results is proposed to close this gap.

The AMFIS tool “photo flight” can provide the lacking information by creating an overview of the site of the incident in a very short time. The application can be used by first rescuers for example directly on site with relative ease. The results provide a huge enhancement to already available information.

Applications include support of fire-fighting work with a conflagration, clarification the debris and the surroundings after building collapses, and search for buried or injured people. Additionally the system can be used to support the documentation and perpetuation of evidence during the cleaning out of the scene at regular intervals.

Non-security related application scenarios are also conceivable, as for example the use of infrared cameras to search large cornfields for fawns before mowing or to document huge cultivated areas or protective areas and biotopes.

The photo flight tool, which was developed from a former research project, shows excellent results in the production of up-to-date aerial situation pictures in ad hoc scenarios. The intuitive and ergonomic graphic user interface allows the operator to define an area of interest and start the photo flight. The results are a number of images depending on the size of the area of interest. They are merged and geo-referenced by suitable tools.

Since AMFIS is capable of controlling and coordinating multiple drones simultaneously [5] the photo flight tool was designed to make use of the advantages and benefits of a UAV swarm. By using more than one single UAV, the same search area can be covered in less time or respectively a bigger area can be searched in the same time.

The biggest problem when working with multiple UAVs is the dwindling clarity for the operator, especially when there are different types of UAVs and payloads. The more drones used in an application the more complicated the control of the single systems gets. That is why it is most essential to reduce the working load on the user as much as possible. Therefore the idea of a self-organizing swarm is transferred to the photo flight application in order to reduce the efforts for controlling this tool to a minimum. The user only has to define the area of interest and decide, which drones he would like to use.

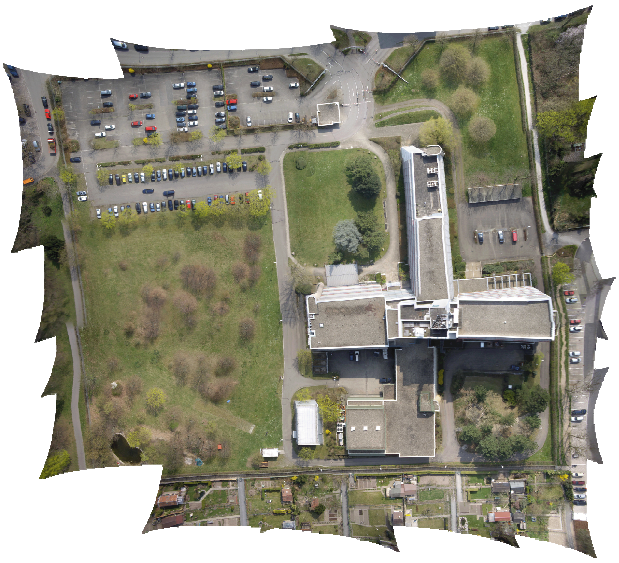


Figure 1. Situation picture from photo flight (ca. 9500 x 9000 pixel)

All additional work as for example the composition of the respective flight routes or the control of the single UAVs including the observation of the aerial security to avoid collisions up to setting the return flight is done by the application.

IV. PLATFORM DESCRIPTION

The primary aim of the photo flight is the clarification of certain areas. The used drones do not necessarily have to be identical. They also can differ in their technical configurations. Nevertheless, in this first research attempt to build a swarm, UAVs of the same type were used.

A lot of effort has been put into the selection of this flight platform. A platform that already comes with a range of sensors, an advanced control system and autonomous flight features significantly reduces the effort necessary to realize a cooperative swarm of micro drones. Furthermore, when it comes to flying autonomously, the system has to be highly reliable and possess sophisticated safety features in case of malfunction or unexpected events.

Other essential prerequisites are the possibility to add new sensors and payloads and the ability to interface with the UAV’s control system in order to allow autonomous flight. A platform that fulfils these requirements is the quadrocopter AR100-B by AirRobot (see Figure. 2). It can be both, controlled from the ground control station through a command uplink and by its payload through a serial interface.

To form a heterogeneous swarm from different UAVs, new systems were gradually integrated. Currently, beside the AR100-B there is also a Microdrones MD-400 as well as a MikroKopter with eight rotors (MK Okto). The user can identify the system by its call sign – the operation of the drone, however, remains identical, rendering the complexity of the heterogeneity transparent.

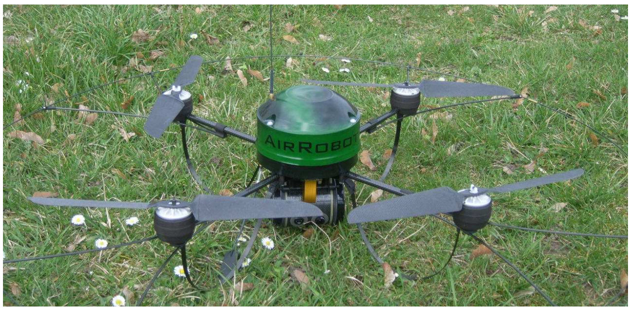


Figure 2. Sensor platform AirRobot 100-B

V. ALGORITHMS

To be able to clarify an area of interest by multiple drones the polygon defining that area must be divided into several subareas, which can then be assigned to the individual UAVs. It is important that each of the branches is economically optimized for its appropriate drone. UAVs with longer endurance or higher sensor payload can clear up vaster areas and should therefore receive longer flight plans than systems with a lower performance.

Besides, the flight routes must consider the behavior of the drones at the single photo points and their flight characteristics. Tests with the multicopter systems have shown that an optimum picture result can be achieved if the system stops at each photo point for two to four seconds to stabilize. Proceeding precisely in such a way, no special flight behavior must be considered, because the drones show identical flight characteristics in every flight direction due to their construction. Indeed, this behavior also decisively affects the operation range, because such stops reduce the efficiency of the drones. To solve this problem, a stabilized camera platform, which compensates the roll, pitch and yaw angles, was developed at Fraunhofer IOSB. Nevertheless, if the photo points are flown by without a stop, an enlargement of the flight radii must be considered at turning points. Since in future versions also fixed-wing aircrafts may be used, further attention must be paid to the fact that the calculated flight paths can also be optimized for systems with different flight characteristics.

The algorithm developed from these demands consists of two main steps. The first part is to break down the given polygon of the search area in suitable partial polygons *(A)*. Then, the optimum flight route per partial polygon is searched for each individual drone *(B)*.

A. Calculation of the partial polygons

Different attempts for decomposing the whole polygon into single sub cells were investigated.

A nice and elegant method to divide an area into subareas is the so-called “Delaunay-Triangulation” [6]. Unfortunately it proved to be very difficult to divide a polygon in such a way that the resulting partial polygons correspond to a certain percentage of the whole area.

In addition to the basic triangulation the single polygon needed to be checked for their neighborhood relations in order to compose them accordingly again. The originating branches would hardly correspond to the targeted area size so that additional procedures would have to be used.

As an alternative the possibility to divide a polygon by using an approximation procedure and surface balance calculation to get partial polygons was investigated. With this variation the polygon is disassembled first into two incomparably large parts by using predefined angles through the surface balance point. Besides it is irrelevant whether the balance point lies outside or within the body. According to the desired size the algorithm can select the bigger or the smaller partial polygon as a source area for any further decomposition. Afterwards the calculated partial polygon is divided again by the surface balance point. This process continues recursive until the requested area size is reached. On this occasion an approximation procedure could be used to calculate a solution as quickly as possible. However, this segmentation method only works with convex polygons. For concave polygons it is necessary to prevent that the area is divided into more than only two parts.

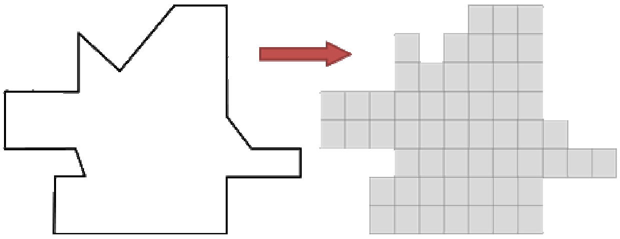


Figure 3. Scanning procedure

A quicker and mathematically less complicated variation to split a polygon is the scanning procedure (see Figure 3). The method is equal to what is called rendering or scan conversion in 2D computer graphics and converts the polygon into a grid of cells. That implies that a higher resolution (i.e. a smaller cell size) will result in a more accurate match of the grid with the originally defined area. To be able to divide the grid afterwards the number of required cells is calculated from the desired area size. With this information and by using a suitable growth algorithm, which extends from any start cell within the grid as long as enough cells have melted, one single continuous area of the desired size can be calculated. This technique resembles the flood-fill algorithm [7] also known from computer graphics. Likewise in this case it is very urgent to know the neighborhood relationship of the cells. Nevertheless, this is quite simple because in contrast to the same problems with the triangulation each of these cells is commensurate and is therefore easy to assign to the co-ordinate system.

To receive a very simple and steady grid polygon, different growth algorithms were compared to each other. A straight growing algorithm turned out to be the most efficient because the results showed more straight edges than other algorithms. In direct conclusion this means a significant reduction of the required rotary and turn maneuvers of the UAV, which leads to a better cost-value ratio in case of using UAVs with a limited turning rate. The generated grid polygons are recalculated into partial polygons just to disassemble them once more into a grid. This time the grid size corresponds to the calculated dimension of the footprint, which depends on the camera specification (focal length and picture sensor) in combination with the desired flight altitude of the drone.

B. *Calculating the flightpath*

To receive an efficient and economically reasonable flight route it is important to find the shortest path that includes all way points and that in addition contains the smallest possible number in turn maneuvers.

The best flight path solution can only be calculated by using a highly complex algorithm and even than an optimal result cannot be achieved in limited time (see the problem of the travelling salesman [8]).

To get acceptable results under the constraint to keep the expenditure as low as possible, different variations were checked mutually.

Because as mentioned earlier a very steady flight route with as few as possible direction changes offers big economic advantages, a method was developed, which processes the polygon according to its expansion in columns or line-by-line similar to a type writer. The so calculated flight route shows a clearer construction in particular with bigger areas.

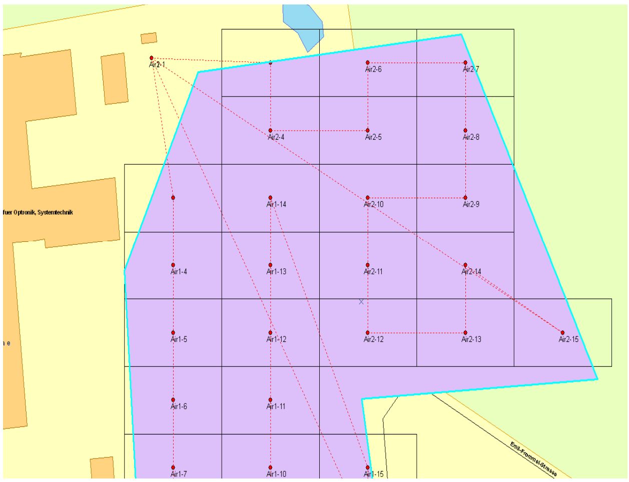


Figure 4. Calculated flight paths for two UAVs

Afterwards the calculated flight route is complemented with safe approach and departure air corridors to avoid collisions between the team members.

VI. Post Processing

The data accumulated by the drones are post processed after the flight to generate an overall result from the single images. Two steps of post processing are done: transferring the images *(A)* and the mosaiking and geo-referencing *(B).*

A. *Transferring the images*

To receive high-quality pictures, high-resolution cameras (10-15 megapixels) are used as payloads for the UAVs. In order to transmit the originating images to the ground station a transport medium must be used that has enough data capacity available. The most elegant method to transfer the images is to use the downlink of the drone. This assumes that the UAV provides an interface, which can be used to feed the data into the downlink of the system. If such an interface is not available other procedures have to be found. During the development of the photo flight, different technologies were tested and evaluated.

To keep the system as simple as possible, the best solution would be to select a communication device that has a great acceptance and is widely used. Therefore the first drafts where done by Wi-Fi. To build such an additional communication line between the UAVs and the ground station a small secure digital memory card was used. This SD card fits perfectly well into the payloads and is able to establish a Wi-Fi connection and to transmit the captured images automatically. The problem with this solution is that in most cases the frequencies for the digital video downlink of the drones is in the 2.4 GHz band which is also used by Wi-Fi for broadcasting. For this reason it can be assumed that at least the Wi-Fi transmission will be disturbed heavily. The best solution for this problem is to move either the digital video downlink or the Wi-Fi to the 5 GHz band. Unfortunately in the current system stage the video downlink is fixed and the used Wi-Fi SD card is not capable of using the 5 GHz band.

For now, the images have to be transferred manually to the ground station.

B. *Mosaiking and geo-referencing*

To benefit from the advantages of the photo flight in full extent, the images taken must be merged to an overall situation picture according to the demand and can be brought into the correct geographical position.

For this purpose different tools were compared to each other. The application of different freeware products showed completely good results but the integration into the overall system proved to be difficult. The software ABUL, (Automated image exploitation at the example of the UAV LUNA), which was also developed at Fraunhofer IOSB would be a possible candidate. However, it is on account of its application aim a too mighty software, which would need, in addition, immense hardware capacities.

Due to the fact that the geo information system used in the photo flight application is based on the software by ESRI [9], which also provides different methods for mosaiking and geo referencing, this variation was also examined. The first results did not show the same quality as previous experiments but they are promising.

The main problem, which leads to disturbances in the final mosaic is, on this occasion, the divergence or inaccuracy of the calculated footprints of the sensor payload. These calculations are based on the GPS position of the drone and the aperture angle of the optic or the size of the used sensor.

To reach more exact values a way must be found to bring the calculated positions of the UAVs in consistency with their real positions. Because these problems are not trivially solvable, another approach could be more promising. Thus it is examined at the moment to what extent the calculation of the corner co-ordinates of the single images can be improved by taking into account certain a priori knowledge and the use of various filters.

VII. Conclusions

The described algorithms were implemented as a software library and are integrated into a geographic information system based on ESRI software specially provided for test purposes. The photo flight tool is an independent software module whereas the logic behind it is interchangeable and thus can be used in other software modules like the situation representation module of AMFIS mentioned above. The results of the algorithm and its ability to adapt to new flight systems with other flight characteristics are currently evaluated. The software was integrated into a three-dimensional simulation tool and the first real test attempts with homogeneous and also with small heterogeneous swarms have taken place.

This research project resulted in a complex prototype system, which is able to form a fully autonomous swarm of UAVs on the basis of several drones and a standard PC or mobile computer at almost any place in very short time that allows acquiring a highly up-to-date aerial image. The sustained data can also make it possible to understand complex blind scenarios quicker. It permits a more exact planning and simplifies the contact with the situation. The deployment of a swarm with a theoretical unlimited number of UAVs means thereby a huge advancement in the field of local just-in-time reconnaissance.

VIII. Extensions and Further Work

In parallel to the work on the photo flight algorithms a small gas sensor, which can also be carried as a payload by an UAV was developed in cooperation with an industrial partner (see Figure 5).

The gas sensor is designed as a very light and compact payload and has been built as a prototype. It can be equipped with up to five different gas sensors and contains, in addition, a sensor to detect universal inflammable gases and a photoionisation detection sensor. Future versions will also be able to detect temperature and humidity. The selection of the five gas sensors can be changed to fit different applications at any time. A supplementation or a further development of the photo flight, in which at least one UAV is equipped with a gas sensor, is planned. Because the aim of this application differs from the original task - visual reconnaissance - above all the geometry of the flight routes must be adapted. This can be assumed from the fact that either the propagation of the gases or the concentration at certain places is of interest. That means that a meandering flight path over a relatively small area makes no sense.

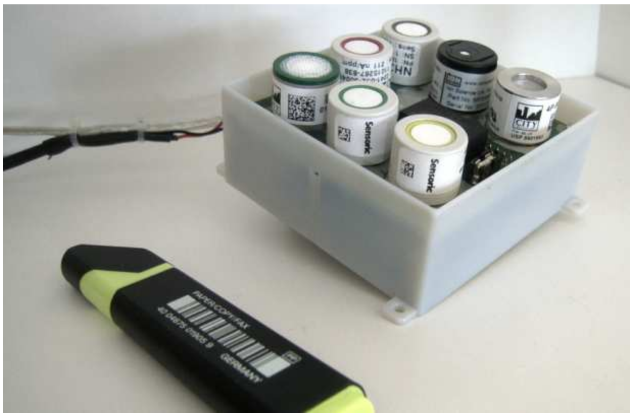


Figure 5. Gas sensor to detect inflammable gases, Ammonia, Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Chlorine

To recognize the propagation of gases certain a priori knowledge like origin, wind force and direction is necessary. With the help of these data a propagation model can be provided as a basis for the calculation of optimum flight routes to validate the estimated results.

Acknowledgment

The authors would like to thank their colleagues and students, who have contributed to the work presented in this paper, especially Sven Müller, Steffen Burger and Thorsten Ochsenreither.

References

[1] S. Leuchter, T. Partmann, L. Berger, E. J. Blum, and R. Schönbein, “Karlsruhe Generic Agile Ground Station,“ In: J. Beyerer (ed.), Future Security. 2nd Security Research Conference 2007, 12th - 14th September 2007, Karlsruhe, Germany. Fraunhofer Defense and Security Alliance (pp. 159-162). Karlsruhe, Universitätsverlag.

[2] Universität der Bundeswehr München, Germany, http://www.unibw.de/lrt13_2/Forschung/Projekte/UAVPF, 2010

[3] K. Daniel, B. Dusza, A. Lewandowski, and C. Wietfeld, “AirShield: A System-of-Systems MUAV Remote Sensing Architecture for Disaster Response”, IEEE International Systems Conference (SysCon), Vancouver, 2009

[4] K. Daniel, B. Dusza, and C. Wietfeld, “Mesh Network for CBRNE Reconnaissance with MUAV Swarms”, 4th Conference on Safety and Security Systems in Europe, Potsdam, 2009

[5] A. Bürkle, F. Segor, and M. Kollmann, “Towards Autonomous Micro UAV Swarms,” In: Proceeding of the International Symposium on Unmanned Aerial Vehicles, Dubai, UAE, 2010.

[6] B. N. Delaunay, “Sur la sphere vide,” In: Bulletin of Academy of Sciences of the USSR 7, Nr. 6, S 793-800, 1934

[7] D. Hearn and M. P. Baker, “Computer Graphics, C version, 2nd Ed,” Prentice Hall, 1997.

[8] D. L. Applegate, R. E. Bixby, V. Chvátal, and W. J. Cook, “The Traveling Salesman Problem. A Computational Study,” Princeton University Press, Februar 2007.

[9] Esri Enterprise, USA, http://www.esri.com, 2010.

Multifactor Authentication Device

Jaroslav Kadlec, Radimir Vrba, Radek Kuchta
Faculty of Electrical Engineering and Communication Brno University of Technology
Brno, Czech Republic
kadlecja | vrbar | kuchtar@feec.vutbr.cz

Abstract— **This paper describes a Multifactor authentication device, the main features of the device and implementation to the Microsoft Windows Credential Provider. The newest, more robust and more secure solutions for user's or system's authentication brings new challenges and requests to design completely new, advanced and highly trustworthy authentication devices. Nowadays one or two factor authentication can be in some cases too risky and vulnerable for security attacks. Therefore more authentication factors are required to increase authentication process security and minimize possible identity forge. Due to this reason a new authentication device, extended authentication process by another two factors, was developed and proposed in this paper.**

Keywords- Multifactor authentication; authentication device; credential provider**.**

I. Introduction

Authentication and authorization are asked almost everywhere in the today's world. People must be identified when they download emails, read newspaper over the Internet, fill out forms for the government, access company private information, etc. When servers communicate to each other, they have to create trusted connection. Before the connection is created, it is necessary to identify servers. There are different ways how to identify a user and a server. For the user authentication some private credentials are usually required. In many cases users are using their unique identification number or username, and password. If one of these values is wrong, a new enter is required. If more than selected number of attempts has been done, user's account is locked.

When a user or a server needs to authenticate a server, the most common way is using of certificates. In this scenario trusted authority issues a certificate that is used for asymmetric cryptography.

Especially scenario with user credentials is sometime insufficient and some extra information is required for many situations and systems. The information should be a user certificate, a user biometric identification or current user position.

This paper describes a new authentication device that allows determining of user position by GPS or wireless communication network, using PIN and fingerprint to authenticate a user and using user's certificate to sing these credentials before sending to an authentication server.

The paper also describes a new possibility of using position data as one of the authentication information. When an information system has information about authenticated person's position, it can change access rights or show only part of accessible data according to predefined access rules.

The first part of the paper describes the main aspects of current user position information using. Next section is a description of basic cryptographic methods used in user authentication process. Following section describes a new Multifactor Authentication Device (MAD) and the main parts of this device. Next sections describe Microsoft Windows Credential Provider and authentication process with connected Multifactor Authentication Device.

II. The Main Aspects of Position Information

Nowadays, many papers discuss using of user's location as a new factor of authentication process. Location-based authentication can be useful in many cases. The advantages of location-based authentication are presented in [1, 2]. One of the possible places for location-based authentication usage can be hospital sector. A doctor shouldn't handle with patients' privacy information out of hospital's border. Another example of location-based authentication we can find in the financial branch. If a user (account owner) would like to operate on his account, it should prove his location at first. If the user is at home or in a bank office, he will get access. If he is out of acceptable locations, he won't get the access to his bank account.

When the user's position is coming into the authentication process some aspects have to be taken into account.

The user's position is very sensitive information that can be abused in many cases. User's position can be also exploited for the position-targeted spam. For these reasons it should be operated very carefully with position information over whole its lifecycle. The way how to achieve user's privacy protection is presented in [3]. Position information should be anonymous as much as possible. The level of anonymity is dependent on required accuracy of position information. For instance, if the service requires position information for country determination, the position information shouldn't be interpreted in accuracy with a few meters.

The second aspect of position-based authentication is user's mobility. In the model situation, a user is authenticated upon its actual position (time t_0). From now the user has granted access, but the user is moving and the position condition can be disturbed (time t_1). The situation is illustrated in Fig. 1.

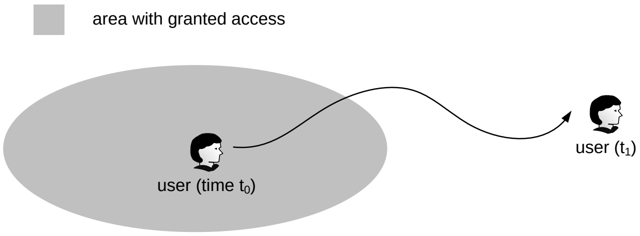


Fig. 1. Position condition collision

The way how to solve this problem is to re-authenticate the user's position periodically. This way is unusable for huge systems for large network resources requirements. Another way is depicted in [4]. Speed and direction of movements are used as additional information.

III. Basics of cryptographical background

The cryptography methods are very helpful and irreplaceable tools in authentication techniques. In Fig. 2 are listed, the main principles of cryptography systems that can be used in authentication techniques.

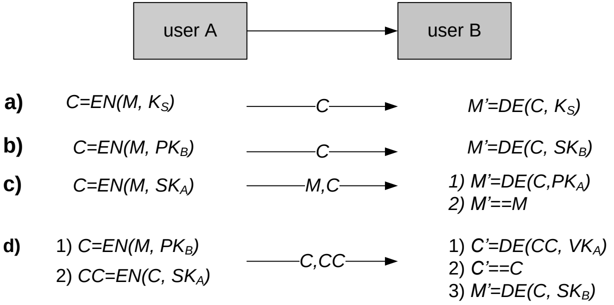


Fig. 2. Basic cryptographycal principles

In Fig. 2 a) there is the principle of symmetrical cryptography system. Message *M* is encrypted by function *EN* and key K_S to cryptogram *C*. After reception *C* is decrypted by decryption function *DE* and same key K_S. The group of symmetrical cryptosystem represents AES (Advantage Encryption System) [5].

In asymmetrical cryptography systems two keys belong to everyone's entity. The first one is public key, known to each entity of the system, the other key is private and its *id* is known only for the appropriate entity. A message can be encrypted by any key (public or private), but for decryption the other key has to be used. In other words, a message encrypted by public key has to be decrypted by private key, and vice versa. Asymmetrical cryptosystem performs two basic functions. Confidentiality of message is performed when the message is encrypted on the sender's side by the recipient's public key PK_B and decrypted by private key SK_B on the recipient's side, Fig. 2 b). If the message is encrypted by sender's private key SK_A, the other side can prove sender's identity when decrypted by public key of the sender PK_A see Fig. 2 c). So, when we encrypt by private key we perform authenticity of the message, respectively encrypting by public key means confidentiality of the message. When we need authenticated confidential message we have to do every operation two times, encryption and decryption, as shown in Fig. 2 d).

The above described principles are used in the proposed techniques for mutual authentication between authentication terminal and AAA server (authenticator). Especially, the principle of Fig. 2 c) is used in newly designed techniques. To prove the origin of a message (authenticity) it has to consist of two parts. The first part is a plain text of a message, the other part is a cryptogram created by encrypting with sender's secret key. On the recipient's side there is a cryptogram decrypted by sender's public key.

Mutual authentication preserves system form third side's attack as depicted in the Fig. 3.

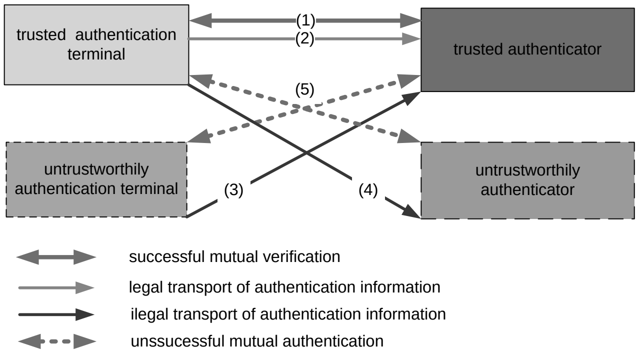

trusted authentication terminal
trusted authenticator
untrustworthily authentication terminal
untrustworthily authenticator
(1)
(2)
(3)
(4)
(5)
successful mutual verification
legal transport of authentication information
ilegal transport of authentication information
unssucessful mutual authentication

Fig. 3. Possible attacks to the authentication systems

Messages should be sent just between trusted pair (paths 1 and 2 from Fig. 3) (authentication terminal and authenticator – part of the server AAA).

Third side's device can emulate each device in the system, but we strongly rely on elimination of other possibilities as are depicted in Fig. 3.

IV. Wireless networks and their possibilities for position determination

When position of a subject in the space is needed to know three basic conditions have to be fulfilled. First, the space where the subject is found should be described. The most often coordinates are used to describe of the space [6]. Secondly, enough of anchor points with known position have to be had. And finally, distance between the subject and anchor points have to be found. Number of used anchor points depends on dimension of the space [7]. Example for two dimensional systems is depicted in Fig. 4.

General equation used for position determination is

$$(x-m)^2+(y-n)^2=r^2, \tag{1}$$

where *m* and *n* are coordinates of the center of a circle (position of an anchor point), *x* and *y* are coordinates of points on the circle (possible position of the subject) and *r* is radius of the circle (distance between the anchor point and the subject). Then we can get equation system for getting position of our subject [8]

$$
\begin{aligned}
&(x-6)^2+(y-14)^2=5{,}83^2\\
&(x-7)^2+(y-7)^2=5{,}65^2\\
&\underline{(x-17)^2+(y-11)^2=6^2}\\
&x=11;\, y=11
\end{aligned}
\qquad (2)
$$

When previous equation system is solved, coordinates of position of the subject will be gotten (for our example 11,11).

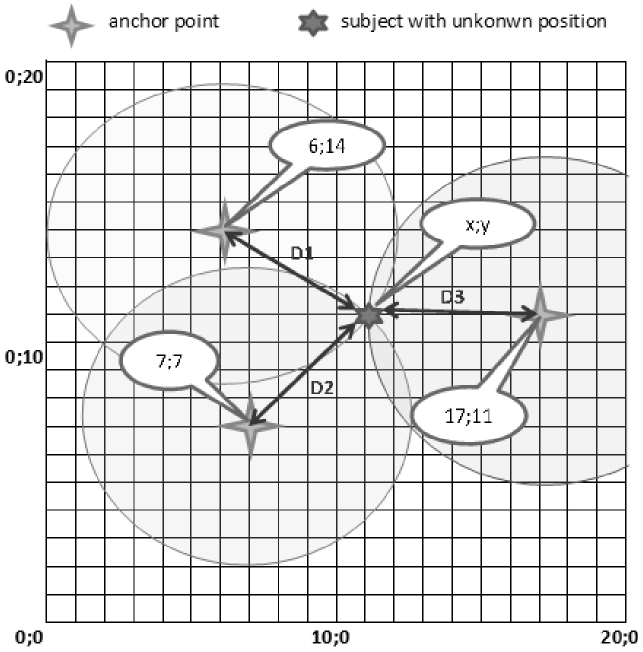


Fig. 4. Position determination with three wireless connection access points

When we use wireless communication network to determine current position, we use the same equation. In many cases it is not important to know exact position. If we would like to know that user is in corporate building, it is enough that he is in range of corporate wireless network or specific wireless transmitter.

V. Multifactor authentication device

For described methods of position determination some user device is needed.

Block diagram of the first generation of the user authentication device connected to the user terminal is shown in Fig. 5. User is using user terminal to connect to the authentication server in this scenario. This terminal is interconnected with the user authentication device via USB data bus. The core of the multifactor authentication device (MAD) is low powered central processor unit. The device contains secured data repository (SDR), where user's credentials are stored.

SDR is also a place, where user's certificate is stored. Each data, that are send to Authentication server are signed out by the user certificate.

SDR also stores trusted server certificates, or in the other scenario, the certificate of a trusted certification authority that issued server certificate. Whole device sends credentials only to the server, with valid and trusted certificate.

The third certificate stored in SDR is a device certificate. This certificate has to be valid and issued by certification authority that is trusted by authentication server.

Server certificate and device certificate are used to authenticate server and device together and to create secured communication channel.

Before data stored in SDR are accessed, the user has to open access to it by fingerprint login through fingerprint biometrics to PIN authentication.

The authentication device is connected to the user terminal with appropriate software that allows interconnection between the user authentication device and the authentication server, over the USB.

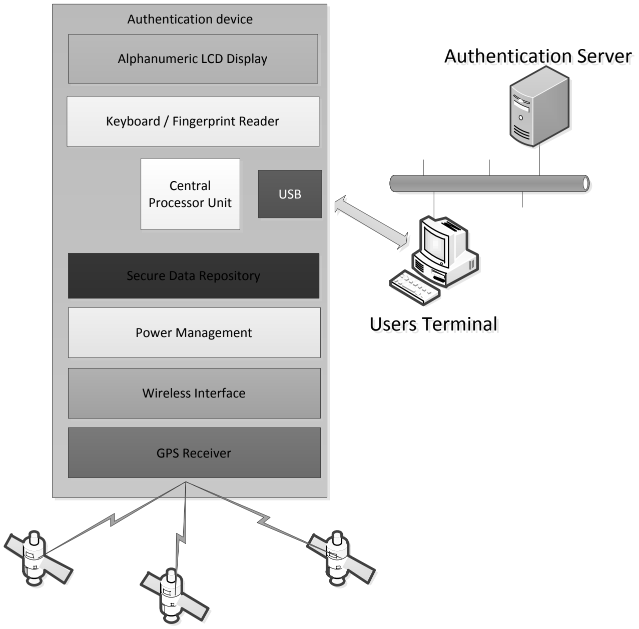


Fig. 5. Block structure of the user authentication device

If user is using standard computer without installed software or public computer, the authentication device also contains alphanumeric display where authentication instructions are shown. Thus, web access can be used for the user's authentication.

Current position of the device is determined by integrated GPS receiver. Current position is periodically stored to the internal memory with time stamp. When authentication process requires the device position and device is out of GPS signal, stored position with time stamp is used.

Different wireless interfaces are available on MAD to determine current position with company wireless network. There two main implementation. The first one is using WiFi and connection to wireless access points. The second one is using proprietary wireless communication platform IQRF. Principe of position determination is always the same how is described above.

VI. Windows Logon Implementation

Authentication protocols are implemented in Windows by security service providers. Windows Vista introduces a new authentication package called the Credential Security Service Provider, or CredSSP, that provides a single sign-on (SSO) user experience when starting a new Terminal Services session. CredSSP enables applications to delegate users' credentials from the client computer (by using the client-side security service provider) to the target server (through the server-side security service provider) based on client policies [9].

Credential providers [9] are in-process COM objects that are used to collect credentials in Windows Vista and run in local system context. In summary, the logon UI provides interactive UI rendering, Winlogon provides interactive logon infrastructure, and credential providers help gather and process credentials.

After all providers have enumerated their tiles, the logon UI displays them to a user. The user interacts with a tile to supply his or her credentials. The logon UI submits these credentials for authentication. Combined with supporting hardware, credential providers can extend the Microsoft Windows operating system to enable users to logon through biometric (fingerprint, retinal, or voice recognition), password, PIN, smart card certificate, or any custom authentication package a third-party developer wants to create.

Credential providers are not enforcement mechanisms. They are used to gather and serialize credentials. The LSA and authentication packages enforce security [12, 13, 14, 15].

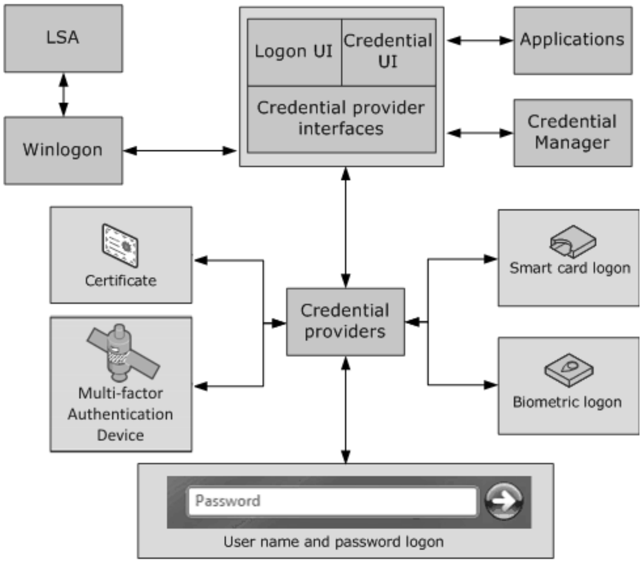


Fig. 6. Windows Vista hybrid credential provider architecture with integrated MAD CredSPP [10]

Credential providers are registered on a Windows Vista computer and are responsible for:

- Describing the credential information required for authentication.
- Handling communication and logic with external authentication authorities.
- Packaging credentials for interactive and network logon.

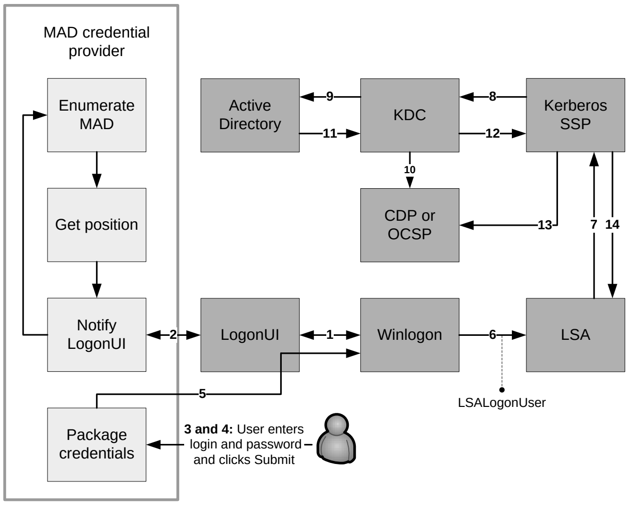


Fig. 7. Modified logon flow [10] for MAD authentication process

The hybrid credential provider architecture is shown in Fig. 6. The hybrid credential provider API does not design UI (User Interface) but describes which controls need to be rendered to windows logon screen. The hybrid credential provider interfaces with the Windows Smart Card API or Biometric API both directly and indirectly. The direct interface is via public routines, which allow the detection of connected biometrics or smartcard devices or even detection of inserted card. The indirect interface is via the custom APIs specific for each connected devices, which allow the credential provider to read a user credential directly from the device. The MAD credential provider uses own MAD API for low-level communication. Obtained user's credential from MAD through MAD API are combined with password from logon UI and sent to credential provider interface.

VII. Authentication Process

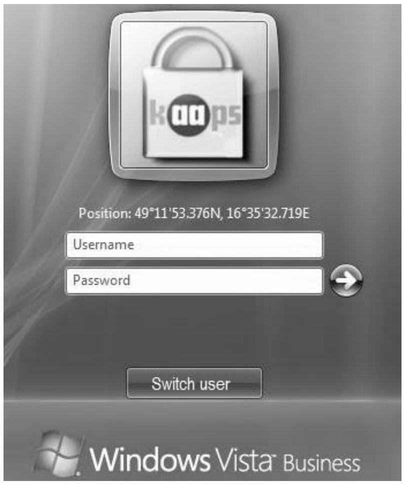
Authentication process with connected MAD is a combination of standard Windows logon process with custom scripts executed by the Active Directory (AD) user's policies [11]. Flow sequence of logon process within Multifactor Authentication Device (see Fig. 7):

1. Winlogon requests the logon UI credential information. Asynchronously, our multifactor authentication resource manager starts. The multifactor authentication credential provider:
   a. Gets a list of multifactor authentication devices (uses our MAD API).
   b. Get position information from connected multifactor authentication devices, the MAD credential provider copies it into a temporary secure cache on the terminal.
   c. Notifies the logon UI that new credentials exist.

2. The logon UI requests the new credentials from the MAD credential provider. As a response, the MAD credential provider provides to the logon UI actual position information. The user selects a multifactor authentication device logon title, and Windows displays a logon dialog box.
3. The user enters his login and password and clicks Go button.
4. The credential provider that resides in the LogonUI process (system) collects login, password and position. As part of packaging credentials in the MAD credential provider, the data is packaged in a KERB_INTERACTIVE_LOGON structure. The main contents of the KERB_INTERACTIVE_LOGON structure are User Name, Domain Name and Password.
5. The credential provider now wraps the data (such as encrypted PIN, container name, reader name, and position information) and sent them back to LogonUI.
6. Data from Logon UI are now presented by Winlogon for LSALogonUser.
7. LSA calls Kerberos Authentication Package (Kerberos SSP) to create a Kerberos Authentication Service Request (KRB_AS_REQ) containing a pre-authenticator [12].
8. The Kerberos SSP sends an authentication request [12] to the Key Distribution Center (KDC) service that runs on a domain controller, to request a Ticket Granting Ticket (TGT).
9. The KDC finds the user's account object in the active directory and uses the user's credentials to verify the user identity.
10. The KDC validates the user's key to ensure that the credential information come from a trusted source.
11. The KDC service retrieves user account information from Active Directory. The KDC constructs a TGT based on the user account information that it retrieves from Active Directory. The TGT includes the user's security identifier (SID), the SIDs for universal and global domain groups to, which the user belongs, and (in a multi-domain environment) the SIDs for any universal groups of, which the user is a member. The TGT's authorization data fields include the list of SIDs.
12. The domain controller returns the TGT to the client as part of the KRB_AS_REP response.
13. The response is as per RFC 4556 [12].
14. The client validates the reply from the KDC (time, path and revocation status).
15. Now that a TGT has been obtained, the client obtains a Service Ticket to the local computer in order to log on to the computer.
16. On success, LSA stores the tickets and returns success to the LSALogonUser. On this success message, user profile, last logon time and position information are obtained.
17. Custom login script for multifactor authentication device is called from AD login policies. The MAD custom script serves as an intelligent decision algorithm, which compares current position with last logon position and last logon time with current time on AD authentication server from Kerberos authentication packet. Using authentication server time prevents changing time cheating. Based on these comparisons user access is allowed or denied.
    a. In case of successful authorization logon process continues normally according to user's policies. Last login time and position in AD is actualized to current values.
    b. If user access is denied Winlogon returns to original state and waits for another user logon attempts.

Preconditions for successful login into AD are customized user's properties in AD extended by login position and time information. These values are validated against position and time of MAD used for user authorization.

Logon UI for the thin client with implemented multifactor authentication is shown in Fig. 8. The thin client doesn't obtain user's credentials from MAD but allows only weakest authorization by three factors. User is challenged for his username and password. These credentials are expanded by the fixed position information of the thin client and AD authorization authority runs modified authorization process, which was described before. Difference of thin and thick client Logon UI implementation is the thick client offers only password input box for entering password. All other necessary information is read from connected MAD (position, username obtained by the biometric validation).



Position: 49°11'53.376N, 16°35'32.719E

Fig. 8. Microsoft Windows Vista logon screens with integrated support of Multifactor Authentication Device (MAD connected-obtained position information, dialog used for user login)

VIII. CONCLUSION AND FUTURE WORK

Common multifactor authentication processes combine in most cases two or three unique authentication factors. Typical scenario is biometric reader connected or inbuilt to personal computer and user is verified by biometrics, and by password knowledge. Similar example is using personal tokens with certificates, when user has to proof knowledge of PIN which secures personal certificate stored in token.

Our newly developed solution combines new multifactor authentication device with currently used technologies as an authentication system based on the Microsoft Credential provider in combination with corporate network with Active Directory services. Main advantages of presented solution are added position information as an extra authentication factor, increased security level due to the position restrictions and rules for allowing or denning access from predefined areas, compatibility with widely used Microsoft technologies and systems and possibility to implement into current corporate networks based on the Active Directory services.

Our future work will be focused on the finishing Multifactor Authentication Device prototype realization and testing in connection with Microsoft Credential provider itself and in combination with Active Directory services under real corporate network.

ACKNOWLEDGMENT

This research has been supported by the Czech Ministry of Education, Youth and Sports in the frame of MSM 0021630503 MIKROSYN New Trends in Microelectronic Systems and Nanotechnologies Research Project, partly supported by ARTEMIS JU in Project No. 100205 Process Oriented Electronic Control Units for Electric Vehicles Developed on a multi-system real-time embedded platform and by ENIAC JU in Project No. 120001 Nanoelectronics for an Energy Efficient Electrical Car, partly by the Czech Ministry of Industry and Trade in projects FR-TI1/057Automatic stocktaking system and FR-TI1/057 Intelligent house-open platform.

REFERENCES

[1] Ray, I., and Kumar, M., *Towards a location-based mandatory access control model*, Computers & Security, vol. 25, pp. 36-44, Feb 2006.

[2] Denning, D. E., and MacDoran, P. F., *Location-based authentication: Grounding cyberspace for better security*, Computer Fraud & Security, vol. 1996, pp. 12-16, 1996.

[3] Schilit, B., et al., *Wireless location privacy protection*, Computer, vol. 36, pp. 135-137, Dec 2003.

[4] Tikamdas, P. S., and El Nahas, A., *Direction-based proximity detection algorithm for location-based services*, in Wireless and Optical Communications Networks, 2009. WOCN '09. IFIP International Conference on, 2009, pp. 1-5.

[5] Menezes A., et al., *Handbook of Applied Cryptography*, CRC Press, 1997.

[6] Cutler, T. J., *Dutton's Nautical Navigation*, 2003. 664 pages. ISBN 155750248X.

[7] Monahan, K., and Douglass, D., *GPS Instant Navigation: A Practical Guide from Basics to Advanced Techniques*. 2nd edition. Fine Edge Productions, 2000. 333 pages. ISBN 0938665766.

[8] Larson, R., *Geometry*. Houghton Mifflin Harcourt, 2006. 1003 pages. ISBN 0618595406.

[9] Kiaer, M., *Multifactor authentication in Windows - Part 2: Preparing Devices on XP and Windows 2003*. WindowSecurity.com. [Online] 12. 2 2008. [Cited: 17. 6 2009.] http://www.windowsecurity.com/articles/Multifactor-authentication-Windows-Part1.html.

[10] Mysore, S. H. *Windows Vista Smart Card Infrastructure. Microsoft Download Center*. [Online] 16. 8 2007. [Cited: 17. 6 2009.] http://www.microsoft.com/downloads/details.aspx?familyid=AC201438-3317-44D3-9638-07625FE397B9&displaylang=en.

[11] Griffin, D., *Create Custom Login Experiences With Credential Providers For Windows Vista*. MSDN Magazine. [Online] 1 2007. [Cited: 5. 6 2009.] http://msdn.microsoft.com/en-us/magazine/cc163489.aspx.

[12] Zhu, L., and Tung, B., *Public Key Cryptography for Initial Authentication in Kerberos (PKINIT)*. RFC4556. http://www.ietf.org/rfc/rfc4556.txt: Microsoft, June 2006.

[13] Microsoft. *How the Kerberos Version 5 Authentication Protocol Works*. Microsoft TechNet. [Online] May 2008. [Cited: 17. 6 2009.] http://technet.microsoft.com/en-us/library/cc772815.aspx.

[14] Harrison, E. R., *Lightweight Directory Access Protocol (LDAP): Authentication Methods and Security Mechanisms*. RFC:4513. http://www.rfc-editor.org/rfc/rfc4513.txt: Novell, Inc., 2006.

[15] Melnikov, A., and Zeilenga, K., *Simple Authentication and Security Layer (SASL)*. RFC4422. http://www.ietf.org/rfc/rfc4422.txt: OpenLDAP Foundation, 2006.

Protect Critical Information Infrastructure Systems in Financial and Healthcare Sectors: Actor Network Theory

Cheng-Chieh Huang
Dept. of Information Management
National Taiwan University
Taipei, Taiwan
d94725007@ntu.edu.tw

Ching-Cha Hsieh
Dept. of Information Management
National Taiwan University
Taipei, Taiwan
cchsieh@im.ntu.edu.tw

***Abstract*—CIIP is not only instrumental means but also social relevance. Most of studies on CIIP neglect considering social groups that CIs or CIIs serve and social characteristics of technology. In this article, it proposes actor network theory to analyze interdependence of CIIs and CIIP policy. Finally, it indicates social dimension in interdependence analysis, government's broker role and social-technical service system approach in critical information infrastructure protection studies.**

Keywords-critical information infrastructure protection; actor network theory; service systems

I. INTRODUCTION

Critical infrastructure protection (CIP) is currently seen as an essential part of national security in numerous countries around the world and a broad range of political and ad-ministrative initiatives and efforts is underway in the US, in Europe, and in other parts of the world in an attempt to better secure critical infrastructures (CIs).

The CI delivers a range of services that individuals, and society as a whole, depend on. That is, any damage to or interruption of the CI could cause ripples across the technical and the societal systems. Moreover, the critical information infrastructure (CII) underpins many elements of the critical infrastructure, as many information and communication technologies (ICT) have become all-embracing, connect other infrastructure systems, and make them more interrelated and interdependent[1]. Critical information infrastructure protection (CIIP) forms new issues for policy research and technology studies.

Past literature on CIP or CIIP divided into two angles, one from technology or system angle to discuss interdependence of technology, infrastructure or cyber security protection [2][3]. Another discusses about public-private-partners or government role from policy research [4] [5].

These studies neglect that current digital economic society is an interwoven socio-technical seamless web, consisting of heterogeneous, changing formations of actor networks, meanings, work practices and institutional and organizational arrangements [6]. They forget social groups that these critical infrastructures or technology serve and technology finally become a part of the relevant practical, symbolic and cognitive spaces of the actors involved [9]. Thus, it seems to consider social, technology or infrastructure simultaneously in CIIP.

In this paper, we propose a social-technical perspective, Actor Network Theory (ANT) to understand the complex social, technology interwoven phenomenon of critical information infrastructures and their protection. Using ANT, we can get a deeper understanding of the interrelationships of heterogeneous actor groups and of the mediating roles played by humans and technologies, and the critical information infrastructures. Further implications to interdependence of critical information infrastructure, government involvement, and public-private-partnerships issues are also discussed.

In the following section, we first review the literature of CIIP. Second, we review ANT Theory. Third, the financial and healthcare sectors of our cases are illustrated. Fourth, we present analysis and discussion and fifth, we identify contributions, limitations and suggestions for future research.

II. CRITICAL INFORMATION INFRASTRUCTURE PROTECTION

CIs are part of a larger set of services and products that are considered essential to the functioning of our modern economies and societies. These include but are not limited to energy, information technology, telecommunications, healthcare, transportation, water, government and law enforcement, and banking and finance.

Critical Infrastructure Assurance Office (CIAO), an interagency office created under Presidential Decision Directive 63 to assist in coordinating the federal government's initiatives on critical infrastructure protection in US [3]. The CIAO defined infrastructure as:

The framework of interdependent networks and systems comprising identifiable industries, institutions (including people and procedures), and distribution capabilities that provide a reliable flow of products and services essential to the defense and economic security of the United States, the smooth functioning of governments at all levels, and society as a whole.

Most of the CI relies on a spectrum of software-based control systems for smooth, reliable, and continuous operation. In many cases, information and communication technologies (ICT) have become all-embracing, connecting other infrastructure systems and making them interrelated and interdependent. These ICTs underpin many elements of the critical infrastructure, called Critical Information Infrastructures (CIIs)[1].

Complex and interdependence of CIs or CIIs are difficult to manage, even protect. Most of CIP or CIIP studies discuss systems of system or interlinks to protect. They define physical interdependency, cyber interdependency, geographic independency, logical independency to analyze or protect [2][3].

These studies provide good analysis tools or methods to protect technology or infrastructure systems. But they forget the final objective of CIP or CIIP to protect people or society that CIs or CIIs serve not just technology or infrastructure themselves. And these services that technology or infrastructure provide are different meanings for different users in their different sectors or stages of usage.

Moreover, there are different institutionally fragmented environments in different sectors of different countries such as healthcare, transportation, water and they serve different user groups and interests groups[1][4]. If we cannot understand meanings of users or interests groups of critical infrastructure in different sectors it serves, it is difficult for government representatives to persuade CEOs or CIOs of critical infrastructure companies to invest.

That is, it is more meaningful to consider society, technology and their interwoven simultaneously in the CIP or CIIP studies.

III. Actor Network Theory

The social-technical approach encapsulates a wide range of perspectives and concepts. They attempt to explain the relationship and interactions between technology and society.

Actor Network Theory (ANT) is one of the most important social-technical perspectives in recently years. It was developed in the sociology of science and technology school [10]. ANT helps describe how actors form alliances and involve other actors and use non-human actors (technology) to strengthen such alliances and to secure their interests. ANT consists of two concepts: translation and inscription.

When an actor-network is created, consists of four processes of translation [11]:

- Problematization: The focal actors define interests that others may share, establishes itself as indispensable resources in the solution of the problems they have defined. They define the problems and solutions and also establish roles and identities for other actors in the network. As a consequence, focal actors establish an “obligatory passage point” for problem solution which all the actors in an actor-network must pass.
- Interessement: The focal actors convince other actors that the interests defined by the focal actors are in fact well in line with their own interests. Through interessement the developing network creates sufficient incitement to both lock actors into networks.
- Enrollment: Enrollment involves a definition of roles of each of the actors in the newly created actor-network. It also involves a set of strategies through which focal actors seek to convince other actors to embrace the underlying ideas of the growing actor-network and to be an active part of the whole project.
- Mobilization: The focal actors use a set of methods to ensure that the other actors act according to their agreement and would not betray. With allies mobilized, an actor network achieves stability.

In addition to the four stages of translation, the process of inscription is critical to building networks, as most artifacts within a social system embody inscriptions of some interests. As ideas are inscribed in technology and as these technologies diffuse in contexts where they are assigned relevance, they help achieve socio-technical stability.

IV. Case Study

A. Case Background

This case is a three years research project to understand the current critical information protection status of every sector in Taiwan. In every year, the research project team will generate CIIP strategies and policy implications reports for government.

In 2009, the first year of research project, the project team decided financial, healthcare sectors as first priority to examine. The project team adopted “sector roundtables methodology” [7] and table-top exercise to understand protection status and dependence or interdependence between sectors.

In every sector analysis, the project team introduced four steps: 1. select CIIs, 2. analyze threats, weakness and interdependence, 3. design exercises, 4. execute table-top exercises and get evaluations from experts.

Followings are the project experiences and reflection in financial and healthcare sectors in first year.

B. Financial Sector

In the financial sector, there are more than forty large banks and institutions in Taiwan. Most banks and institutions are privacy, but they are supervised by Financial Supervisory Commission of Taiwan government (FSCEY). Moreover, the most critical institutions, such as stock exchange institution, futures exchange institution, depository or cleaning institution are government funded.

There two subsystems or mechanism that serve different users. One is the stock exchange system that supports the investors to exchange stocks or futures in the stocks or futures markets. The institutions and information systems must make sure the fair trade and the price is sensitivity to any crisis or news. The actors or technology system should response quickly to any crisis.

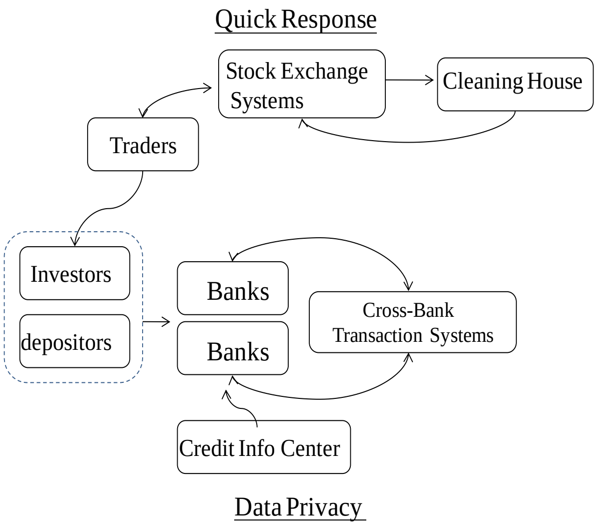


Figure 1. Subsystems in Financial Sectors

TABLE I. COMPARISONS OF ENERGENCE AND PRIVACY ISSUES IN FINANCIAL SECTORS

Institutions	Emergence Issues	Privacy Issues
Banks	Low	High
Clearning Hourse	Medium	Medium
Stock Exchange Institutions	High	Low
Credit Info Center	Low	High

Another subsystem is the banking system. The banks provide their customers to deposit or loan. Thus, compared to stock exchange system, data privacy is more critical for the banking system than quick response. Even the cross-bank transaction system is shutdown; the customers still can withdraw money by branch of banks.

Thus, the two subsystems represent their different protection angles. The stock exchange system focus on when and what sequences they should stop the exchange markets and announce to people or inventors. The banking system protects their data and prevents data leak from their inner control mechanisms.

No matter banks, stock exchange instructions, they very depend on information systems. They are also very actively response to any events that will impact confidence of people or their investors.

For financial sector, inscriptions of their CIIs are "confidence" and their cultures are responsively to serve their customers.

Thus, they expect the electrical power or telecommunication could recover quickly and response the exactly recover time to satisfy their customers' expectation.

C. Healthcare Sector

TABLE II. COMPARISONS OF ENERGENCE AND PRIVACY ISSUES IN HEALTHCARE SECTORS

Institutions	Emergence Issues	Privacy Issues
Hospitals/Clincals	Low	High
Heathcare insurance institution	Low	Medium

Although in the healthcare sector, such as hospitals or clinics are also the service institutions to serve their customers. But they are less IT resources to operate because of most resources are invested in clinical instruments. A Chief Information Officer (CIO) of a hospital said, "Our IT problem is less IT resources. We do not have enough IT investments and also we do not have qualified IT people to join".

"One day, our data center flooded because of a typhoon. We wait for one week to get the electrical power engines to recover our electronic power and data center!" the CIO updated. Also, if some accidents that let MIS people cannot work, such as H1N1 infection could be a problem to run IT operation.

The hospital information systems (HIS) are critical information systems, but it will bring just inconvenient, such as register or submit prescription slowly while IT breaking down. Also, clinical information is very important; it is the data protection issue.

Other important institutions in health sector are health insurance institutions. There is only one health insurance institution in Taiwan and governed by government. While the healthcare insurance system is shutdown, the patients still can see a doctor, the hospitals or clinics can record their insurance numbers and issue later when the system is recovered. The health insurance institution holds partial clinical information in their database; it has some data privacy protection issue.

For healthcare sector, inscriptions of their CIIs are "convenient". And they spend little money on IT investments.

V. ANALYSIS AND DISCUSSIONS

It summarizes different inscriptions, problems, interests and requirements for dependent sectors in financial and healthcare sectors in table III. We can understand that different institutions concern their interests in every sector. Moreover, the technical elements, CIIs also represent their different meanings, such as 'confidence' or 'convenient. Before protection policy carried out, we should understand relevant social groups in sectors, their meanings and social inscriptions of technology.

Following sections, we discuss new insights and implications of CIIP from the ANT theory.

TABLE III. ANT ANALYSIS IN FINANCIAL AND HEALTHCARE SECTORS

ANT Analysis/ Sectors	**Financial-Stock Exchange Subsystem**	**Financial-Banking Subsystem**	**Healthcare**
Inscription	confidence	confidence	convenient
Translation (problems and interests)	reover market confidence quickly	data protection and confidence	smooth healthcare process, less IT resources

A. *Social Relevance, then Protection*

Most CIP or CIIP literature focuses on analyzing different interdependences of CIIs in different sectors. But they neglect how they collaborate under their different interdependence or dependences relationships.

For example, in our project experiences, the stock exchange subsystem in financial sector very depend on electrical power, and request that every shortage power events, the electrical power company should response planned recovery time quickly.

For institutions in financial stock exchange subsystem, transparent and confidence information they provide to their investors are important in financial markets. But for the electrical power company, they concern not only recovery quickly, but electrical power stable and quality.

Usually, the CIIs are dependent or interdependent, but their social concerns are inconsistent.

Thus, how to enroll the "confidence" CIIs of financial sectors and "stable" CIIs of electrical power sectors is not only from system functional views but from the social-technical angle.

B. *Redefine Government's Role in CIIP*

It is difficult for policy makers to enroll the private sectors to join the critical information infrastructure protection actor-network. Especially, private sectors realize the governments want they invest in security and reliability beyond their normal business continuity requirements [4]. Moreover, what is the government's role in CIP or CIIP?

From ANT analysis, the different sectors have their different competition environments, relevant social groups, social-technical configurations, and meanings. It is not suitable for policy makers to persuade or enroll these actors from only "national security" reasons or strategies.

ANT theory argues that focal actors establish an "obligatory passage point" for problem solution which all the actors in an actor-network must pass. It means governments as the focal actors, should consider obligatory passage points for sectors or institutions to enroll.

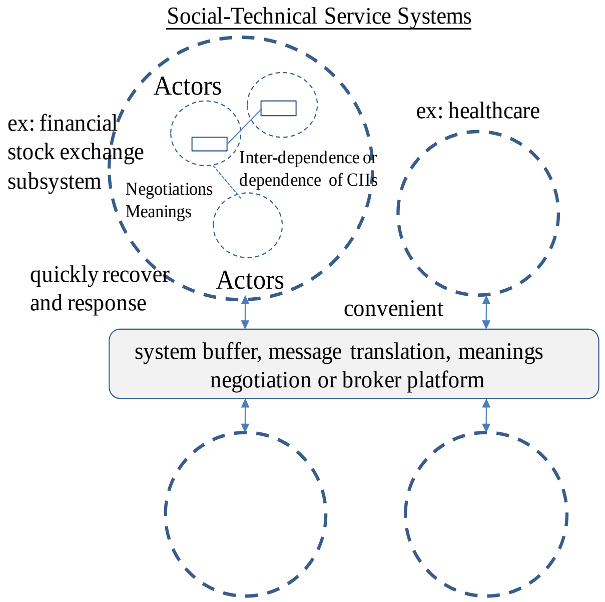

Social-Technical Service Systems
Actors
ex: financial stock exchange subsystem
Negotiations Meanings
Inter-dependence or dependence of CIIs
ex: healthcare
quickly recover and response
Actors
convenient
system buffer, message translation, meanings negotiation or broker platform

Figure 2. The Social-Technical Service System Approach to CIIP

For example, "confidence" is the financial sector's consideration to their IT investment. The policy makers should persuade or assist they to strength these mechanisms or technologies. Few resources are the problems to healthcare sectors, how to allocate different resources while crisis happened is solution to enroll healthcare sectors.

Moreover, as our cases showed, these sectors depend on other sectors' technologies or infrastructures, but pursue themselves interests. Government could provide the broker role or build up boundary infrastructures, to help different sectors to collaborate or negotiate.

Thus, we call this kinds of relationships should be private-public-private partnerships.

C. *Methodology of CIIP*

Developing a comprehensive architecture or framework for interdependency modeling and simulation is very challenge. Moreover, it is difficult to resolve different sectors' problems using a single methodology or analysis tool.

No matter what kinds of tools or methods, it should consider more on social-political dimensions. Moreover, these methods should not try to sacrifice sectors' or actors' interests, but how to strength or broker their interests or meanings.

In summary, we propose the "Social-Technical Services Systems" view to guide the methods design (see figure 2). It provides the service system view to understand every sector or subsector service meanings, requirements. Also, governments or institutions can provide the broker roles or information system platforms to help negotiate different meanings, translate requirements to help every sectors achieve their obligatory passage points.

VI. Conclusion

In this article, we propose the actor network theory to understand complex social, technology interwoven phenomenon of critical information infrastructures and their protection. Through our financial and healthcare sectors experiences and social-technical analysis, it implicates that include social dimension in independence analysis. Moreover, governments should also consider sectors' interests, and play a broker role to balance or negotiate their different interests. Finally, we propose social-technical service system approach to protect critical information infrastructure.

Future research, we will take more in-depth analysis of different sectors about their meaning, negotiations, interactions, interests, and inscriptions. Finally, we will design more comprehensive methods for CIIP from a social-technical service system approach.

References

[1] Brunner, E.M. and M. Suter, International CIIP Handbook 2008 / 2009, ETH, 2008.

[2] E. Bagheri and A.A. Ghorbani, "The State of the Art in Critical Infrastructure Protection: a Framework for Convergence", International Journal of Critical Infrastructure, 2007, pp. 1-36.

[3] S. M. Rinaldi, J.P. Peerenboom, and T.K. Kelly, "Complex Networks: Indentifying, Understanding, and Analyzing Critical Infrastructure Interdependencies", IEEE Control Systems Magazine, 2001, pp. 11-25.

[4] M.B. Bruijne and M. Eeten, "Systems that Should Have Failed: Critical Infrastructure Protection in an Institutionally Fragmented Environment", Journal of Contingencies and Crisis Management, 2007, pp. 18-29.

[5] M. Dunn, "The Socio-Political Dimensions of Critical Information Infrastructure Protection (CIIP)", International Journal of Critical Infrastructure, 2005, pp. 258-268.

[6] R. Kling and R. Lamb, "IT and Organizational Change in Digital Economies: A Socio-Technical Approach", Computer and Society, 1999, pp. 17-24.

[7] Dunn, M. and V. Mauer, International CIIP Handbook 2006, ETH, 2006.

[8] T. F. Pinch and W. E. Bijker, "The Social Construction of Facts and Artifacts: Or How the Sociology of Science and Technology Might Benefit Each Other", Social Studies of Science, 1984, pp. 399-411.

[9] K. H. Sorensen and R. Williams, Shaping Technology, Guiding Policy, MA: Edward Elgar Publishing, 2002.

[10] M. Callon and B. Latour, "Unscrewing the big Leviathan", in Advances in Social Theory and Methodology, K. Knorr-Cetina, and A.V. Cicourel, Eds, London: Routledge & Kegan, 1981, pp. 277–303.

[11] M. Callon, "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St. Brieuc Bay", in Power, Action and Belief, J. Law, Eds, London: Routledge & Kegan, 1986, pp. 197–233.

Mobile RFID Mutual Authentication and Ownership Transfer

Ming Hour Yang
Information Computer Science
Chung Yuan Christian University
mhyang@cycu.edu.tw

Jia-Ning Luo
Information and Telecommunication
Ming Chuan University
deer@mail.mcu.edu.tw

Abstract — **In this paper, we propose an ownership transfer scheme that applies in mobile RFID networks. The scheme includes a mutual authentication protocol and a role-based ownership transfer protocol. A tag will decide what actions are allowed for a reader according to the reader's role class, and the back-end server will send to the reader the requested information about the tag. Keyed-hash functions are used to secure the protocols. Last, we prove that our protocol can do against the threats of replay attacks, distributed denial of service (DDoS), Man-in-the-Middle (MITM) attacks that change users' data, interception of data and location privacy, and tracking of tags' ownership transfer.**

Keywords-RFID;authentication; ownership transfer

I. INTRODUCTION

RFID features mass identification, large data size, modifiable identification and data, and effective scanning of tags by batch processing at long distance. Nowadays, mobile RFID [1][11] integrating reading chips, passive RFID tags and mobile phones enables users to access information. Mobile RFID can be applied in business transaction; through the transfer of tagged products' ownership, each transaction can be done with mobile RFID. The transfer of a tagged product's ownership suggests whoever is registered in the tag is the one entitled to the item.

To protect the privacy of both the former and current owners of a tagged item, RFID protocol designers have to make sure that when the item's ownership changes, its tag's ownership has to change accordingly and simultaneously. Former owners, therefore, will no longer be able to access the tag, whereas the current owners have no way to track the privacy history that was kept in the tag, either.

Due to the limitation of tags, there are only 2000 logic gates in a passive tag to do security functions [4][7]. In 2006, John Ayoade[5] proposed an authentication-control framework, creating a table on back-end authentication server (AS) to control the reader-tag authentication. When a reader accesses a tag, the tag will send out its identifier and encrypted messages to the reader, and the reader sends a reading request to the AS. The AS checks the reader's identity and gives a key to the reader to decrypt the message and grant the ownership of the tag.

The authorization and ownership transfer process, the delegation, should be done securely to protect the owner and the tags [3][8][9][10]. If the delegation process is incomplete, the former owner could still access the tag [2]. Fouladgar proposed a delegation protocol to deal with incomplete ownership transfer [8][9][10]. In the protocol, the delegated reader can verify the digital certificate of a current owner's reader through a certificate authority (CA) during the ownership transfer process, and the key stored in the tag is updated by the AS to ensure only the current owner can access the tag.

Although delegated readers reduce the computation load of the AS, the reader's computation resources such as CPU and memory are limited. When a reader has too many delegated tags, it can no longer afford the authentication task because it does not have enough memory to keep tags' information. Fouladgar's protocol uses counters to limit delegated readers seemed to fail to take good control of reading limits.

When malicious users sent a large number of queries to the tags, the tags will keep asking AS to update the keys. If the update message was lost or abandoned by attackers, Foudladgar's protocol will fail and the owner's reader will lost the tag. To prevent this kind of DoS attack, Osaka[6] proposed another ownership transfer scheme. In Osaka's scheme, the tag confirms the ownership transfer is completed with AS in every session. However, in Osaka's scheme, a reader should have large memory to keep the tags' keys, and it is suffer from man in the middle attack.

In this paper, we propose a protocol for ownership transfer and reader-tag mutual authentication in a mobile RFID environment. Unlike traditional RFID, mobile readers are usually put under the presupposition that they might be malicious devices and their communication with back-end server is not secured. In our protocol, the ownership of a tag is transferred to the new reader by updating the tag's key after a mutual authentication process between the read and the tag. Our protocol can not only reduce tags' computational load effectively but also allow readers to access tags without storing any shared keys. Furthermore, our protocol provides location privacy, data privacy and forward security. Our protocol can prevent replay attacks, man in the middle attack, the DoS attack, and protects the tag location and the ownership transfer history.

This paper is organized as follows: in the next section, we proposed a Mobile Access Control and

ISBN:978-1-61208-114-4

Ownership Transfer protocol (MACOT) to deal with mutual authentication and ownership transfer in mobile RFID environment. Section 3 deals with the security analysis of our protocol. Section 4 analyzes the protocol's performance. Conclusion is drawn in the Section 5.

II. Mobile Access Control and Ownership Transfer

In this section, we propose a Mobile Access Control and Ownership Transfer (MACOT) protocol to deal with mutual authentication and ownership transfer in mobile RFID. Our ownership transfer scheme consists of three stages. The first one is mobile mutual authentication procedure (MMAP), the second ownership transfer procedure (OTP), and the third RC-Action Table update procedure.

The mutual authentication protocol requires the readers obtain the corresponding information of a tag from back-end authentication server according to reader's authority. The ownership transfer protocol updates the tag's key with the authorized owner after the mutual authentication process. And the RC-Action table update procedure is used by a current owner to grant control of a tag.

A. Preliminary

With the high mobility, a mobile reader has wide-range accessibility. Subsequently, tags within its access range could probably belong to a different authority. An authentication scheme is required to identify tags and locates their corresponding back-end servers. According to H. Lee and J. Kim's mobile RFID infrastructure [11], the authentication process with 7 steps is shown in Figure 1. :

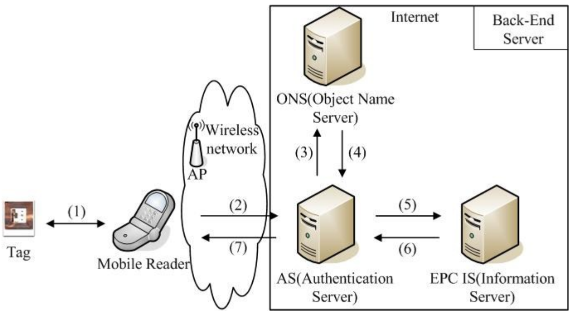


Figure 1. Mobile RFID infrastructure

Step 1. A mobile reader sends a reading request to the tag, and gets a responding message from it.

Step 2. The reader forwards the message to AS to verify the identity of the tag.

Step 3. AS verifies tag's identity and queries Object Name Server (ONS) to get the detailed information of the tag.

Step 4. ONS sends the tag's URL of EPC IS to AS.

Step 5. According to the URL, AS requests the tag's information from EPC IS, which is the back-end database of tags.

Step 6. EPC IS sends the tag's information to AS.

Step 7. AS sends the tag's information to the reader.

Because a passive tag's computation resources is limited, the packets uses a keyed-hash function $h_x()$, generated with the key x shared by the tag and the authentication server to prevent eavesdropping. The traffic between the readers and the AS is protected by traditional symmetric encryption algorithm $E_k()$. Back-end server, including AS, ONS and EPC IS, are trusted by tags and readers.

To manage a reader's authority over a specific tag, AS and the tag must store the authorizing information. Figure 2(a) stored the corresponding actions of readers of different role classes (RCs) to the tag *TID* 80 in a RC-Action table. In the table, the tag owner has the highest privileges to modify the actions of each RC. The role in an upper row has higher privileges. As shown in Figure 2(a), readers with an owner-level RC are entitled to Action 3, which means they are also authorized to do Action 1 and Action 2. In addition, the relevant information of tags is stored in different EPC ISs, as in Figure 2(b):

(1) Readers' Access Control List: each row indicates each reader's RC class. For example, the reader *RID* 312's authority over *TID* 80 is B-class RC.

(2) Action Table: the AS decides what command could be send from the reader to the tag. For example, the reader with *RID* 312 can access *TID* 80's public (general) data and private (personal) data.

RC-Action Table
for TID 80

RC	Action
A	1
B	2
C	2
Owner	3

(a) Information Table in Tag

Reader's Access Control List

TID / RID	79	80	730
214	A	A	
312		B	Owner
666	A	C	A

Action Table

TID	Action	Command	Data
79	1	Read_pubilc	
	2	Ownership transfer	
80	1	Read_pubilc	
	2	Read_private	
	3	Ownership transfer	
730	1	Read_pubilc	
	2	Read_private	
	3	Ownership transfer	

(b) Information Table in Back-End Server

Figure 2. Data stored in (a) tag (b) back-end server

We assume all readers are not trusted and they do not need to store any tags' keys. In the initialization stage, the keys and secret of back-end server, readers and tags are shown in Figure 3. The server stored each tag's identifier *TID*, two shared keys K_x and K_y between tags and the server, a *PIN* shared with a tag and the owner's reader, and a shared secret *C*.

The tag's owner, the reader, which owns a tag, stored the *TID* of the tag, and it's *PIN* and *C* values. In each tag's memory, stored the K_x, K_y, *PIN* and *C*.

Tag's Information Table

TID	K_x	K_y	PIN_i	C
79	99	96	94	90
80	11	22	33	44
730	55	66	77	88

(a) Keys and Secret Values Stored in the Server

Tag Owner of TID 730

PIN	77
C	88

(b) Keys and Secret Values Stored in the Reader

TID	730
K_x	55
K_y	66
PIN	77
C	88

(c) Keys and Secret Values Stored in the Tag

Figure 3. Shared keys and secret values stored in (a) Server (b) Reader (c) Tag

B. Mobile Mutual Authentication Procedure (MMAP)

We assume the back-end server can verity the reader's identity and exchange a session key K_{dr} between them. When a reader read a tag, the Mobile Mutual Authentication Procedure (MMAP) is preforms with 8 steps:

Step 1. When a reader queries a tag, the tag generates a random number r_1 and creates a secret value *S* by XOR-ing r_1 and its own identifier *TID*, and computes $S = h_{K_x}(TID \oplus r_1)$. The tag sends *S* and r_1 to the reader.

Step 2. The reader generates another random number r_2, and sends $E_{k_{dr}}(S, r_1, r_2, RID, Command)$ to the server.

Step 3. The server decrypts the message with K_{dr} and computes $h_{K_x}(TID \oplus r_1)$ for all tags to obtain *TID*.

Step 4. The server looks up Reader's Access Control List to find out the reader's RC, access level, and generate a random number r_3. It computes $T = h_{K_y}(TID \oplus r_1 \oplus r_2 \oplus r_3, RC, Command)$ and $p = h_{K_y}(r_3 \oplus TID)$. The server encrypts *T*, *p*, r_2 and r_3 with a session key K_{dr}, i.e. $E_{K_{dr}}(T, p, r_2, r_3)$, and sends the result to the reader.

Step 5. The reader decrypts the message with the session key K_{dr} and verifies r_2. If it's correct, the reader forwards *T*, r_2 and r_3 to the tag.

Step 6. The tag verifies T by searching all the possible values of RC and commands. It computes $p = h_{K_y}(r_3 \oplus TID)$, $G = h_{K_x}(TID \oplus r_3 \oplus Act)$ and sends them to the reader.

Step 7. The reader verifies *p*, and forwards G and r3 to the server by computing $E_{K_{dr}}(G, r_3)$.

Step 8. The server verifies *G* to find a matched *Act*, and searches the action table in Figure 2. to find a matched *Command*. If the *Command* matches the *Act*, the reader is authorized, and the server sends the requested tag's data to the reader.

The complete mutual authentication protocol is illustrated in Figure 4:

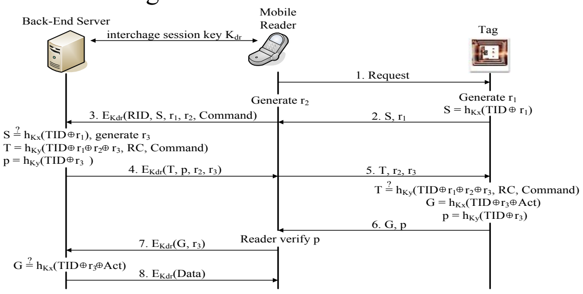

Back-End Server
Mobile Reader
Tag
interchage session key K_{dr}
1. Request
Generate r_2
Generate r_1
$S = h_{Kx}(TID \oplus r_1)$
3. $E_{Kdr}(RID, S, r_1, r_2, Command)$
2. S, r_1
$S \overset{?}{=} h_{Kx}(TID \oplus r_1)$, generate r_3
$T = h_{Ky}(TID \oplus r_1 \oplus r_2 \oplus r_3, RC, Command)$
$p = h_{Ky}(TID \oplus r_3)$
4. $E_{Kdr}(T, p, r_2, r_3)$
5. T, r_2, r_3
$T \overset{?}{=} h_{Ky}(TID \oplus r_1 \oplus r_2 \oplus r_3, RC, Command)$
$G = h_{Kx}(TID \oplus r_3 \oplus Act)$
$p = h_{Ky}(TID \oplus r_3)$
6. G, p
Reader verify p
7. $E_{Kdr}(G, r_3)$
$G \overset{?}{=} h_{Kx}(TID \oplus r_3 \oplus Act)$
8. $E_{Kdr}(Data)$

Figure 4. Mobile Mutual Authentication Procedure

C. Ownership Transfer Procedure (OTP)

After the server, the reader and the tag authenticate themselves to each other, the Ownership Transfer Procedure (OTP) is performed to transfer the ownership between the former owner and the current owner, as shown in Figure 5. The two owners should authenticate each other through a trust third party before perform the OTP.

The OTP is divided into two parts:

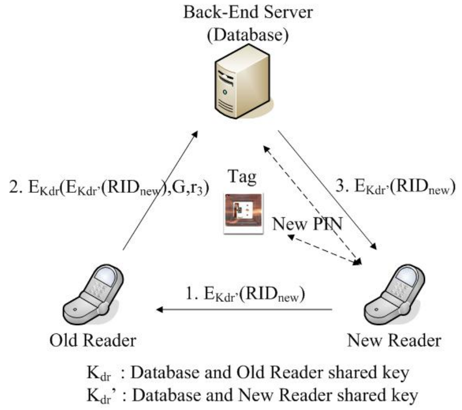

Back-End Server (Database)
2. $E_{Kdr}(E_{Kdr'}(RID_{new}), G, r_3)$
Tag
3. $E_{Kdr'}(RID_{new})$
New PIN
1. $E_{Kdr'}(RID_{new})$
Old Reader
New Reader
K_{dr} : Database and Old Reader shared key
K_{dr}' : Database and New Reader shared key

Figure 5. Diagram of Ownership Transfer

1) Part 1

In Part 1, the server authenticates the former and current owners and tag. The two owners have to exchange session keys K_{dr} and K_{dr}' with the server. Next, the current owner encrypts his identifier RID_{new} with K_{dr}' and send it to the former one. The former owner uses the MMAP protocol to authenticate him with the tag, as shown in messages 2-7 of Figure 6. . The former owner encrypts $E_{Kdr'}(RID_{new})$, r_3 and G to back-end server in message 8. The server adds the current owner RID_{new} into the Access Control List of the current owner's reader into the tag and marks RID_{new}'s RC as Owner.

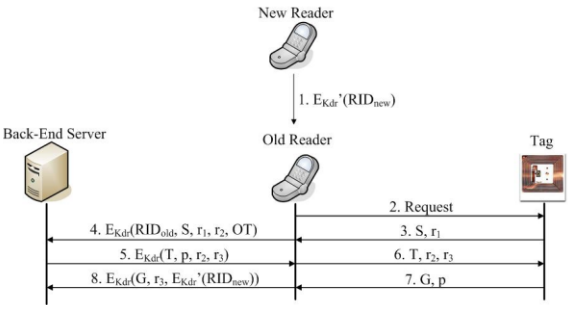


Figure 6. The First Part of OTP

2) Part 2

The second part of the OTP is shown in Figure 7. After receiving $E_{K_{dr}'}(RID_{new})$ from the server, the current owner verifies the derived RID_{new}. The current owner uses the MMAP protocol to authenticate him with the tag, as shown in messages 2-6 in Figure 7. The rest of the protocol (steps 7-13) is as follows:

Step 7. The tag generates $G' = h_{K_x}(TID \oplus r_3 \oplus Act, PIN_i)$ and sends it to the server via the reader.

Step 8. The reader encrypted G' with the random number r_3', and sends it to the server.

Step 9. The server verifies G' and updates the tag's PIN and secret value C. the server computes $h_{PIN_i}(r_1', C)$, and sends it to the reader

Step 10. The reader forward the message to the tag.

Step 11. The tag verifies r_1' and C, and generates a new $PIN_{i+1} = h_{K_x}(PIN_i \oplus K_y, r_3')$ and $C' = h_{K_x}(C \oplus K_y, r_3')$. The tag computes $h_{PIN_{i+1}}(r_3', C')$ with the new PIN and C, and sends it to the reader.

Step 12. The reader forwards the message to the server. The server uses the same function to generate the new PIN and C and verifies $h_{PIN_{i+1}}(r_3', C')$. If the comparison is the same, the server modifies the reader's Access Control List to change or delete the former owner's tag identifier and reader's RC.

Step 13. The server sends the new PIN and C to the current owner.

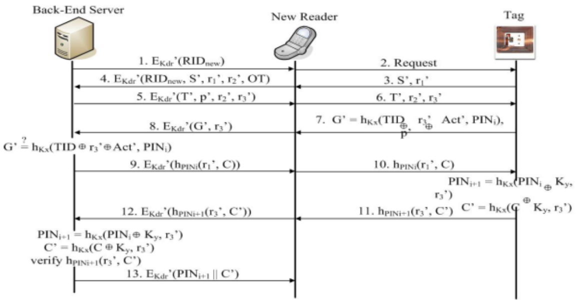

Back-End Server
New Reader
Tag
1. E_Kdr'(RID_new)
2. Request
4. E_Kdr'(RID_new, S', r_1', r_2', OT)
3. S', r_1'
5. E_Kdr'(T', p', r_2', r_3')
6. T', r_2', r_3'
7. G' = h_Kx(TID ⊕ r_3' ⊕ Act', PIN_i), p'
8. E_Kdr'(G', r_3')
G' ?= h_Kx(TID ⊕ r_3' ⊕ Act', PIN_i)
9. E_Kdr'(h_PINi(r_1', C))
10. h_PINi(r_1', C)
PIN_i+1 = h_Kx(PIN_i ⊕ K_y, r_3')
C' = h_Kx(C ⊕ K_y, r_3')
12. E_Kdr'(h_PINi+1(r_3', C'))
11. h_PINi+1(r_3', C')
PIN_i+1 = h_Kx(PIN_i ⊕ K_y, r_3')
C' = h_Kx(C ⊕ K_y, r_3')
verify h_PINi+1(r_3', C')
13. E_Kdr'(PIN_i+1 || C')

Figure 7. The Second Part of OTP

The missing of message 10 and 11 could lead to asynchronous update of data between back-end server and the tag. Our protocol is designed to tackle such asynchrony in OTP and requires that the reader re-access the tag after it sends G' to the server. Meanwhile, the server uses PIN_i for computation to generate G' and check if this is the same as the G' from the reader. If they are different, the server will compute again with other secret values to generate PIN_{i+1} and re-queries G'. If the two G's are the same, it means PIN_{i+1} is the key of the tag and it has updated its key. Therefore, the server no longer needs to update the tag, and will send PIN_{i+1} and C' to the current owner directly. If the G' that the server generates with PIN_i is identical to the one from the reader, the tag has missed message 10 in the communication and has not yet updated PIN_i. Consequently, back-end server begins to generate $h_{PINi}(r_1', C)$ and update the tag's key with it. If the tag returns $h_{PINi+1}(r_3', C')$, the update has been completed. After verification, the server will send PIN_{i+1} and C' to the current owner. The procedure is illustrated as below:

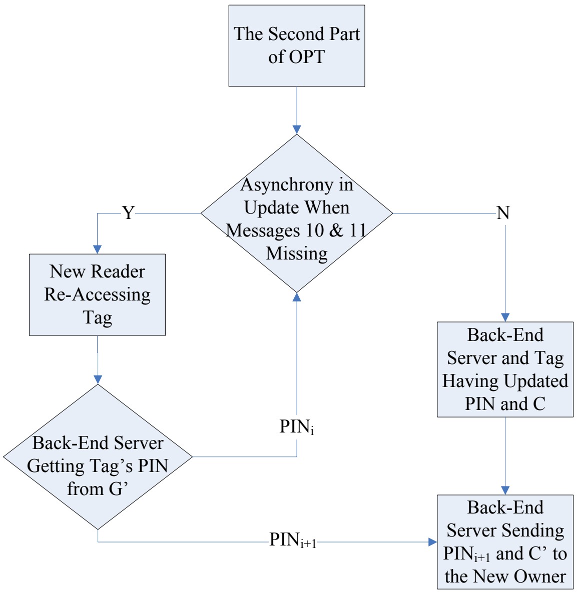

The Second Part of OPT
Asynchrony in Update When Messages 10 & 11 Missing
Y
N
New Reader Re-Accessing Tag
Back-End Server and Tag Having Updated PIN and C
PIN_i
Back-End Server Getting Tag's PIN from G'
PIN_i+1
Back-End Server Sending PIN_i+1 and C' to the New Owner

Figure 8. Diagram of OTP When Messages Missing

In mobile RFID, tag owners can transfer a tag's ownership to others through OTP. The following is an instance to exemplify OTP. Figure 2 outlines the initiation stage, the owner of reader *RID* 312 transferring his ownership over the tag *TID* 730 to a current owner of reader *RID* 666. First, mutual authentication is achieved between back-end server and *RID* 312. Then the server looks up the reader's Access Control List (ACL) and confirms *RID* 312's RC to *TID* 730 is Owner. Next, the server generates *p* and *T* before sending them to *RID* 666 and *TID* 730 respectively. For the server now, *RID* 666's RC to *TID* 730 has been updated as Owner, as the ACL in Figure 9 (b) indicates. Following these steps, the server will begin its mutual authentication with *RID* 666 and accordingly verifies it as the data receiver of *TID* 730. Subsequently, the server encrypts *T* and *p* with a key shared with *RID* 666 and then sends the encrypted *T* and *p* to *RID* 666. Now, the mutual authentication between *TID* 730 and *RID* 666 must be achieved before *TID* 730 generates *G'* with its *PIN* 77 and sends it to the server. Next, the server generates new *PIN* with 77 (see Figure 3 (b)) and a new shared secret *C* with the old one 88 (see Figure 3 (b)), and sends them to *RID* 666 and *TID* 730 respectively. As a result, both the tag and back-end server update *TID* 730's *PIN* and *C* in their own information tables, as illustrated in the highlighted cells of Figure 9. Ownership, therefore, is transferred to the reader *RID* 666.

TID	730
K_x	55
K_y	66
PIN	$h_{55}(77 \oplus 66,15)$
C	$h_{55}(88 \oplus 66,15)$

(a) Table in Tag

Table for Tags' Information

TID	K_x	K_y	PIN_i	C
79	99	96	94	90
80	11	22	33	44
730	55	66	$h_{55}(77 \oplus 66,15)$	$h_{55}(88 \oplus 66,15)$

Reader's Access Control List

RID \ TID	56	80	730
214	A	A	
312		B	A
666	A	C	Owner

(b) Tables in Back-End Server

Figure 9. After Ownership Transfer, Tables in (a) Tag (b) Back-End Server

D. RC-Action Table Update Procedure

As the ownership is transferred, the current owner, with the *PIN* and *C* from back-end server, is able to renew the tag's RC-Action Table to set the allowed actions for other readers. The steps are as follows:

Step 1. The current owner's reader generates a random number r_1, sends it to the tag and begins to update the RC-Action Table. The tag receives r_1 and generates r_2 and then XORs them. Further, it computes a hash with *PIN* and *C* before sending it to the reader.

Step 2. The reader queries the hash function. If it is valid, this message is sent by a legal tag. Next, the reader uses r_2 and *C* to compute a hash with *PIN* $h_{PIN}(r_2,C)$ before sending it to the tag.

Step 3. The tag queries $h_{PIN}(r_2,C)$. If it is not valid, the reader does not belong to the owner. If valid, then the reader does. Next, the tag puts the values of *RC* and its corresponding *Act* into a keyed-hash function for computation one row after another before sending it to the owner. Receiving the message, the owner computes the hash values one after another and therefore is able to restore the RC-Action Table. r_2 here is used to prevent MITM attacks on one hand, and for the owner to query whether the message is sent by the tag on the other.

Step 4. The owner XORs the *RC* and its corresponding *Act* in the RC-Action Table one row after another, then put each of them into a keyed-hash function with r_1 for computation, and finally sends them to the tag. In addition, the tag computes the hash values one by one and updates its RC-Action Table.

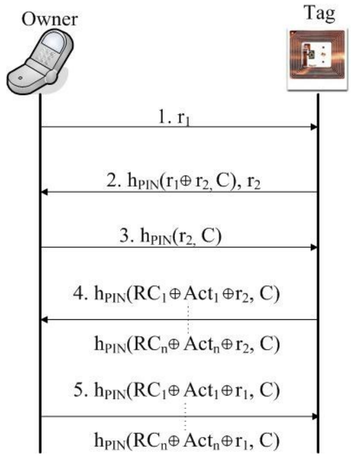

Owner
Tag
1. r_1
2. $h_{PIN}(r_1 \oplus r_{2,} C), r_2$
3. $h_{PIN}(r_{2,} C)$
4. $h_{PIN}(RC_1 \oplus Act_1 \oplus r_2, C)$
$h_{PIN}(RC_n \oplus Act_n \oplus r_2, C)$
5. $h_{PIN}(RC_1 \oplus Act_1 \oplus r_1, C)$
$h_{PIN}(RC_n \oplus Act_n \oplus r_1, C)$

Figure 10. RC-Action Table Update Procedure

Since a tag owner can modify its RC-Action Table at will, we will take the following example to show how an owner updates a tag's RC-Action Table, enabling the RC-A users to access private data. Its initial state is shown in Figure 11 (a) and the tag's identifier *TID* is 730. After the reader-tag mutual authentication, *TID* 730 verifies this reader as Owner. Now, the reader is able to modify *TID* 730's RC-Action Table, updating RC-A's Action from 1 to 2. That is to say, users authorized as RC-A can access not only public data but also private one.

RC-Action Table

RC	Action
A	1
B	2
C	2
Owner	3

(a) Before Update

RC-Action Table

RC	Action
A	2
B	2
C	2
Owner	3

(b) After Update

Figure 11. TID 730's RC-Action Table (a) Before Update (b) After Update

By controlling the RC-Action Table, a tag owner is also able to decide what level of data is accessible to what readers, according to their RCs. Thus, readers with lower authority is not entitled to the data that requires high authority, while readers with higher authority can fully access the tag at will.

III. Security Analysis

In this section, we will prove that our OTP is able to secure ownership transfer against replay attacks, DoS from asynchronous update and MITM attacks that change messages; to achieve mutual authentication; and to protect the privacy of tags' data and location, even though the valid readers have been attacked. Since we have assumed that the communication between a reader and AS is secured, we will just focus on the security of the reader and tag.

A. Against MITM Attacks' Modification of Messages

In our protocol, messages between a reader and tag are protected by keyed-hash functions. For instance, a tag generates *S* and sends it to the back-end server. The server uses *TID* and K_x, shared with the tag, to verify *S*.

B. Against DoS from Asynchronous Update

We use *PIN* to synchronously update the keys and secret values between a tag and back-end server, that is included in G', as sent by the tag in the message 7 in Figure 7, is used for the server to verify a tag's keys. If message 11 is abandoned by malicious users, which could lead to only a tag's update of *PIN* and *C* unilaterally, the server can derive PIN_{i+1} from PIN_i found in this tag's information table and from G' sent by the tag so as to query whether the key PIN_{i+1} is exactly identical to that stored in the tag. This scheme can, therefore, prevent DoS attacks that result from asynchronous update of keys.

C. Against Replay Attacks

As every query between a tag and reader carries a random number in each session, attackers are not able to launch replay attacks by simply coping the last verified message and resending it to back-end server. Our authentication scheme will fail their attempts in this style. For example, if attackers resend to a tag a verified message that contains the value *T* consisting of r_1 generated by the tag, such as the message 6 in Figure 7. , the tag will query *T* with current r_1 in the current session. If the two are different, the authentication procedure will not go further and attackers cannot access any data from the tag, either.

D. Security of the Data Privacy of Tags

We secure the messages between a reader and tag with keyed-hash functions h_{Kx} and h_{Ky}.
If attackers launch replay attacks or try interception, they can only get hashed values sent by a tag, e.g. $h_{Kx}(TID \oplus r_1)$. They cannot obtain a tag's identifier *TID* from those hashed values. Thus, the privacy of tags' data is secured.

E. Security of the Location Privacy of Readers and Tags

Normally, if attackers record a couple of messages between a reader and tag, they can probably find the connection in these messages and accordingly are able to track the location of the reader and tag. In our OTP, a tag sends out three messages, i.e. *S'* in the message 3, *G'* and *p'* in the message 7 and $h_{PINi+1}(r_3',C')$ in the message 11, as illustrated in Figure 7. Because the three messages all contain random numbers, their results change in every session. In doing so, attackers can no longer track a tag's location from these messages and its location privacy is secured. Similarly, the messages 6 and 7 in Figure 7, which are forwarded to back-end server by a reader, also change in every session because of the random numbers that the three messages (3, 7 and 11) carry along. Consequently, attackers cannot find the connection between these messages that the reader forwards and track its location.

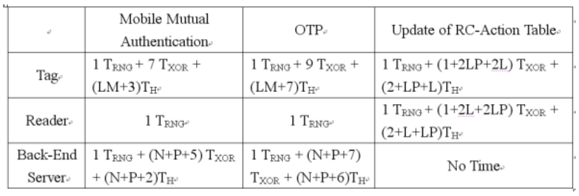
F. Security of Ownership Transfer

To secure ownership transfer, back-end server sends and updates a tag's secret values *PIN* and *C* via the current owner, who then encrypts them with a symmetric key K_x. Because of the keyed-encryption, owners' privacy is protected and the former owner can no longer modify a tag's RC-Action Table with the old *PIN* and *C*. Therefore, with the deprivation of former owners' access authority and the protection from the threats mentioned above, we can say the ownership transfer is secured.

IV. Performance

The performance of our schemes will be analyzed in this section and their results will be illustrated in detail in Table 1. T_H represents the time that a hash function takes in one computation; T_{XOR}, the time that an XOR takes in one computation; T_{RNG}, the time it takes to generate a random number; N, the total tags that back-end server stores; L, the levels of a RC; M, the actions of a tag; P, the actions that a tag is entitled to.

TABLE I. PERFORMANCE OF TAG, READER AND BACK-END SERVER IN EACH SCHEME



	Mobile Mutual Authentication	OTP	Update of RC-Action Table
Tag	1 T_{RNG} + 7 T_{XOR} + (LM+3)T_H	1 T_{RNG} + 9 T_{XOR} + (LM+7)T_H	1 T_{RNG} + (1+2LP+2L) T_{XOR} + (2+LP+L)T_H
Reader	1 T_{RNG}	1 T_{RNG}	1 T_{RNG} + (1+2L+2LP) T_{XOR} + (2+L+LP)T_H
Back-End Server	1 T_{RNG} + (N+P+5) T_{XOR} + (N+P+2)T_H	1 T_{RNG} + (N+P+7) T_{XOR} + (N+P+6)T_H	No Time

Table 1 indicates that the performance of the three items (tag, reader and back-end server) is based on the numbers that users design for L, M and P, whereas a reader does not need to store any keys to access a tag, e.g. in mobile mutual authentication and ownership transfer, except in the update of RC-Action Table.

V. CONCLUSION

In the foreseeable future, RFID readers will not be confined by locations anymore. The combination of reading chips and mobile devices has made mobile readers come true and paved the way for the development of mobile RFID. However, security issues remain a pain for RFID engineers, traditional and mobile alike. As for mobile RFID, the security is even at more serious stake because malicious users might take unauthorized readers to access people's tags and this could endanger the privacy of users and their data. For this reason, we propose a mutual authentication scheme for mobile RFID, using back-end server to verify readers and then find out their RCs. Besides, we require that back-end server send RCs via readers so that a tag can always obtain the current reader's RC before being accessed. Tag owners subsequently look up the information tables stored in tags and decide what actions are allowed for a reader. Eventually, following a tag's final decision, back-end server sends to a reader the requested information of this tag. Apart from these, this scheme is also capable of ownership transfer by updating tags' keys. We use keyed-hash functions in the messages between tags and readers and therefore secure the tag-reader mutual authentication and ownership transfer against replay attacks, DoS from asynchronous update, MITM attacks' modification of messages and malicious users' tracking of tags' location and ownership transfer history, and, last but not least, enhance the privacy of tags' data and location.

ACKNOWLEDGEMENT

This work was supported by the National Science Council (NSC 99-2219-E-033-001 and NSC 99-2221-E-130-007), Republic of China.

REFERENCE

[1] M. H. Yang, "Lightweight authentication protocol for mobile RFID networks," *International Journal of Security and Networks*, vol.5, no.1, pp. 53-62, 2010.

[2] B. Toiruul and K. Lee, "An advanced mutual-authentication algorithm using AES for RFID systems," *International Journal of Computer Science and Network Security,* vol.6, no.9, pp. 156-162, September 2006.

[3] D. Molnar, A. Soppera, and D. Wagner, "A scalable, delegatable pseudonym protocol enabling ownership transfer of RFID tags," *Selected Areas in Cryptography,* pp. 276-290, 2006.

[4] H. Y. Chien, "Secure access control schemes for RFID systems with anonymity," *Mobile Data Management, 2006. MDM 2006. 7th International Conference on,* pp. 96-96, 2006.

[5] J. Ayoade, "Security implications in RFID and authentication processing framework," *Computers & Security,* vol. 25, no.3, pp. 207-212, 2006.

[6] K. Osaka, T. Takagi, K. Yamazaki, and O. Takahashi, "An efficient and secure RFID security method with ownership transfer," *Proceedings of the 2006 International Conference on Computational Intelligence and Security,* vol. 2, pp. 1090-1095, 2006.

[7] S. A. Weis, S. E. Sarma, R. L. Rivest, and D. W. Engels, "Security and privacy aspects of low-cost radio frequency identification systems," *The First International Conference on Security in Pervasive Computing,* pp. 201–212, March 2003, Revised Papers, 2004.

[8] S. Fouladgar, F. Evry, and H. Afifi, "An efficient delegation and transfer of ownership protocol for RFID tags," *Proceedings of the First International EURASIP Workshop on RFID Technology*, September 2007.

[9] S. Fouladgar and H. Afifi, "A simple delegation scheme for RFID systems (SiDeS)," *RFID, 2007. IEEE International Conference on,* pp. 1-6, 2007.

[10] S. Fouladgar and H. Afifi, "A simple privacy protecting scheme enabling delegation and ownership transfer for RFID tags," *Journal of Communications,* vol. 2, no. 6, pp. 6-13, 2007.

[11] N. Park, H. Lee, H. Kim and D. Won "A security and privacy enhanced protection scheme for secure 900MHz UHF RFID reader on mobile phone," *IEEE International Symposium on Consumer Electronics,* pp. 692–696, 2006.

Intrusion Detection System for wide Automation Network Based on the Ethernet Compatible Communication Protocols

Jaroslav Kadlec, Radimir Vrba, Radek Kuchta
Faculty of Electrical Engineering and Communication Brno University of Technology
Brno, Czech Republic
kadlecja | vrbar | kuchtar@feec.vutbr.cz

***Abstract*— This paper is focused on the description of importance, design, and implementation of the Intrusion Detection Systems for a new automation system based on the Ethernet communication protocol. Newly developed and designed automation networks for complex factory control are composed from several types of automation communication links with different communication protocols, but most of the factory middle layer and top layer communication networks are based on Ethernet communication protocol. Wide use of Ethernet communication protocol not only in IT, but also in automation field, brings not only advantages of easy implementation and interoperability between different automation communication networks, but also brings risks and vulnerabilities, well known form IT. Therefore security incidents are becoming more serious and more common not only in computer networks, but also in automation networks. Actual trends in automation networks are among others wide automation networks covering several manufacture divisions or remote controlling of automation networks through the Internet. Necessity of a remote connection to the automation networks covers all security vulnerabilities and risks, which originate from the Internet. Analogically with IT, an automation network can be secured by the conventional way through firewalls and VPN tunnels, but automation networks have several specific requirements on the QoS, against the IT networks. For this reason a new automation firewall device was defined, designed and tested. The new automation firewall includes messaging system for logging all events and alerts originates form automation network. IDMEF (Intrusion Detection Message Exchange Format) is used, as a basis for automation firewall messaging system.**

*Keywords- **Intrusion detection; automation network; IDMEF**.*

I. INTRODUCTION

When a standard security mechanism is taking some actions to prevent the system from a threat, the engineering or a local intrusion detection system might be interested in such information. For this a policy has to be defined, when and how alerts and logging messages are processed.

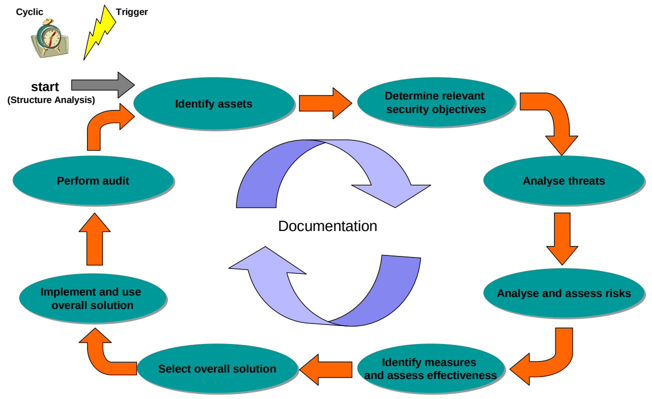


Fig. 1: Basic Security analysis procedure model [3].

The handling of messages for alerts or logging information should be in line with standardized mechanisms and methodologies, to be compatible with possible present intrusion detection systems. For this case the Intrusion Detection Message Exchange Format (IDMEF) is suitable. This format is defined in the RFC 4765 [1]. The purpose of the IDMEF is to define data formats and exchange procedures for sharing necessary information between intrusion detection and response systems and also to the management systems that may need to interact with them. Within that specification the data model is described to represent the information and the implementation in the Extensible Markup Language (XML) is presented. To realize this, a XML Document Type Definition (DTD) was developed for the specification. Beside the normative DTD a XML schema for the structure is also given in the specification, providing a definition for XML data, which is mostly used nowadays [2]. The requirements for this communication mechanism are specified in the RFC 4766 [2].

Based on this format, a new messaging system was defined. The main functions of the new messaging system are alarm creation, and logging of information. The XML structure for these messages was derived as a specialization of the general IDMEF-Message structure. The IDMEF-Message data structure definition is briefly shown in Fig. 2.

After short introduction in first section implementation section describes specification and implementation Intrusion Detection System to a heterogeneous network. Last section concludes reached results, and discusses future ways of implementation in real secured networks.

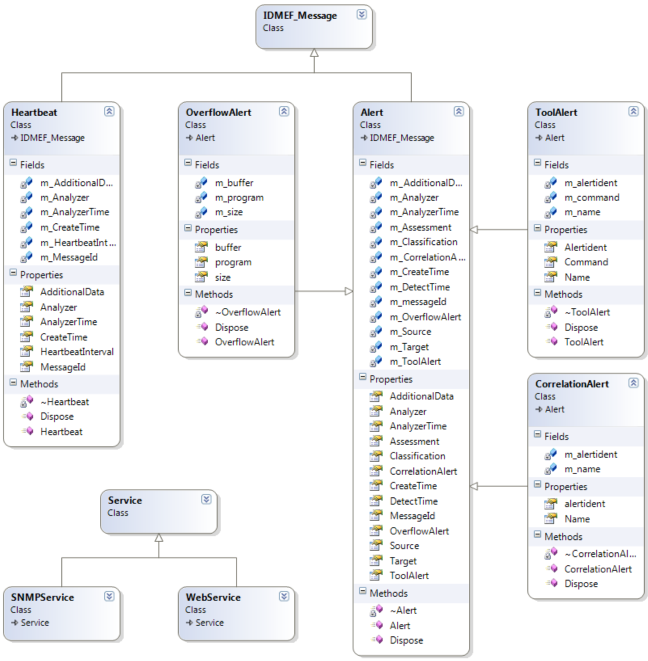


Fig. 2: Basic IDMEF-Message class diagram

II. IMPLEMENTATION

Implementation of Intrusion Detection System can be split into two implementation areas. The first area is implementation of IDS into firewall firmware and the second is developing of IDS message logger software. Both areas are interconnected by alert messages according to IDMEF-Message standard.

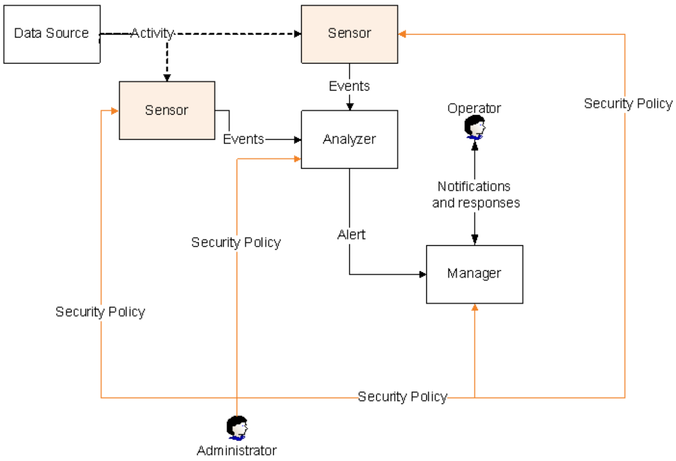
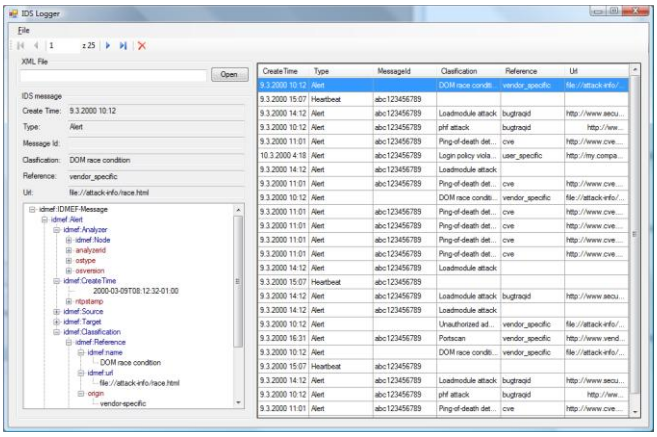


Fig. 3: Block diagram of the Intrusion Detection System [3]

The new automation firewall generates alert messages in XML file type according to RFC 4766 [2] and sending it to IDS message logger. For processing and logging of created messages by the Intrusion Detection System a software tool was developed. The IDS Logger offers GUI (Graphical User Interface) for displaying IDMEF messages and processing these messages to database for later evaluation. It also offers possibility to evaluate stored messages from database and export stored alerts back to the XML file.

Main window of the IDS logging software tool is shown in Fig. 4.



IDS message

Create Time: 9.3.2000 10:12

Type: Alert

Message Id:

Clasfication: DOM race condition

Reference: vendor_specific

Url: file://attack-info/race.html

```
idmef:IDMEF-Message
  idmef:Alert
    idmef:Analyzer
      idmef:Node
      analyzerid
      ostype
      osversion
    idmef:CreateTime
          2000-03-09T08:12:32-01:00
      ntpstamp
    idmef:Source
    idmef:Target
    idmef:Classification
      idmef:Reference
        idmef:name
          DOM race condition
        idmef:url
          file://attack-info/race.html
        origin
          vendor-specific
```

Create Time	Type	MessageId	Clasfication	Reference	Url
9.3.2000 10:12	Alert		DOM race conditi...	vendor_specific	file://attack-info/...
9.3.2000 15:07	Heartbeat	abc123456789			
9.3.2000 14:12	Alert	abc123456789	Loadmodule attack	bugtraqid	http://www.secu...
9.3.2000 10:12	Alert	abc123456789	phf attack	bugtraqid	http://ww...
9.3.2000 11:01	Alert	abc123456789	Ping-of-death det...	cve	http://www.cve....
10.3.2000 4:18	Alert	abc123456789	Login policy viola...	user_specific	http://my.compa...
9.3.2000 14:12	Alert	abc123456789	Loadmodule attack		
9.3.2000 11:01	Alert	abc123456789	Ping-of-death det...	cve	http://www.cve....
9.3.2000 10:12	Alert		DOM race conditi...	vendor_specific	file://attack-info/...
9.3.2000 11:01	Alert	abc123456789	Ping-of-death det...	cve	http://www.cve....
9.3.2000 11:01	Alert	abc123456789	Ping-of-death det...	cve	http://www.cve....
9.3.2000 11:01	Alert	abc123456789	Ping-of-death det...	cve	http://www.cve....
9.3.2000 11:01	Alert	abc123456789	Ping-of-death det...	cve	http://www.cve....
9.3.2000 14:12	Alert	abc123456789	Loadmodule attack		
9.3.2000 15:07	Heartbeat	abc123456789			
9.3.2000 14:12	Alert	abc123456789	Loadmodule attack	bugtraqid	http://www.secu...
9.3.2000 14:12	Alert	abc123456789	Loadmodule attack		
9.3.2000 10:12	Alert		Unauthorized ad...	vendor_specific	file://attack-info/...
9.3.2000 16:31	Alert	abc123456789	Portscan	vendor_specific	http://www.vend...
9.3.2000 10:12	Alert		DOM race conditi...	vendor_specific	file://attack-info/...
9.3.2000 15:07	Heartbeat	abc123456789			
9.3.2000 14:12	Alert	abc123456789	Loadmodule attack	bugtraqid	http://www.secu...
9.3.2000 10:12	Alert	abc123456789	phf attack	bugtraqid	http://ww...
9.3.2000 11:01	Alert	abc123456789	Ping-of-death det...	cve	http://www.cve....

Fig. 4: Main window of IDS logging software [4]

An IDMEF-Message can be either an alert or a heartbeat message. In case of security device or application it won’t send any regular status information, then the *Heartbeat* is not necessary and only the *Alert* message is used. For keeping backward compatibility with full IDMEF-Message specification we implemented also *Heartbeat* message. Each IDMEF-Message element has a *version* attribute that must contain string "1.0" for the value according to the specification.

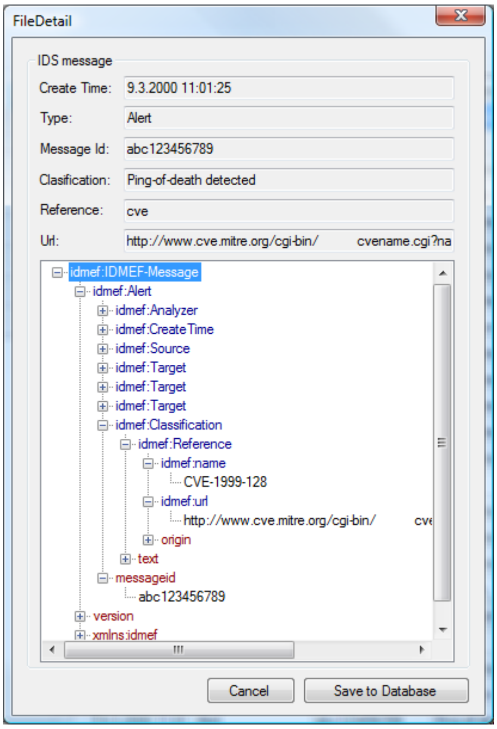
The *Alert* message is generated every time an event occurs. In this case it means, a packet filter rule is fired with a higher severity level, the access to a security relevant function is denied or a similar event occurred. This element has an attribute called *messageid* that contains a unique identifier for this message.

The *Alert* message contains exactly one *Analyzer* element, holding the information about the analyzer, from which the alert originates. The most statements within this element are optional. Because there can be several security measures in an automation network, it is recommended to provide some basic information. The element contains the attribute *analyzerid*, which specifies the unique identifier for this application. The alert message also contains exactly one *CreateTime* element. This attribute specifies the time when this message was created. This element is the only time related element within IDMEF which is required. The value of this entry is the time, a regulative rule of an automation network security system is fired, which caused the creation of this message. Because there is no difference between the occurrence of the event and the creation of the message the optional element *DetectTime* is not used. The values of those time elements are the date and the time according to the dateTime data format and attribute *ntpstamp* which contains the time in the NTPSTAMP data type.

The next element within the alert message is the *Source* element, which contains the information about the sources of the event leading to the alert. This element is an array of *Source* elements for defining unlimited number of alert sources. It contains the *Node* element, which holds the address information about the source of the packet causing

ISBN:978-1-61208-114-4

the filter rule to fire. To notify of the port number and the protocol of the source, the *Service* element can be used. The *Service* node is also part of *Source* attribute.



FileDetail

IDS message

Create Time: 9.3.2000 11:01:25

Type: Alert

Message Id: abc123456789

Clasification: Ping-of-death detected

Reference: cve

Url: http://www.cve.mitre.org/cgi-bin/cvename.cgi?na

```
idmef:IDMEF-Message
  idmef:Alert
    idmef:Analyzer
    idmef:CreateTime
    idmef:Source
    idmef:Target
    idmef:Target
    idmef:Target
    idmef:Classification
      idmef:Reference
        idmef:name
          CVE-1999-128
        idmef:url
          http://www.cve.mitre.org/cgi-bin/        cve
        origin
      text
    messageid
      abc123456789
  version
  xmlns:idmef
```

Fig. 5: Detail view of the IDMEF message

To send the information about the intended targets of the event that caused the alert the element *Target* is used. It has almost the same structure as *Node* and *Service* attributes.

The *Node* element is used to represent information about a host or a device. This element can hold a *Name* and unlimited *Address* elements. In the new automation network firewall application, this node holds address information.

The *Address* element is used to represent any form of address of hardware or an application. It has an attribute *Category* to represent the type of address. The following keywords are used:

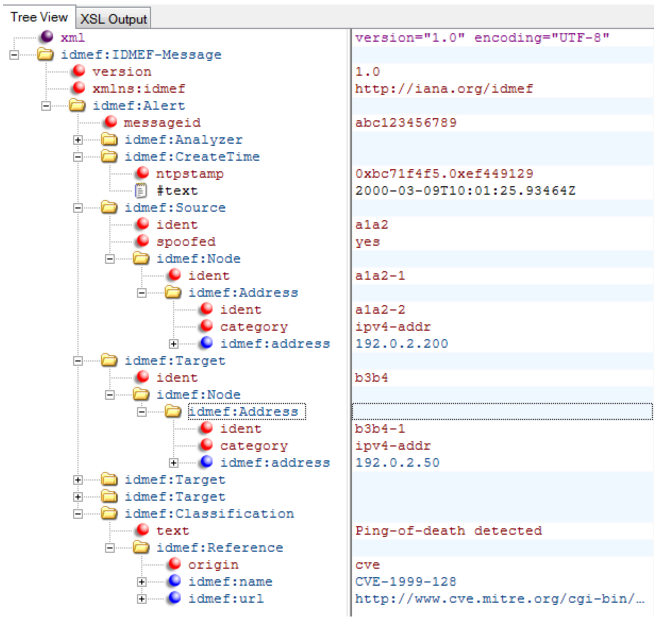
- *unknown* - for a not specified address type,
- *mac* – for the hardware Media Access Control address,
- *ipv4-addr* – for an IP address in the dotted-decimal address,
- *ipv4-net* – for an IP network address range in dotted-decimal with slash and significant bits (e.g. 123.123.123.123/24).

Beside this attribute it contains the string type element *Address*, holding the actual address.

The *Service* element is used to identify services for the source or the target by name, port, and protocol. This element has the optional attributes *iana_protocol_number* and *iana_protocol_name*.

The alert message also has exactly one *Classification* element, to provide a naming for the alert. It contains a text attribute with the name which illustrates type of the alert. This element can be used to distinguish a remote logging from a real alert.

Another elements are not required and not useful for the new automation network security messaging system but to keep IDMEF standard, IDS logger is fully compatible with other IDMEF messages and can process all attributes and nodes defined in RFC 4765 [1].



```
xml                                 version="1.0" encoding="UTF-8"
idmef:IDMEF-Message
  version                           1.0
  xmlns:idmef                       http://iana.org/idmef
  idmef:Alert
    messageid                       abc123456789
    idmef:Analyzer
    idmef:CreateTime
      ntpstamp                      0xbc71f4f5.0xef449129
      #text                         2000-03-09T10:01:25.93464Z
    idmef:Source
      ident                         a1a2
      spoofed                       yes
      idmef:Node
        ident                       a1a2-1
        idmef:Address
          ident                     a1a2-2
          category                  ipv4-addr
          idmef:address             192.0.2.200
    idmef:Target
      ident                         b3b4
      idmef:Node
        idmef:Address
          ident                     b3b4-1
          category                  ipv4-addr
          idmef:address             192.0.2.50
    idmef:Target
    idmef:Target
    idmef:Classification
      text                          Ping-of-death detected
      idmef:Reference
        origin                      cve
        idmef:name                  CVE-1999-128
        idmef:url                   http://www.cve.mitre.org/cgi-bin/…
```

Fig. 6: Example of alert message according to IDMEF-Message definition

All captured messages from the IDS are stored in the MS SQL database. Due to the complexity of IDMEF are all alert messages stored in one database table instead of relational database tables. In this database table are stored most important parameters along with whole XML alert string. This solution saves processing time and allows us potential recreation of original alert message.

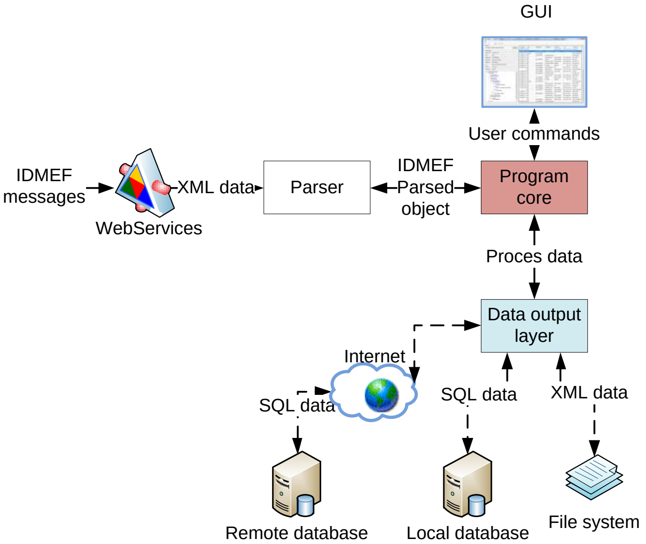


Fig. 7: Block structure of developed IDS

III. CONCLUSION

A complete implementation of IDMEF into the new automation messaging system was described. Messaging system based on the IDMEF is perfectly suitable for a new automation network firewall due to its compatibility with other similar devices. Identified and captured alert message is processed and stored by IDS Logger tool to MS SQL database. This solution allows controlling of all used firewalls in large automation network, over one database and in one easily managed place. Stored alerts are the most important sources of security information. User can monitor all attempts to un-allowed connections and intrusions to the internal secure area of the automation network. Based on the logged alerts, user can have perfect knowledge about security risks and can effectively react and protect all automation devices within automation network.

ACKNOWLEDGMENT

This research has been supported by the Czech Ministry of Education, Youth and Sports in the frame of MSM 0021630503 *MIKROSYN New Trends in Microelectronic Systems and Nanotechnologies* Research Project, partly supported by ARTEMIS JU in Project No. 100205 Process Oriented Electronic Control Units for Electric Vehicles Developed on a multi-system real-time embedded platform and by ENIAC JU in Project No. 120001 Nanoelectronics for an Energy Efficient Electrical Car, partly by the Czech Ministry of Industry and Trade in projects FR-TI1/057Automatic stocktaking system and FR-TI1/057 Intelligent house-open platform.

REFERENCES

[1] H. Debar, D. Curry, and B. Feinstein: *The Intrusion Detection Message Exchange Format IDMEF*; RFC 4765; IETF;1. March 2007, http://tools.ietf.org/html/rfc4765.

[2] M. Wood and M. Erlinger: *Intrusion Detection Message Exchange Requirements*; RFC 4766; IETF;1. March 2007; http://tools.ietf.org/html/rfc4766.

[3] ANSI/ISA, ANSI/ISA-99.02.01-2009 *Security for Industrial Automation and Control Systems*, Establishing an Industrial Automation and Control Systems Security Program, 2009.

[4] VAN – *Virtual Automation Network, Real Time for Embedded Automation Systems including Status and Analysis and closed loop Real time control*, Real-time for Embedded Automation Systems deliverable, 6th Framework Program, 1.8. 2008, http://www.van-eu.org/sites/van/pages/files/D04.1-1_FinalV1_2_060702.pdf

ISBN:978-1-61208-114-4

End-User Development Success Factors and their Application to Composite Web Development Environments

David Lizcano, Fernando Alonso, Javier Soriano and Genoveva López
DLSIIS, Dept. of Computer Science
Universidad Politécnica de Madrid
Madrid, Spain
Email: dlizcano@fi.upm.es

***Abstract*—The Future Internet is expected to be composed of a mesh of interoperable Web services accessed from all over the Web. This approach has not yet caught on since global user-service interaction is still an open issue. Successful composite applications rely on heavyweight service orchestration technologies that raise the bar far above end-user skills. The weakness lies in the abstraction of the underlying service front-end architecture rather than the infrastructure technologies themselves. In our opinion, the best approach is to offer end-to-end composition from user interface to service invocation, as well as an understandable abstraction of both building blocks and a visual composition technique. In this paper we formalize our vision with regard to the next-generation front-end Web technology that will enable integrated access to services, contents and things in the Future Internet. We present a novel reference architecture designed to empower non-technical end users to create and share their own self-service composite applications. A tool implementing this architecture has been developed as part of the European FP7 FAST Project and EzWeb Project, allowing us to validate the rationale behind our approach.**

***Keywords*-End-User Development, User-Centred Service-Oriented Architectures, Service Front-Ends, Composite Applications, Future Internet, Internet of Services**

I. Introduction

Service-Oriented Architectures (SOA) have attracted a great deal of interest over the last few years. In fact, SOAs increase asset reuse, reduce integration expenses and improve business agility in responding to new demands [1].

Nonetheless, mainstream development and research into SOAs have until now focused mainly on middleware and scalability, service engineering and automating service composition using business modelling process (BPM) technologies. Little or no attention has been paid to service front-ends, which we view as a fundamental part of SOAs [2]. As a result, SOAs remain on a technical layer hidden away from the end user.

The evolution of Web-based interfaces bears testimony to the progress made towards improving service usability. However, existing, web-based service front-ends do not come at all close to meeting end-user expectations [3]. Applications and information portals are still based on monolithic, inflexible, non-context-aware, non-customizable and unfriendly user interfaces (UIs). Consequently end users do not really benefit from the advantages promoted by service orientation in terms of modularity, flexibility and composition [4]. In addition service front-ends are constructed ad hoc without formal engineering methods and tools that could accelerate the time to market.

The vision presented here is an early result of the Service Front End (SFE) Open Alliance initiative[1]. This initiative aims to integrate results from several relevant R&D projects in the field to produce open specifications and an open source reference implementation of components of an envisioned Web platform to access services, contents and things in the Future Internet. This would enable end-user development (EUD) of software solutions based on user-centred services. The SFE Open Alliance initiative was setup under the umbrella of activities within the Service Front Ends Collaboration Working Group created in the FP7 call and currently involves projects such as FAST or EzWeb.

Section II of the paper states the service front-end problem. Section III presents a framework for studying EUD success in current solutions in order to elicit vital guiding principles to drive our search for the shortcomings of existing service front-end technology. Then Section IV proposes a novel reference model and architecture that empowers end users and supports this vision. This architecture will enable the creation of new ecosystems, where all stakeholders will be able to collaboratively develop capabilities and innovate new operating procedures by mixing and integrating already available services. Then, the above ideas are briefly validated in Section V. Finally, Section VI discusses the main conclusions of this research.

II. Shortcomings on the Road Towards an Internet of Services enabling EUD

The provision and consumption of information-intensive electronic services across corporate boundaries has attracted considerable interest over recent years. Particularly the Web services technology stack [1] was expected to act as efficient and agile "plumbing [. . .] for information systems to interact

[1] Open Alliance for Service Front Ends, http://sfe.morfeo-project.org

without human involvement" [5]. Following the design principles suggested by SOAs, Web services provide a uniform, system-independent way for interlinking dispersed electronic services. While technology and standards are important for achieving the vision of a globally networked, agile service economy, it has been widely recognized today that they are not sufficient on their own [3]. Analyses of today's cross-organizational service interconnections following the SOA paradigm have resulted in the identification of the following major weaknesses:

- ***Rigid and process-oriented composition***. Not all the potential of SOAs has been unleashed yet. Adherence to merely process-oriented design principles leads to rigid applications that cause huge reprogramming efforts in the event of changes. As in the 1970s, where the prevalence of Spaghetti-code-like software programming led to unmanageability and unchangeability of applications (the software crisis [6], [7]), the application of inflexible service orchestration techniques (e.g., based on BPEL (Business Process Execution Language)) prevents SOAs today from being truly agile.
- ***Deficient interoperability***. A second major issue of today's SOAs concerns service interoperability. Sometimes referred to as the "corporate household problem", information objects defined as input or output messages of services are based on highly proprietary specifications. The resulting semantic and syntactical heterogeneity causes significant mapping efforts when different services are to be interconnected and often leads to errors and increased costs.
- ***Limited retrievability***. In today's Internet, a lack of comprehensive, trustworthy and widely accepted service registries is another roadblock on the way to a networked service economy. In fact, a number of intermediaries are required to provide rich navigation to users, as well as to improve transparency and thus fulfil institutional functionality.
- ***Mute and autistic service interfaces***. Technologies such as the above Web services stack aim at supporting the setup of loosely coupled application interconnections especially in a professional context and assume users to be technically qualified. WSDL-based interfaces, for example, do not allow for rich interaction between machines and human users, but rather focus on automated machine-to-machine interoperation.

All these weaknesses prevent current services from being really useful for non-tech end users, who are unable to easily exploit them to develop and compose their own solutions. Therefore, Section III will review and analyse major current EUD solutions with the aim of exploring the design principles, factors and issues to be dealt with in order to achieve EUD success.

III. Existing successful EUD solutions

Nowadays there are several applications empowering end users without programming skills to develop their own software solutions, fitted to their unique and instant requirements. These applications, like spreadsheets, e-mail filter creators or mashup Web platforms, focus on outputting different software solutions, each oriented to a specific problem domain, such as calculation requirements, spam filtering or visual Web widget composition.

Of all these approaches, several studies state that spreadsheets are the most relevant and successful EUD solution existing at present [8]. In an empirical study carried out by Wu et al. (2007), 100% of a huge sample of end users had at some time used a spreadsheet program in their daily work to solve some problem or other. Other publications, like Boehm et al. (2005), establish that more than 55 million people in the United States do this kind of EUD programming, whereas professional programmers account for about 2.75 million of the country's population. This gap is actually widening, as Scaffidi et al. (2005) predict that the EUD population (users of spreadsheets and other EUD approaches) at workplaces in the United States will be 90 million by 2012. These studies and publications indicate that EUD is about to take control not only of personalizing computer applications and writing programs but of designing new computer-based applications without ever seeing the underlying program code.

For this reason many researchers around the globe have begun to study EUD success factors, focusing above all on the most relevant EUD solution, spreadsheets, to understand which of their principles and factors are used and accepted by end users. Studies like [8] focus on successful human EUD factors, since other studies like [9] focus on HCI (human-computer interaction) factors or on aspects of specialization and functionality [10].

The feedback that we get after reviewing all referenced studies is that EUD success is related to *human factors*, *HCI factors* and the *specialization-functionality relationship*. However, there are no publications that put all these ideas together to offer a general conception of EUD success factors. In this paper we will review the most relevant publications and scientific results in the EUD domain and then take a step further by combining them in an innovative EUD success framework. This way, we will be able to study each EUD solution and form a general idea of how successful it is likely to be among end users based on the studied factors. Additionally this framework will be useful for obtaining which design and architectural decisions are relevant for achieving end-user satisfaction, something that is vital for improving future EUD solutions such as the one proposed in this paper and fostering EUD success.

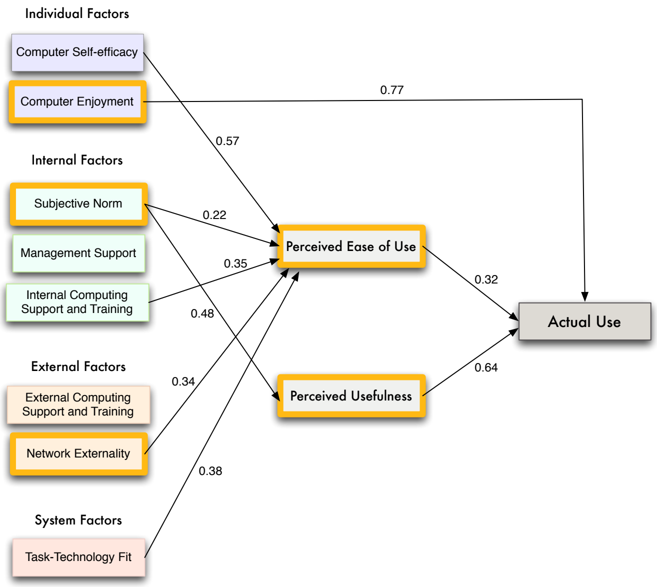


Figure 1. Empirical results of study about human factors related to EUD success

A. Successful Human Factors

All software development tools should be well accepted by their target users if they are to be considered a successful solution. However, this acceptance, as Wu et al. (2007) show in their empirical survey, is not down to the choice of particular technologies or architectural decisions, but because they preserve and take care of a number of human factors.

End-user computer acceptance (EUC) has been established by Wu et al. as one of the critical success factors in achieving business success, and is defined as the adoption and use of information technology by personnel outside the IT domain to develop software applications in support of organizational tasks. Davis [11] proposed the technology acceptance model derived from the reasoned action theory that has been tested and extended by numerous empirical researchers. In these studies the actual use of any application is derived from several human factors as perceived ease of use, perceived usefulness, and so on, and these ideas were the basis for Wu et al.'s research.

The empirical study carried out by Wu et al. relied on 800 people testing programs and evaluating software solutions. The evaluation showed that actual software use follows the causal relationships illustrated in Figure 1. This diagram establishes what factors are related to the actual end use, and what is the weight or strength of this causal relation, expressed by a correlation coefficient. The most relevant factors are explained in detail below.

- Perceived Ease of Use: The degree to which a person believes that using specific software would be effortless.

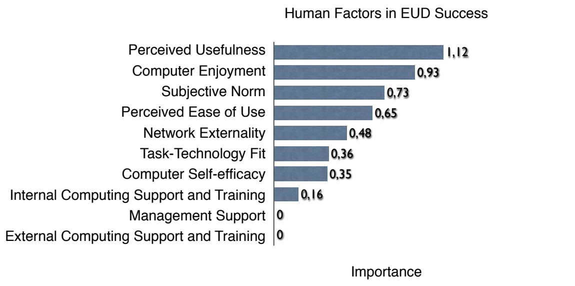


Figure 2. Importance of each human factor related to EUD success

- Perceived Usefulness: The degree to which a person believes that using specific software will increase his or her job performance.
- Computer Self-Efficacy: A person's perception of their ability to use computers to complete a task.
- Computer Enjoyment: Individuals that experience immediate pleasure and joy from using software.
- Subjective Norm: The degree to which a person believes that people that are important to him or her think that he or she should do the thing in question.
- Internal Computing Support and Training: Technical support and the amount of training provided inside the company.
- Network Externality: The utility of software use increases if the number of users increases.
- Task-Technology Fit: The degree to which an organization's application meets the information needs of the task.
- Management support or external training provided from outside the company is not related to end use according to this study.

In Figure 1 there are factors with multiple weighted paths to the final concept "actual use", and therefore the correlation coefficient between each factor and the use of a program is not clear. In Figure 2 all correlation coefficients have been recalculated in order to show the final impact of each factor on the actual use of the software. This way, it is possible to scale every factor and get an idea of its relevance.

This study suggests that an end user will use a program if he or she perceives it to be useful and enjoys the experience of using it. If an application is used in the end user's environment, it is more natural for him or her to accept and use this software too. Finally, ease of use and the fact that application usefulness would increase when it is used by more and more users will cause more actual use of a software tool.

B. Successful HCI Factors

Other studies like Jones et al. (2003) claim that spreadsheets (and other similar EUD solutions) are the programming language of choice for many people because of their human-computer interaction facilities. Spreadsheets are a user-centred approach to language design, focusing on fostering usability through effective human-computer interaction. Specialized research into the psychology of programming and empirical studies of programmers [9] offer a groundwork for human issues in programming, structured as cognitive dimensions, that EUD solutions should consider and optimize in order to be successful among end users. These cognitive dimensions prove to be HCI factors that, if properly taken care of, result in high end-user acceptance on a par with spreadsheets. Green and Petre (1996) defined 13 cognitive dimensions (all of which were of equal importance) that, if well looked after, improve HCI and simplify EUD [12]. The most quoted of these factors are listed below:

- Abstraction gradient: What are the minimum and maximum levels of abstraction? Can fragments be encapsulated?
- Consistency: When some of the language has been learnt, how much of the rest can be inferred?
- Error-proneness: Does the design of the notation induce “care-less mistakes”?
- Hidden dependencies: Is every dependency overtly indicated in both directions? Is the indication perceptual or only symbolic?
- Premature commitment: Do programmers have to make decisions before they have the information they need?
- Progressive evaluation: Can a partially complete program be executed to gather feedback on “How am I doing”?
- Role expressiveness: Can the reader see how each component of a program relates to the whole?
- Viscosity: How much effort is required to make a single change?
- Visibility and juxtaposability: Is every part of the code simultaneously visible, or is it at least possible to compare any two parts side-by-side at will? If the code is dispersed, is it at least possible to know in what order to read it?

According to Jones et al.’s study, software that looks after these factors, like Microsoft Excel or other spreadsheet solutions, enable users without programming skills to implement software in a simple and flexible manner. Therefore, these dimensions must be kept in mind when new EUD approaches are set out.

C. *Successful Specialization/Functionality Trade-off*

For many researchers in the EUD and composite applications domain, the most relevant factor for success among end users is that EUD software accomplishes a good trade-off between the specialization and the functionality of the created solutions [10]. This relationship gives an idea about whether end users could create their own solutions to satisfy their needs. How well suited a developed solution is for a task or real problem could be quantified by two factors:

- Specialization: the degree to which an application exactly matches real requirements, features and details of a real problem.
- Functionality: the sum or any aspect of what a product, such as a software application or computing device, can do for a user. The overall functionality decreases when the solution is overly specialized for a specific problem.

However, these factors are opposite. It is impossible to increase one factor without decreasing the other. For this reason, EUD solutions should adopt a trade-off where both factors are at equilibrium. This will lead to solutions that are very specialized for a problem but could be easily exported and used in other problem domains. This balance is frequently measured on a four-point Likert scale (poor, average, good or optimal specialization/functionality relationship) [10].

D. *Framework for Studying and Eliciting key EUD Success Factors to Improve EUD Solutions*

All the factors explained above are frequently referenced and used in EUD research, but they are always applied individually. In this paper we propose the creation of a complex framework to study key EUD success factors globally by combining all the studied factors.

First of all, they all have to be compared to study interrelations. Because each factor type focuses on one vertex of the HCI domain (human, human and computer interaction and software), we must conclude that human factors, HCI factors and the specialization/functionality factor are orthogonal to each other. Bearing this premise in mind, it is possible to join each family of factors as an independent axis in a 3D plot that represents each independent family of factors in a visual manner (Figure 3). Each axis must be managed as follows:

- X-axis = human factors. In the last section we described eight factors that should be considered to achieve EUD success. These factors had correlation coefficients to indicate their relevance. In the proposed framework, the study of an EUD solution will include an evaluation of every factor according to a three-point Likert scale (0 for low factor rating, 1 for an average rating and 2 for a high rating). This rating will be multiplied by the correlation coefficient (shown in Figure 2) to output a final rating for this factor. Every factor must be evaluated and added together for each EUD solution studied. This will add up to a final rating of from 0 to 9.56 (due to correlation coefficients). Finally, this rating has to be normalized to a standard scale ranging from 0 to 10 (by multiplying by 10/9.56). This final value will be represented on the X-axis and give visual information about how successful the studied solution would be in terms of EUD based on human factors.
- Y-axis = HCI factors. In the previous section we studied thirteen HCI factors that should be improved to achieve

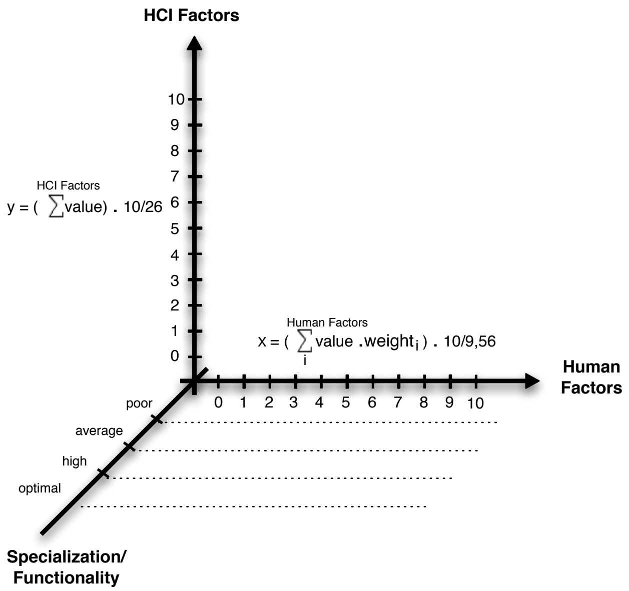


Figure 3. Framework of key EUD success factors

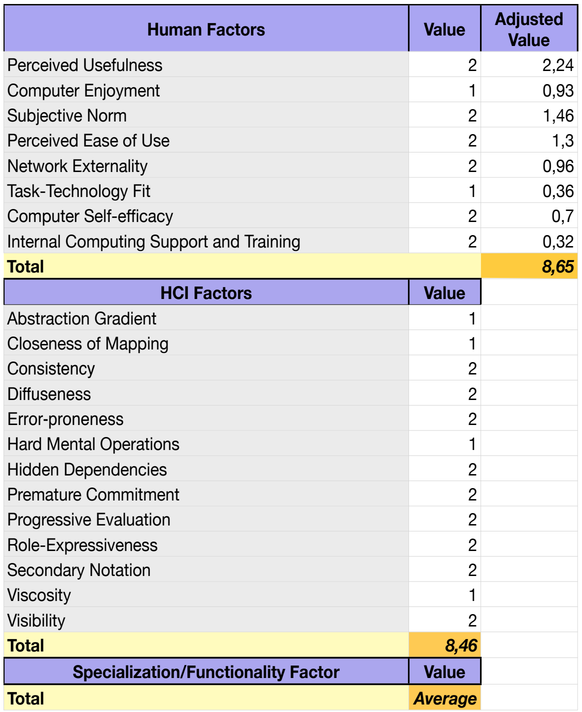
EUD success, all of which were equally important. Therefore, an analysis of these factors for an EUD solution will involve evaluating each factor on a three-point Likert scale (0 for a low rating, 1 for an average rating and 2 for a high rating of the factor) and add up this value for each factor. This process will output an overall rating ranging from 0 to 26 points. Finally, this value has to be normalized to an understandable scale (0-10) by multiplying by 10/26. This final value will be represented on the Y-axis giving a visual idea about how successful the studied solution would be in terms of EUD based on its cognitive dimensions.

- Z-axis = Specialization/functionality trade-off: the four-point Likert score studied in Section III-C could be represented directly on the z-axis, adding a third dimension to the solution's expected EUD success.

This framework is very useful in two ways:

- It is a powerful tool for studying any EUD solution and forming an idea of expected EUD success based on extended and proven principles, founded on factors included in several research papers.
- This framework summarizes all proven factors that are related to actual use and user acceptance, so it is a good starting point for creating new EUD environments or approaches.

Starting from referenced studies about spreadsheets and specific solutions like Excel, we can exploit user impressions and evaluations [8], [9], [10] to plot the expected success of Excel using the presented framework. This way, we could observe the performance of a successful EUD solution in the framework, giving an idea of what is the ultimate goal when new EUD solutions are to be developed. Table 4 presents a user evaluation of Excel, whereas the final rating of Excel is shown in Figure 5.



Human Factors	Value	Adjusted Value
Perceived Usefulness	2	2,24
Computer Enjoyment	1	0,93
Subjective Norm	2	1,46
Perceived Ease of Use	2	1,3
Network Externality	2	0,96
Task-Technology Fit	1	0,36
Computer Self-efficacy	2	0,7
Internal Computing Support and Training	2	0,32
Total		***8,65***
HCI Factors	**Value**	
Abstraction Gradient	1	
Closeness of Mapping	1	
Consistency	2	
Diffuseness	2	
Error-proneness	2	
Hard Mental Operations	1	
Hidden Dependencies	2	
Premature Commitment	2	
Progressive Evaluation	2	
Role-Expressiveness	2	
Secondary Notation	2	
Viscosity	1	
Visibility	2	
Total	***8,46***	
Specialization/Functionality Factor	**Value**	
Total	***Average***	

Figure 4. Excel evaluation from studies [8] and [9]

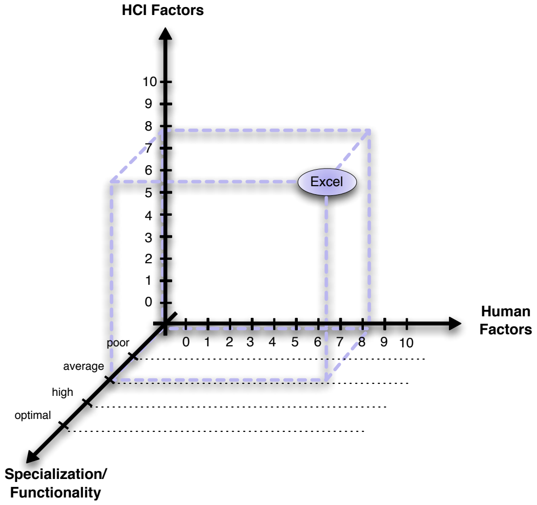


Figure 5. Excel EUD success represented in the proposed framework

E. Guiding EUD Principles Enabling the Internet of Services

From our point of view, the factors and principles explained previously state the need for user-centric SOAs based on a new generation of service front-end technologies in order to achieve EUD success through the Internet of Services. Such technologies will enable the massive deployment of services on the Internet, driven by the following guiding principles:

- ***End-User Empowerment***, enhancing traditional user-service interaction by facilitating the selection, creation, composition, customization, reuse and sharing of applications in a personalized operating environment [13].
- ***Seamless Context-Aware User-Service Interaction***. New-generation service front-ends should have the capability to detect, represent, manipulate, and use contextual information to adapt seamlessly to each situation, supporting human users in a more effective, personalized and consistent way [3]. Novel engineering tools and methods should be devised in order to support context-aware service front-ends.
- ***End-User Knowledge Exploitation***. This principle aims to exploit users' domain knowledge and collective intelligence to improve service front-ends. End users' knowledge can be used to tag resources using light semantics, assist while interacting with services, enrich contextual information (e.g. by means of automatic user profiling) and infer new candidate processes to be later automated (on the back-end) [5].
- ***Universal Collaborative Business Ecosystems***. Enterprise systems should incorporate advanced user-centric, context-aware front-ends to enable their employees and other stakeholders to exploit and share their extensive domain expertise, and their thorough business knowledge [4]. Employees, customers, developers and providers will collaborate to create and improve enterprise applications, sharing, reusing, changing and combining existing context-aware components (services, contents, things...)[14].

The EUD solution presented in this paper has been conceived following the principles and ideas presented in the EUD success framework in order to procure as much end-user acceptance as possible. Finally, our approach is evaluated in the validation section to study its success and compare it with spreadsheet solutions.

IV. Materializing EUD Principles and Success Factors to Reach a User-Centred Internet of Services

In this section we propose a novel architecture for next-generation service front-ends. This architecture was devised in accordance with the presented guiding principles, and applies human, HCI and functionality/specialization factors studied in Section III. This whole complex architecture is being researched and developed as part of several major R&D&i projects like NEXOF-RA, EzWeb, MyMobileWeb and FAST, and its objective is that end-users with no programming skills could create their own software solutions, perfectly adapted to their instant and unique requierements. These projects are currently subsidized by the EU and the Spanish Ministry of Industry, Tourism and Commerce. For the sake of clarity we have separated the authoring and runtime phases of the service front-end lifecycle (see Fig. 6).

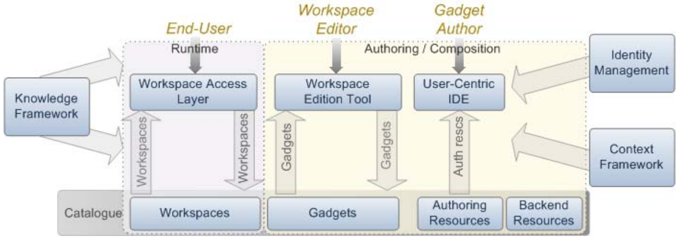


Figure 6. Proposed Architecture for Next-Generation Service Front-Ends (Overview) to Enable EUD

Gadgets will be the main building blocks of such an architecture. A gadget implements the user interface and application logic necessary to interact with one or more underlying services. Gadgets are self-contained front-end components focused on a single goal and, consequently, are of limited complexity. Gadgets can be grouped into workspaces.

Our proposed architecture for the authoring phase includes two main components:

- A "Gadget Authoring Tool", which it is a user-centric IDE dedicated to gadget design and creation. This is a visual tool that assists non-IT-aware users in creating their own service front-end resources [15]. Using this user-oriented IDE gadget, authors will be able to visually design, reuse and share gadget screens, flows and back-end resource compositions or connectors among others. Authors will easily compose a gadget from a series of building blocks (authoring resources) available in a palette. This palette is actually a specific view of the resource catalogue and can contain UI artefacts (screens), operators, screenflows, ready-to-use back-end resources and compositions, etc.
- A "Workspace Editing Tool" intended to design custom user workspaces, as a mashup editor. This tool will permit the visual design, reuse and sharing of user workspaces by selecting, connecting and composing the most suitable gadgets for dealing with a domain problem. The ultimate aim is to create new, modular and anticipated service front-ends (instant applications) by combining smaller pieces (gadgets). Each user can have and share any number of workspaces with other members of the community.

These two tools will be supported, at least, by the following formalisms:

- A Declarative Authoring Language for describing device and modality-independent user interfaces. Traditional user interface development approaches are insufficient for supporting the new-generation service front-ends. Taking one step further, traditional UI platforms and toolkits lack the formalisms necessary to deal with context-aware service front-ends. For example, there are no declarative mechanisms to specify how an interface should adapt according to different delivery contexts. Instead the developer needs to do the adaptation manually, using an ad-hoc and costly approach that does not promote reuse or standardization. We propose a layered approach to the development of UI for the services front-end: abstract UI (device independent), concrete UI (device dependent) and rendering for specific devices. All the layers can be represented in XML and embody a model of behaviour at a gradually finer level of detail.

- A standard format and infoset for the description of gadget metadata (Gadget Template). This template is a machine-readable gadget description that should at least contain information about author names and affiliations, date, a human-readable description, pointers to the gadget source code and, most importantly, publish-subscribe metadata depicting what data items the gadget publishes (including their type, name, semantics, etc.) and what data items are consumed by the gadget (including their type, name, semantics, etc.). Finally, the template should contain context metadata about the gadget.

At runtime we propose a "Workspace Access Layer" that is responsible for giving end users access to one or more workspaces. The layer will be in charge of rendering each user's workspace (and the gadgets it contains) depending on the characteristics of the target context, adapting a workspace and the gadgets it contains to the restrictions dictated by the target context and also implementing all the artefacts needed to support the execution of gadgets at runtime, such as publish and subscribe communication mechanisms or persistence of gadget data and state, providing a runtime environment for gadget execution.

Additionally, there will be a set of horizontal modules dedicated to different aspects that are common to the runtime and authoring phases and that were explained in Section III-E:

- A Resource Catalogue containing all the metadata about the different building blocks of the architecture (gadgets, screens, flows, workspaces, content delivery resources, application data resources, resource compositions, etc.).
- A Context Framework implementing all the concepts, formalisms and artefacts described previously in this paper.
- Identity and Session Management for dealing with user session and identity among others.
- A Knowledge Framework following the approach described in Section III-E of this paper.

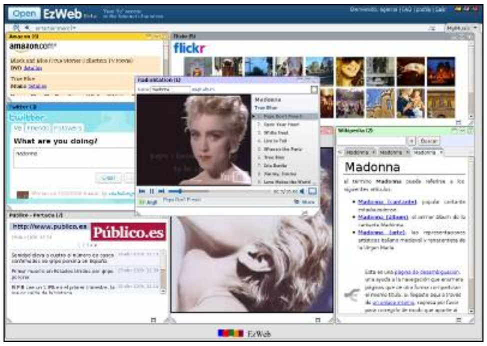
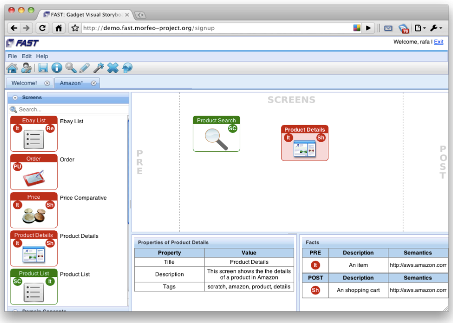


Figure 7. EzWeb platform



Properties of Product Details

Property	Value
Title	Product Details
Description	This screen shows the the details of a product in Amazon
Tags	scratch, amazon, product, details

Facts

PRE	Description	Semantics
It	An item	http://aws.amazon.com
POST	**Description**	**Semantics**
Sh	An shopping cart	http://aws.amazon.com

Figure 8. FAST IDE

V. Implementation and Validation of the Proposed Architecture

The proposed architecture was implemented on the EzWeb/FAST framework. This framework is composed of two tools derived from the EzWeb[2] (Figure 7) and FAST projects[3] (Figure 8). Both tools are publicly available at http://demo.ezweb.morfeo-project.org and http://demo.fast.morfeo-project.org, respectively.

EzWeb is a mashup platform where gadgets could be interconnected and arranged in several workspaces to satisfy instant requirements. Multi-purpose gadgets are published in a collaborative catalogue. However, if end users are unable to find what they need, they can easily create new gadgets using the FAST development environment, designed to enable non-technical users to create gadgets from more specific components called resources, available in public catalogues and around Internet. A short video, available at http://www.youtube.com/watch?v=qFt2LBlxkwU (part 1) and http://www.youtube.com/watch?v=dpoRhnF8_1A (part 2), demonstrates EzWeb/FAST features and teaches end users how to compose their own applications. Figure 9 illustrates the EzWeb/FAST composition model that strictly implements the reference architecture proposed in Section IV.

[2] Morfeo EzWeb, http://ezweb.morfeo-project.org

[3] Morfeo Fast, http://fast.morfeo-project.eu

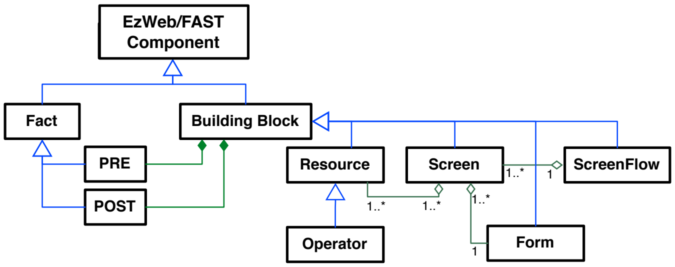


Figure 9. EzWeb/FAST composition model

Now that we have presented the EzWeb/FAST composition model, research focuses on evaluating its use and proving that our premise of enabling end users with no programming skills to build their own composite applications is feasible and true. For the evaluation we used the existing EzWeb/FAST framework that strictly implements the proposed architecture. If this framework is validated, then we will also have validated the underlying architecture model. EzWeb/FAST evaluation aims to test whether the developed user-centred composition system satisfies its usability, functionality and performance requirements. We present some of the findings of the research we conducted within the 2009/2010 FAST project reporting period [16].

In the above report we took a holistic approach to EzWeb/FAST system analysis from the information systems research viewpoint [17]. This perspective focuses on solving practical problems in the interaction between the organisation, people and information technology. Consequently, we conducted the research from the organisation, user, and information technology angles, aiming at gathering and structuring feedback about both the EzWeb/FAST system and the underlying architecture as a basis for improvement.

We experimented with three different approaches for running the holistic study, thus covering all the evaluation perspectives:

- An **expert evaluation** from a business consultant perspective. The expert was familiar with traditional business process modelling approaches, and represented the EzWeb/FAST system target user group.
- A **case study** as a strategy of empirical enquiry representing a specific real-life situation. It addresses the organisation and IT perspectives.
- A **laboratory experiment** under controlled conditions, covering the user perspective.

Based on the expert evaluation approach, we identified three key findings regarding the improvement of the EzWeb/FAST service composition system usability:

- **Design and runtime convergence**. This involves the composition of real-time data by non-experts. Composite applications are built from real data sources, consequently traditional test systems will have to evolve in new ways to support this convergence.
- **The organisation is similar to communities**. User-centred composition environments are structured similarly to a community. Building block sharing within the community is a key issue.
- **The need for user guidance**. Our composition system aims at supporting consultants in their daily work. Consultants usually have limited programming experience. This point uncovers the fact that users need guidance to create composite applications and extensive sample libraries in order to reuse existing solutions.

From an organizational perspective, the case study showed that the use of user-centred service composition systems has a number of interesting business benefits. The most important findings were that the use of EzWeb/FAST leads to higher employee productivity, as they can use the time they save to complete other tasks. It also improves user flexibility, involvement and satisfaction.

Last but not least, we conducted a laboratory experiment which accounted for the user perspective. This was the core and most important part of the evaluation, because it is the users that determine the success of the community- and user-oriented service composition paradigm

No.	Question
	Usability
Q1	FAST was easy to use first time round
Q2	I would imagine that most people would learn to use FAST very quickly
Q3	I felt confident using FAST
Q4	I didn't need to learn a lot about FAST before I could use it effectively
	Functionality
Q6	Screens were easy to find
Q7	The screenflow of a composite application was easy to model
Q8	I had no problem defining inputs and outputs
Q9	The designed composite applications were easy to publish
	Performance
Q14	The FAST system responded too slowly to inputs
Q15	The system ran stably
	General
Q16	I was able to set up screens for the B2B scenario
Q17	The evaluation task was too difficult

Table I
MAIN QUESTIONNAIRE QUESTIONS

One of the tasks set for the experiment participants was to create parts of a complex and real-world B2B scenario. To do this, they used the EzWeb/FAST components (i.e. *screenflows*, *screens*, *gadgets*, and so on.). This way, we were able to evaluate design and runtime convergence. They also had to fill in a *questionnaire* stating their individual opinion and impression of the EzWeb/FAST system. In a semi-structured *focus group*, they were asked about their attitude to EzWeb/FAST. And, apart from this, they were observed doing the main evaluation task. In sum, we applied three data collecting instruments (both quantitative and qualitative) to ensure that the quality of the results of the evaluation was beyond all question and to assure that the evaluation had a solid basis. The results of this user evaluation are summarized below.

We received feedback from a total of 41 participants, where 21 had little or no programming experience —business users—, whereas the other 20 were expert programmers —technical users. The results reveal that users have a positive impression of the EzWeb/FAST system. Figure 10 represents the statistical feedback based on the questionnaire data. The questions are listed in Table I. The EzWeb/FAST usability was rated as neutral with a positive tendency. We observed that participants found the components that they needed and did not have to do a lot of learning to use EzWeb/FAST. From a functionality perspective, we found that participants had some difficulties in finding the right screen to use. Additionally, input and output definitions were not self-explanatory. On the other hand, users found the procedure for publishing designed composite applications on a target platform easy to follow. Regarding the performance, EzWeb/FAST was stable and there were no critical incidents throughout the whole evaluation time frame.

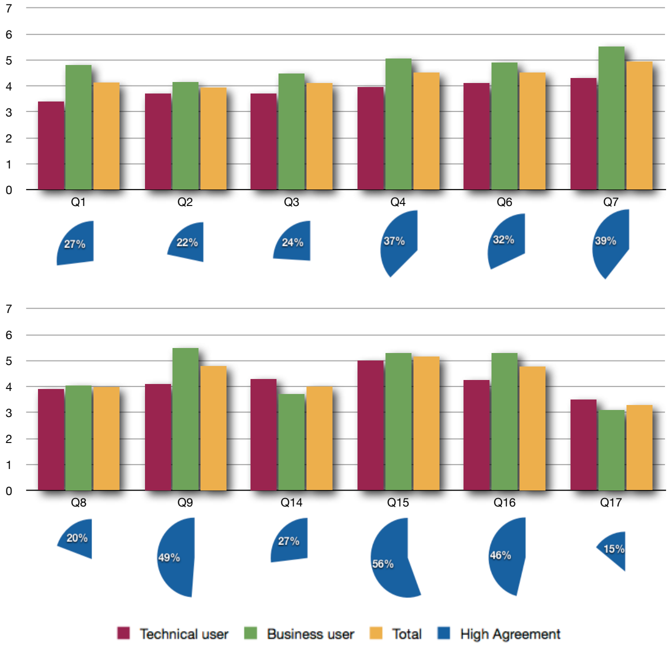
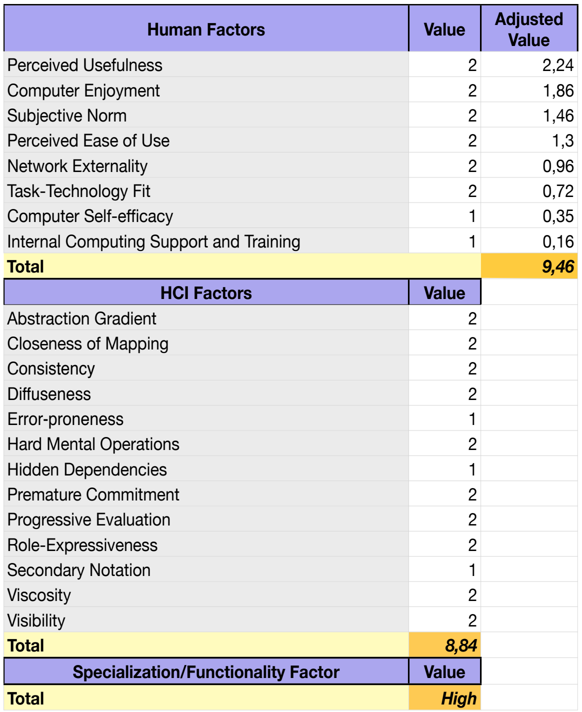


Figure 10. Questionnaire about EzWeb/FAST

Looking at the overall impression of EzWeb/FAST, it is noteworthy that more than 70% of the participants rated EzWeb/FAST as good or excellent, recognizing its EUD potential. A comparison of the impressions of technical and business people revealed an interesting point: business people are more enthusiastic about EzWeb/FAST than technical users. A possible explanation for this is user empowerment. These results indicate that users with no programming skills are able to create composite applications on their own, and, consequently, demonstrate that our composition model is valid too.



Human Factors	Value	Adjusted Value
Perceived Usefulness	2	2,24
Computer Enjoyment	2	1,86
Subjective Norm	2	1,46
Perceived Ease of Use	2	1,3
Network Externality	2	0,96
Task-Technology Fit	2	0,72
Computer Self-efficacy	1	0,35
Internal Computing Support and Training	1	0,16
Total		***9,46***
HCI Factors	**Value**	
Abstraction Gradient	2	
Closeness of Mapping	2	
Consistency	2	
Diffuseness	2	
Error-proneness	1	
Hard Mental Operations	2	
Hidden Dependencies	1	
Premature Commitment	2	
Progressive Evaluation	2	
Role-Expressiveness	2	
Secondary Notation	1	
Viscosity	2	
Visibility	2	
Total	***8,84***	
Specialization/Functionality Factor	**Value**	
Total	***High***	

Figure 11. EzWeb/FAST EUD success evaluation based on validation study

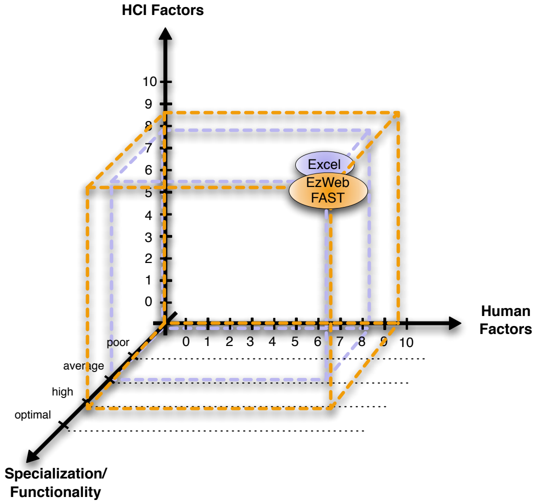


Figure 12. EzWeb/FAST EUD success represented in the proposed framework (and compared to Excel)

Finally, EzWeb/FAST was evaluated following the EUD success framework presented in the Section III-D. End users were asked about human, HCI and specialization/functionality issues, and Figure 11 presents the results of the questionnaires.

Analysing these data, EzWeb/FAST can be compared with other successful EUD solutions, like Excel (Figure 4), where it has proved to be even more suited for end-user requirements. Therefore, it should be successful for EUD (Figure 12).

VI. Conclusion

The first effect of the advent of user-centric approaches for constructing next-generation service front-ends such as the one proposed in this paper will be to unleash unprecedented potential with respect to the consumption of electronic services by different stakeholders. As a result, large enterprises will be able to capitalize on faster application development (thereby reducing application development backlogs in IT departments), a more agile system landscape, and the empowerment of their employees to contribute to the design of the applications that they are supposed to use. Small and medium-sized enterprises will be enabled to select and compose resources hosted by a wealth of third parties rather than paying for pre-determined, inflexible and potentially heavyweight solutions. Finally, private individuals will benefit from intuitive, unsophisticated ways to discover, remix and use the Web-based services that they consider interesting and useful to build a EUD environment based on a user-centred ecosystem of services.

Besides the discussed benefits for different user groups, the novel user-centric approach will also abet the large-scale proliferation of what is often referred to as the Internet of Things (IoT). Not until information gathered from the multitude of dispersed sensors is made accessible and usable through an agile service front-end architecture will the envisioned Internet of Things become reality.

Acknowledgment

This work was supported in part by the European Commission under the first call of its Seventh Framework Program (FAST STREP Project, grant INFSO-ICT-216048), the European Social Fund and the UPM under its researcher training programme.

References

[1] G. Alonso, F. Casati, H. Cuno, and V. Machiraju, *Web Services Concepts, Architectures and Applications*, ser. Data-Centric Systems and Applications. Germany: Springer, 2004.

[2] C. Schroth and O. Christ, "Brave new web: Emerging design principles and technologies as enablers of a global soa," in *Proceedings of the IEEE International Conference on Services Computing, 2007. SCC 2007*. Los Alamitos, CA, USA: IEEE Computer Society Press, 2007, pp. 597–604.

[3] D. Lizcano, J. Soriano, M. Reyes, and J. J. Hierro, "Ezweb/fast: Reporting on a successful mashup-based solution for developing and deploying composite applications in the upcoming "ubiquitous soa"," *Mobile Ubiquitous Computing, Systems, Services and Technologies, International Conference on*, vol. 0, pp. 488–495, 2008.

[4] A. P. McAfee, "Enterprise 2.0: The dawn of emergent collaboration," *MIT Sloan Management Review*, vol. 47, no. 3, pp. 21–28, 2006.

[5] ——, "Will web services really transform collaboration," *MIT Sloan Management Review*, vol. 46, no. 2, pp. 78–84, 2005.

[6] R. M. Balzer, "Imprecise program specification," *Report ISI/RR-75-36, Information Sciences Institute*, December 1975.

[7] M. D. McIlroy, "Mass produced software components," in *Software Engineering, Report on a conference sponsored by the NATO Science Committee, Garmisch, Germany*, October 1968, pp. 138–155.

[8] J.-H. Wu, Y.-C. Chen, and L.-M. Lin, "Empirical evaluation of the revised end user computing acceptance model," *Computers in Human Behavior*, vol. 23, no. 1, pp. 162 – 174, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/B6VDC-4CF5BTH-2/2/1c5438462f07b8c0745efb54bee4fed1

[9] S. P. Jones, A. Blackwell, and M. Burnett, "A user-centred approach to functions in excel," in *In ICFP 03: Proceedings of the eighth ACM SIGPLAN international conference on Functional programming*. Sweden, EU: ACM Press, 2003, pp. 165–176.

[10] H. Lieberman, F. Paternò, M. Klann, and V. Wulf, *End-User Development*, ser. Human-Computer Interaction Series. Germany: Springer, Nov. 2006, vol. 9, ch. End-User Development: An Emerging Paradigm, pp. 1–8.

[11] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "Extrinsic and intrinsic motivation to use computers in the workplace¡sup¿1¡/sup¿," *Journal of Applied Social Psychology*, vol. 22, no. 14, pp. 1111–1132, 1992. [Online]. Available: http://dx.doi.org/10.1111/j.1559-1816.1992.tb00945.x

[12] T. Green and M. Petre, "Usability analysis of visual programming environments: A cognitive dimensions framework," *Journal of Visual Languages and Computing*, vol. 7, no. 2, pp. 131–174, 1996.

[13] R. Smith, "Enterprise mashups: an industry case study," in *Keynote at New York PHP Conference and Expo*. NY, USA: IBM Software Group Press, Jun. 2006.

[14] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More*. NY, USA: Hyperion, July 2006.

[15] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, University of California, Irvine, 2000.

[16] V. Hoyer, A. Fuchsloch, S. Kramer, K. Moller, and J. López, "Evaluation of the implementation," FAST Consortium, Tech. Rep. D6.4.1, February 2010.

[17] A. S. Lee, "Remarks from mis quarterly editor - inaugural editor's comments," *MIS Quarterly*, vol. 23, no. 1, 1999.

Towards building health systems

Elaine Lawrence, Karla Felix Navarro
mHealth Laboratory, iNEXT
University of Technology Sydney, FEIT
Sydney, Australia
{Elaine.Lawrence;Karla.Felixnavarro}@.uts.edu.au

Jaime Andres Garcia Marin, Christian Sax
mHealth Laboratory, iNEXT
University of Technology Sydney, FEIT
Sydney, Australia
Jaime.GarciaMarin@uts.edu.au,
Sax.Christian@gmail.com

***Abstract*—This paper reports on a series of interviews with three mainstream and three alternative/complimentary health professionals about the use of interactive technologies as a tool to improve the physical and mental well-being of the elderly. The questions are based around an Analytic Framework for investigating Interactive Technologies and the elderly. Four balance games using the Wii were demonstrated to the interviewees and their opinion of the suitability or otherwise of these games are discussed. The chosen games did not meet with universal approval but did provide us with useful insights on how to ensure the development of appropriate games for the elderly.**

Keywords-interactive games; balance; elderly; healthcare.

I. INTRODUCTION

The percentage of aged persons over 65 is increasing dramatically both in Australia and worldwide and unfortunately, as people age, their mental and physical health deteriorates and impacts negatively on their quality of life. This paper specifically investigates ways in which Interactive technologies, namely the Nintendo *Wii* could help to overcome functional decline, maintain independence and preserve social connectivity and engagement among seniors [1]. We report here on the results of semi-structured interviews with six health professionals who have specialized in caring for the elderly and infirm. Each of the interviewees answered a series of questions before being shown a demonstration of four Wii balance games. They were asked to comment on the suitability or otherwise of the games for their clients. Our project should, in time, reveal (a) the potential impact of interactive computer technologies on client and health carer outcomes and satisfaction levels, (b) the potential impact of clients' abilities to adapt to the introduction of new technologies, and (c) the potential benefits and obstacles to the application of interactive technologies in aged care environments, both at client's homes and at aged care facilities [2] [3]. Because of word limit constraints this paper only addresses some of the comments of the interviewees.

Research has shown that it is vital to encourage the elderly and infirm to do physical and mental exercises and support them throughout the process in order to maintain good health outcomes [1] [2] [3]. Current information and communication technologies on assistive healthcare mainly focus on remote sensing using the Internet and wireless sensor networks for collecting health condition data of elderly people. Examples include the *ReMoteCare* system [4] [2] as well as the *Personal Health Monitor* developed at our university [5]. Very little effort has been made to support collaborative planning, implementation procedures for physical training (such as callisthenics and cardiovascular exercise), mental training and assessment such as described by [6]. Some further research on the use of the Wii for the elderly has been undertaken by [7] [8].

We have used an Analytic Framework based on work by [1] to investigate the aged population cohort, to study the economic and environmental factors and to help in assessing whether interactive technologies are useful for the elderly, chronically unwell and the infirm [3].

In Section 2 of the paper, we provide background information on the Analytic Framework and the interviewees. In Section 3, we describe the four chosen games and comments by the interviewees about their suitability or otherwise. We discuss the findings in Section 4 and draw our conclusions in Section 5, while pointing the way ahead for future research.

II. BACKGROUND

The authors have modified the Analytic Framework used by [1] in their landmark study on *Consumer Health Information Technologies* used by the elderly, chronically ill and underserved. Our modifications limited the technologies to Interactive Game Technologies, in this case, four Wii games that concentrate on balance. Figure 1 below sets out the Analytical Framework which served as a basis for our semi-structured interviews and demonstrations.

In July and August 2010 we conducted semi-structured interviews with six mainstream and alternative/complimentary health professionals who are used to dealing with the aged (See Table 1).

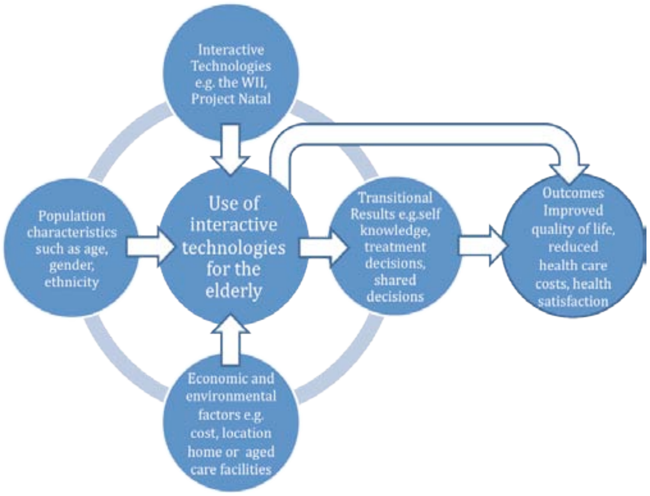

Interactive Technologies e.g. the WII, Project Natal
Population characteristics such as age, gender, ethnicity
Use of interactive technologies for the elderly
Transitional Results e.g.self knowledge, treatment decisions, shared decisions
Outcomes Improved quality of life, reduced health care costs, health satisfaction
Economic and environmental factors e.g. cost, location home or aged care facilities

Figure 1: Analytic Framework (based on [1])

TABLE I. INTERVIEWEES' DETAILS AND PSEUDONYMS

Pseudonym	Details
Professor Aged Care: Sydney Hospital Researcher and director of Health and the Aged Centre.	Professor of Aged and Extended Care Nursing at a Sydney University examining aged and dementia care, health technologies, models of aged and dementia care, health technologies and the aged. **Mainstream Health Professional.**
Feldenkrais Practitioner: Certified Feldenkrais Practitioner in Sydney.	The Feldenkrais method is designed to improve movement repertoire, aiming to expand and refine the use of the self through awareness, in order to reduce pain or limitations in movement, and promote general well-being[10]. He teaches Feldenkrais method to a diverse range of people such as those with injuries, the elderly, actors, athletes, or those who just want to move more easily. **Alternative or Complimentary Health Professional.**
Physiotherapist: for 28 years the physiotherapist at an large Aged Care Centre in Sydney	Responsible for assessing [the patients]. There are objective assessments for patients when they enter the facility and a care plan is then instigated for use by the rest of the nursing staff and also for the physiotherapy assistants to carry on with walking assistance or passive class exercises etc **Mainstream Health Professional.**
Alexander Technique Practitioner: Certified Alexander Technique Practitioner in Sydney. Own Practice.	The Alexander Technique is a subtle but extremely powerful re-education process that seeks to restore the natural physiological organization for posture, balance and movement to the human system. Deals with patients aged from teenagers to over 90 years old. Also teaches Alexander Technique at a Community College in Sydney. **Alternative or Complimentary Health Professional.**
Music Therapy Specialist: Expressive Arts & Music Therapy Specialist at a University in Sydney as well as having a Sydney private practice.	Twenty years experience in working with single patients and groups in movement therapy, music therapy and personal development. **Alternative or Complimentary Health Professional.**
Associate Professor Chronic Care at a Sydney University.	Self management, secondary prevention and weight management in cardiac and chronic illnesses. Cardiac and Pulmonary Rehabilitation, recovery and psychosocial adjustment. Women and heart disease. **Mainstream Health Professional.**

The questions sought to establish their background in using exercises for the elderly (over 65); their use or non-use of any interactive technologies with the elderly and the reasons; their awareness of the use of interactive technologies in either elderly persons' own homes or at institutional care facilities; the types of exercises they actually used or devised and improvements or lack of improvements in physical and mental health outcomes. After each interview one of our researchers demonstrated the four chosen Wii balance games namely: *Tightrope*, *Bubble*, *Skateboard* and *TableTilt*. In one interview an interviewee performed the exercises.

Professor Aged Care works with people suffering from Parkinson's disease so uses a *whole range of therapeutic activities like land based and aqua based strengthening exercises, balance, agility, memory, focus things.* For her patients recovering from stroke she uses *rehabilitation type exercises*. The only technology she has used is voice training [Lee Silverman Voice Therapy programme – LSVT] for recovering stroke patients. Her patients *do lots of gym work with treadmills, balance machines, getting in and out of beds, and on and off toilets*. They play manual games such as board, table and floor games to improve hand-eye coordination and arm movement to extend the range of movement and coordination. *If we actually teach somebody with Parkinson's disease to dance they can stop that frozen movement and also the shaking movement* because they are *focussing on the neurological biofeedback of the task* rather than the social situation.

Feldenkrais Practitioner recently worked with the elderly in a Community Health Centre in conjunction with their *Stepping [On]* Programme. This program is *a falls prevention programme run by a number of area health services around [NSW] that is designed to teach physical activity leaders how to incorporate falls prevention exercises into their exercise programs. They train older people about improving balance.* She had not used interactive technologies but had seen the Wii used by a colleague for artistic purposes but not with the elderly. *When I [first] started working with the elderly there was a programme developed in Melbourne called* ***Getting Grounded Gracefully****. The researcher [developed] it in conjunction with the National Research Institute on Ageing. They did a trial with the movements he devised to assess how effective they were for the elderly and it got a great, big tick [of approval]*. She believed [the exercises] increased their confidence, enabled them to feel capable to leave their homes and navigate to places they need to go to with confidence. This decreased their feelings of isolation and the fear of being alone. *[The exercises] improved their overall sense of wellbeing, because you do feel well when you move, and it improves their outlook.*

Physiotherapist has worked for the last 28 years as the physiotherapist at an Aged Care Home in the Eastern Suburbs of Sydney**.** There are about 70 men and women (both low and high care) in the facility. Physiotherapist makes *objective assessments when the patients enter the facility and a care plan is instigated for use by the rest of the nursing staff and also for the physiotherapy assistants to carry on with walking assistance or passive class exercises etc.* She has not used Interactive technologies at all either in the Aged Care home or in her business. She states that many of her patients are high care and therefore she uses passive exercises such as helping the patient perform a stretch. She uses Thera Bands for resistance training (see Figure 2).

Alexander Technique Practitioner stated that most people over the age of 65 are displaying *some* infirmities in their balance. He has had considerable experience working through the Alexander Technique with that cohort. *The oldest student I had came to me when he was 96 years of age. He was very shuffly and it took him 10 minutes to get up the stairs to my teaching room, but he found it very beneficial. He had sessions with me for 4 years and passed away 2 months before his hundredth birthday*. Feldenkrais Practitioner does not actually teach people exercises but gets them to understand how their body is organized and how to release the wrong patterns that they have utilized since they went to primary school. He is not au fait with interactive technology but stated that after receiving our phone call and emails, he googled Wii programs and also Project Natal [11] and *I thought 'Wow! This is a pretty exciting field'. I wasn't really aware it existed.*

Music Therapy Specialist attended the same interview as Alexander Technique Practitioner as they work together on a number of levels. Her expertise is in the area of movement therapy, music therapy and personal development. Her comments are related specifically the Wii games as she actually did the exercises and these will be reported in Section 4.

Associate Professor Chronic Care works in Cardiac and Pulmonary Rehabilitation *with people with a history of a cardiac events, of varying kinds including big heart attacks, or heart surgery as well as very small changes so we think of it as secondary prevention.* She does not use interactive technologies with her patients because the physiotherapy gym is nearly always hospital based and it is multipurpose. *It has treadmills, bikes, elastic bands [Thera Bands], and steps for going up and down and that's the kind of the accepted thing.* She is familiar with the Wii as she has her own and stated *the Wii is a perfect way to do home based exercise. Our biggest problem is getting people to Cardiac Rehabilitation* because they must attend the gym at the hospital.

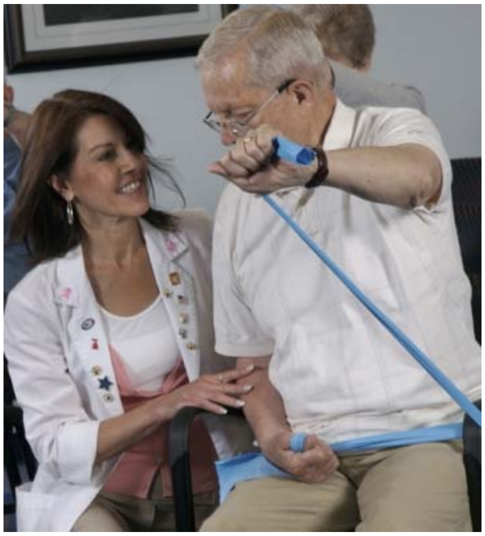


Figure 2: Exercising with Thera Bands. Source: [13]

IV. THE GAMES

As can be seen from the above our experts therefore had little experience in the use of interactive technologies in their work with the aged but **Associate Professor Chronic Care** had used the Wii herself at home. In this section we discuss their reactions to the Wii games. Table 2 sets out details of the games that were demonstrated.

The Tightrope

Professor Aged Care felt that the Tightrope game would terrify the Parkinson patients as the instructions came up far too quickly. She did not feel it would be a problem for stroke patients if they were able walk independently. **Feldenkrais Practitioner** stated the Tightrope was *good for balance and weight shifting but could be a bit confusing for the user as the picture is opposite to what they do e.g. the picture shows you one foot in front of the other; however, on the board your feet are apart*. **Physiotherapist** commented that many of her high care patients would not be able to read

TABLE II. GAMES DEMONSTRATED

Description	Main purpose
Tightrope: Your avatar stands on the balance board and tries to keep upright.	Balancing on a tightrope – use both feet side by side although the avatar appears to place one foot in front of the other. An object may appear on the tightrope and you must jump over it. If you do not reach a goal within a period of time you cannot move up a level.
Bubble: The avatar appears in a bubble; navigate down a river safely.	Again use of the balance board – both feet are used to keep steady – the avatar mirrors what the person is actually doing on the screen. The aim is to avoid bumping into river banks. The longer the walk the higher the score.
Skateboard: Avatar appears on a skateboard and must keep upright	Again use of the balance board but this time the person must occasionally push off with a foot on the ground just as if he/she were on a real skateboard. Must avoid obstacles, jump over skate jumps and improve scoring. If you do not reach a goal within a period of time you cannot move up a level.
Table Tilt: A ball appears on a Table Tilt which has holes on it – the person must tilt the Table Tilt in order to get the ball to flow down the hole.	Person stands on balance board and by shifting weight the table tilts and the ball/s move towards or away from the holes. More balls and more holes appear so it is more difficult. If you do not reach a goal within a period of time you cannot move up a level.

the instruction, however she thought the low care patients would find it *lots of fun* and she suggested they could use rollators (See Figure 3) to keep themselves steady. **Alexander Technique Practitioner** also mentioned that his patients would find this game challenging and upsetting, particularly if their avatar fell. He believed it was difficult to read and do the actions at the same time and suggested audio feedback. **Music Therapy Specialist** stated *that falling from a tightrope between two skyscrapers is anxiety-producing stimulus, which makes people tighten their necks and shoulders and sabotages their balance reflex.* She did the exercises herself and mentioned *that reading instructions spread across the screen may be confusing and less compatible with kinesthetic awareness than auditory feedback*. **Associate Professor Chronic Care** felt it was *a useful exercise because it is about balance and shifting weigh* and commented our instructor was using his quads and a bit of his core. She thought this game would appeal more to men than women.

The Bubble

Professor Aged Care was more impressed with this one stating that it would need to be much slower *so they've got a chance of success. What would probably happen is they would probably keep hitting the side the whole time - it could be frustrating, worrying. They could pick up speed over time. Start very slowly until they get used to these things.* **Feldenkrais Practitioner** indicated this game *allows movement forwards and backwards and side to side, which is a good strategy*. **Physiotherapist** thought this one was fun and could be easily used in the aged care facility or at home. She believed it could be used by people holding onto a balance frame or rollator for support. **Alexander Technique Practitioner** approved of this one also – the movement on the screen matched what the exerciser was doing – unlike the Tightrope. **Music Therapy Specialist** stated that *the side to side movements of the body correspond well with the screen movements, so this would be good for stimulating body awareness of weight distribution on the feet.* She criticized the technique for slowing down as the linking of speeding up of the image with leaning forward and slowing down with leaning back is a very unrealistic (and potentially dangerous for the elderly) since it does not correspond to human movement. *If the elderly are learning from this* game *to lean back in response to wanting to slow down it could actually lead to falling backward.* **Associate Professor Chronic Care** thought this one was *gentle and safe* but would not be useful for her cardiac rehabilitation patients as it did not raise their heart rate sufficiently. She stated it was better than Tightrope as a starting exercise.

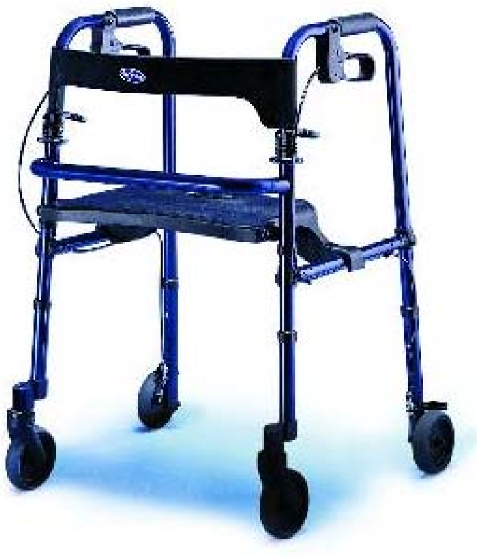


Figure 3: Rollator for Stability(has brakes). Source: [14]

Skateboard

Professor Aged Care stated that her aged patients would not relate to Skateboarding as it is outside their experience. She suggested a game with a scooter or a bicycle would have more appeal to the elderly**.** She did feel it might appeal to males who were in better health or those classified as *High Functioning*. **Feldenkrais Practitioner** decided this one was good for counterbalance and gave locomotion practice as people have to kick off from the floor from time to time as on a real skateboard. **Physiotherapis**t believed the younger and more able of her patients would find it useful and fun. She indicated patients would need good coordination and flexibility and it provide them with opportunities for *varied flexion and rotation*. **Alexander Technique Practitioner** was not keen on this as he felt you could not read and operate the skateboard at the same time. His patients could have issues with putting their foot on and off the board. **Music Therapy Practitioner** felt this game most corresponds *to realistic human movement simulation and the leaning and pushing off one needs to do on a real skateboard*. She too stated that elderly people may not have the experience or inclination to learn this skill because they would think it is for teenagers. She warned it would be necessary *to be careful when first working with this one since turning is sometimes linked to a backward lean, which may be tricky for the elderly*. **Associate Professor Chronic Care** felt this game was *more challenging and provided a better workout* but, although she liked it herself, she did not think her patients would.

Table Tilt

Professor Aged Care endorsed this game as it required *quite fine balance*, as the participant must just move the table slightly. *This one would be quite relevant because it is slow enough to be able to achieve something. You don't want to have activities that are impossible to achieve as it becomes too distressing for people and just give*. She said it was a familiar game for the elderly. **Feldenkrais Practitioner** also liked this game as it was a good exercise which worked well for elderly persons. Counterbalance is necessary. **Physiotherapist** stated that this one was good for concentration and would help her patients to stay focused. She believed it was a fun game as well. **Alexander Technique Practitioner** did not endorse this game as he believed it encouraged patients to tense up and lose balance He stated he wants people relax and loosen up**. Music Therapy Practitioner** did not comment on the Table Tilt**.**

Associate Professor Chronic Care stated that patients would need posture help on this one. She suggested having a wider walking frame around the footpad for balance assistance if the patient *started to fall a bit*. She liked this game for the elderly as it would be familiar.

V. DISCUSSION

As can be seen from the above section the use of the Wii Balance Games we selected for examination by the health professionals did not meet with universal approval. A common criticism concerned the fact that the participants would have to read the instructions while attempting to do the exercise. Many of the elderly have fading eyesight so this would be difficult for them. Some of the patients with whom our professionals worked could not read English as they were from a non-English speaking background. The use of audio instructions was suggested by some of our interviewees but again many elderly people have hearing loss so that too could give rise to problems. Training with the Wii games would be necessary because people do not like to fail or appear silly in front of their peers. **Feldenkrais Practitione**r emphasized that a lot of the movement strategies that she used also included getting up and down from the floor, *because that's what people fear, they fear falling and they fear that they can't get up. So this reduces their confidence too.* The alternative/complimentary practitioners concentrated more on every day type movements trying to get their clients to be more aware of their bodies and their sense of balance. Many of their exercises concerned getting up and down from a chair, from a bath or from the floor. They wanted their clients to feel confident in posture and balance and they could see the benefit of some of the Wii games.

It was also noted that no Interactive Technologies were currently in use in any of the areas in which the interviewees worked. They were however interested in the Wii games and thought that they could definitely be introduced into gyms and/or aged care facilities under supervision and that people would have to be introduced to the games slowly and see rewards or improvements. The mainstream professionals felt that many of the games would appeal more to men than women. They felt that the social aspect of using the Wii would be important. The interviewees were united in their opposition to the Skateboard game and commented that the patients would be attracted to games with which they were familiar e.g. Table Tilt and perhaps riding a bicycle rather than a skateboard.

VI. CONCLUSION AND FUTURE DIRECTIONS

In our ongoing investigation of the use of interactive technologies for the aged and infirm we believe that obtaining specific insights into how professionals look at balance exercises was extremely useful. We are aware now that reading the instructions whilst doing the exercises in probably not the best way for an elderly cohort. The use of a trainer to teach the elderly would be important as would the use of a rollator and/or frame to give the elderly a feeling of support once they were playing the games.

It is apparent to us that these games are more relevant to the young and we are currently investigating how to design games that are more age relevant. We also intend to investigate the new Project Natal [11] (now called Kinect [12]) when it arrives in Australia as it might prove to be more adaptable for the elderly cohort. One Kinect game called 'Your Shape-Fitness Evolved' serves as personal trainer with a variety of exercises including Yoga and could be a possibility for elderly users. We have commenced our next task of testing the games on a cohort of people over 65 years of age to elicit their opinions and to see how whether their opinions accord with the professional professionals' perspectives.

ACKNOWLEDGMENT

The authors would like to thank the six interviewees for their insights and time.

REFERENCES

[1] Jimison H., Gorman P., Woods S., Nygren P., Walker M., Norris S. and Hersh W: "Barriers and Drivers of Health Information Technology Use for the Elderly, Chronically Ill, and Underserved". Evidence Report/Technology Assessment No. 175 (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-02-0024). AHRQ Publication No. 09-E004. Rockville, MD: Agency for Healthcare Research and Quality. November 2008.

[2] Lawrence, E., Sax, C., Felix Navarro, K. and Mu Qiao: "Interactive Games to Improve Quality of Life for the Elderly: Towards Integration into a WSN Monitoring System". In Proceedings of The International Conference on eHealth", Telemedicine and Social Medicine (eTELEMED 2010), St Marteens, February 2010, pp. 10-17, doi: 10.1109/eTELEMED.2010.21.

[3] Lawrence, E., Sax, C, Felix Navarro, K.: "Improving Health Outcomes for the Elderly: An Analytic Framework". 23rd Bled eConference eTrust: Implications for the Individual, Enterprises and Society", Bled, Slovenia, June, 2010, pp. 441 – 454, doi:

[4] Fischer, M., Lim, Y.Y., Lawrence, E., Ganguli, L.K. and Kargl, F.: "ReMoteCare: Health Monitoring with Streaming Video". IEEE 7Th International Conference on mBusiness, July 2008, Barcelona, Spain, doi: 10.1109/ICMB.2008.16.

[5] Leijdekker, P., Gay, V. and Lawrence, E.: "Smart Homecare System for Health Tele-monitoring". International Conference on Digital Society, IEEE Computer Society, March 2007, pp. 3-3, doi: 10.1109/ICDS.2007.37.

[6] Chilukoti, N., Early, K., Sandhu, S., Riley-Doucet, C. and Debnath, D: "Assistive technology for promoting physical and mental exercise to delay progression of cognitive degeneration in patients with dementia", Biomedical Circuits and Systems Conference (BIOCAS 07), IEEE Computer Society, November 2007, pp.235-238, doi: 10.1109/BIOCAS.2007.4463352.

[7] Sugarman, H., Weisel-Eichler, A., Burstin, A. and Brown, R.: "Use of the Wii Fit system for the treatment of balance problems in the elderly: A feasibility study". Virtual Rehabilitation International Conference, July 2009, pp. 111-116, doi: 10.1109/ICVR.2009.5174215.

[8] Gil-Gomez, J.A., Lozano, J.A.; Alcaniz, M. and Perez, S.A.: "Nintendo Wii Balance board for balance disorders**".** Virtual Rehabilitation International Conference, July 2009, pp.213–213, doi:10.1109/ICVR.2009.5174251.

[9] IJsselsteijn, W., Nap, H.H., Poels, K. and de Kort, Y.: "Digital Game Design for Elderly Users". Proceedings of the 2007 conference on Future Play, May, 2007, pp. 17-22, doi: 2007 ISBN:978-1-59593-943-2.

[10] Strauch, R.: "An overview of the Feldenkrais Method®". Sourced from: http://www.somatic.com/articles/feldenkrais_overview.pdf.

[11] Project Natal: "You Are the Controller". Sourced 2nd May, 2010 from http://www.xbox.com/en-US/live/projectnatal/.

[12] Anonymous Kinect gets gamers off the couch, sends Wii packing,Technotes, http://www.apcmag.com, August 2010, pp. 11

[13] "Exercising with Thera Bands" . Sourced August, 2010 from: http://www.stoningtonvna.com

[14] Sourced August, 2010 from: http://www.gobilitymobility.com/Rollator.html

Comparative Analysis of the Practice of Telecom Operators in the Realization of IPTV Systems Based on ITIL V3 Recommendations for the Supplier Management Process

Anel Tanovic
Department for IT development of multimedia services
BH Telecom d.o.o. Sarajevo
Sarajevo, Bosnia and Herzegovina
anel.tanovic@bhtelecom.ba

Fahrudin Orucevic
Department of Computer Science and Informatics
University of Sarajevo, College of Electrical Engineering
Sarajevo, Bosnia and Herzegovina
forucevic@etf.unsa.ba

Abstract – **For the business of an organization to be at a high level, it is necessary to define the relationships with an external company (partner) which is going to be helpful in the complete or partial implementation of some project. The main motive for hiring a company as a partner in the process of realization is for finding a high quality solution, and saving your own human resources. The process which deals with the business relationships between your company and an external company in the IT industry, according to ITIL V3 methodology of leading IT services, is Supplier Management. The aim of this article is to describe Supplier Management in the development of IPTV systems for a Telecom Operator, through the creation of a contract between Telecom Operator and an external company which needs to implement and install IPTV systems, and also through the guidelines for performance control of a partnership company by the Telecom Operator, during the project of application and installation of the IPTV system. The result of the comparative analysis has to be a sequence of recommendations for the improvement of relationships with an external company which has implemented the IPTV system and which is responsible for the initial maintenance of the IPTV system, for the purpose of enabling high quality IPTV services to the end users of this Telecom Operator. The measuring of the implementation of recommendations from the Supplier Management process in a comparative analysis was performed as recommended by the Balanced ScoreCard method.**

Keywords - Service Management; Information Technology Infrastructure Library; Service Strategy; Service Design; Service Transition; Service Operation; Continual Service Improvement; IPTV; Supplier Management.

I. INTRODUCTION

Every company that wants to increase its level of work and business depends on the IT Service Management. If the IT processes and services are led successfully, the operation of the company will become more fortunate and successful, which can be noticed in the decrease of costs, and increase of revenues and achievement of contacts with other business partners. For the IT processes and services to be successfully led, it is necessary for the company to define a gathering of specialized organizational skills which are offered to clients in the form of a service. That set of specialized skills makes up the Service Management of a company [1], [8].

There are many standards of Service Management practice of which the most important is ITIL [1], [2]. Information Technology Infrastructure Library or ITIL represents the best environment for the practice of a company which offer IT services as their main business function. ITIL poses a tool for implementing a service which one organisation will be able to fully use with realization of the implementation of all the processes or partially use through the implementation of just a few of their processes which are considered to be helpful in developing their business results. According to version 3, ITIL has 5 life cycle stages: Service Strategy, Service Design, Service Transition, Service Operation and Continual Service Improvement [1], [2], [8].

Section II describes the basic meaning and goals of the Supplier Management process. Section III describes the criteria that the Telecom Operator needs to set for the IPTV consultant during the process of selecting a consultant who needs to implement and install the IPTV system. Section IV lists the steps that need to be taken by the Telecom Operator and the IPTV consultant for the implementation of IPTV systems and defines five levels for the maintenance of the IPTV system by the consultant after the system is available for production. Section V is the central part of the paper, which presents a comparative analysis between the actual ITIL V3 recommendations Supplier Management processes and implemented ITIL V3 recommendations Supplier Management in the actual IPTV system, where BH Telecom's IPTV system is taken as the reference model. The conclusion of the contribution should give a result of applications of the ITIL V3 recommendations for Supplier Management in BH Telecom's IPTV system and give suggestions for improving recommendations that are poorly implemented.

ISBN:978-1-61208-114-4

The last chapter of the article includes an analysis of the implementation of Supplier Management through the IPTV system of BH Telecom (the leading Telecom Operator in Bosnia and Herzegovina) with the recommendation of implementation, according to ITIL V3 standards, where throughout 6 main recommendations of Supplier Management according to ITIL V3 standards a comparative analysis has been performed regarding the application of ITIL V3 recommendations related to Supplier Management in the IPTV system of BH Telecom.

This contribution represents a continuation of scientific research and application of ITIL V3 standards which is lead in BH Telecom in order to adapt all of the systems under this standard with regard to previously published contributions in this field, a contribution entitled "Implementation of the Information System of the Telecom Operators Using the ITIL V3 Service Methodology for the Service Design Phase" which was published during the SERVICE COMPUTATION 2010 conference.

II. SUPPLIER MANAGEMENT

Supplier Management process ensures that the external company and services, that they provide, can support the goals of the IT services and business expectations of a company. The aim of this process is to emphasize the importance of working with partner companies, and to provide guidance on how the business can best be directed towards the business benefit of the company by establishing a contract with another company that is used for complete or partial implementation of one or more projects which are of primary importance for the same company [9], [10], [11], [12], [18].

The main goals of Supplier Management processes are to [2], [4], [18]:

- Maintain the value of money from the partner companies and the contracts with them.
- Manage relationships with partner firms.
- Manage the products of partner firms.
- Categorize partner companies according to the quality of the products they offer.
- Examine, renew and terminate agreements with partner companies.
- Manage the performance of partner companies.
- Implement services and plans for improving the partner firms.

When contracting business with partner companies, it is recommended to make a fully formal agreement with the partner company, with clearly emphasised and documented responsibilities and goals which the contract bears during certain phases in its life cycle, starting by defining business needs all the way to terminating agreements. Key activities of Supplier Management are [4], [8], [10], [16]:

- Identifying business needs and preparation for business activities.
- Assigning new agreements with partner companies.
- Managing the performances of partner companies and contracts with them.
- Defining the final requests with the existing partner companies.
- Categorizing partner companies and contracts with them.
- Evaluating and assigning new contracts with partner companies.

III. REALIZATION OF THE CONTRACT BETWEEN TELECOM OPERATOR AND EXTERNAL COMPANY WITH THE PURPOSE OF IMPLEMENTING AND INSTALING IPTV SYSTEM

The first step in the realization of IPTV systems of Telecom Operator is the selection of a partner company that will design, implement and install IPTV system and also maintain the same for a certain period [1], [16]. Agreements with external companies-consulate (contracts) are accomplished after a public process with which a company is chosen for projecting the installation of the IPTV systems. The selection of IT consultants is performed through two phases: the prequalification phase and the final phase where the IPTV consultant is chosen. In order for the candidates to pass the phase of prequalification’s it is necessary for them:

a) To not have any legal impediment in their participation in the contest for the choosing the best IT consultant.
b) To have the right to perform professional services and to be registered in the proper professional registers.
c) To have economical and financial condition to realize a successful implementation of the contract.
d) That their technical and professional capacity ensures successful realization of the contract.

Regarding the technical and professional abilities, candidates must fulfil the following minimum requirements which must be confirmed through certified documents:

a) Employed at least 30 certified IPTV consultants.
b) A minimum of three references to the implementation of IPTV systems in the telecom industry.
c) Employed at least 20 certified IPTV consultants that will participate in the implementation of the project provided that at least 10 consultants provide a 12 month software support in the maintenance of the system.
d) The existence of at least 10 IPTV consultants who have taken part in at least 3 projects of design and installation of IPTV systems.
e) Possession of own hardware and software infrastructures for the start of the mentioned project

f) The offered IPTV software has to enable the end users functions such as: watching live TV channels, browsing, the electronic program guide (EPG), recording shows, shopping and an overview of all genres of movies, possibility of integration with VOIP platform, internet access, listening to radio channels, access to live games and encrypt live TV channels and movies.

g) The offered IPTV software has to provide the expansion of the number of IPTV users in any number, expanding the number of movies to an unlimited number and the expansion of live TV an unlimited number.

h) The offered IPTV solution must be such that it can be integrated with any type of Set Top Box.

i) The offered IPTV solution must have a monitoring system that will monitor the activity of the whole system, live TV channels in certain regions of the Telecom Operator and the work of each Set Top Box that is connected to the IPTV system.

j) The offered IPTV solution must be compatible with the existing network architecture of the Telecom Operator.

The second phase is the phase of final selection for the IPTV consultant between the consultants who met legal, economical, technical and professional skills. There are three main parameters by which an IT consultant is selected from all other consultants: lowest price (with a percentage share of 50%), the quality of offered the solution (with a percentage share of 40%) and a deadline for the shipment of the offered solution (with a percentage share of 10%). The contract must be assigned to the consultant that has submitted the top rated acceptable offer. In the event that two or more firms had the same rating for acceptability in the final assessment, the IPTV consultant with the shortest deadline time for the delivery of the offered solutions is chosen.

The Telecom Operator can terminate the transfer from the prequalification stage into in the final selection phase of the IPTV consultant if the number of providers who have made it into the prequalification stage is less than 3 and in this case can again call a public process for the choice of IPTV consultant for the implementation and installation of IPTV solutions.

IV. MANAGING THE PERFORMANSE OF THE PARTNER COMPANY BY THE TELECOM OPERATOR IN REALIZATION OF THE IPTV SYSTEM

With the levels of maintenance of agreement between the Telecom Operator and the partner company it is necessary to define the levels of problems, define malfunctions, response time and eliminating problems which the partner firm needs to solve according to the contract [16] (Table I).

The first level represents the level of noticing problems, second level represents categorization of problems and setting priorities for solving the same, level three is solving the major problems that are essential for running the system and that must be solved in a matter of hours, level four is solving problems that need to be addressed over longer periods and are not relevant to the operation of the whole system, and the fifth level consists of a team of software engineers who are responsible for changing the functionality of the whole IPTV system, which usually takes place over a period of months or years [4], [16].

According to the signed contract the total time of the project must be 18 months, provided that the time for the design and installation of the system by the partner company is 6 months, and the time for completing the obligations as well as offering software support provided by the partner company is 12 months. After the end of providing software support, after 12 months, Telecom Operator and the partner company can resign the contract again on the software maintenance system which again lasts 12 months. After this period of time, the Telecom Operator should be able to maintain with its own workforce its own IPTV system and to solve any incident or problems that may arise in it.

After defining the levels and types of problems, it is possible to define the workflow for solving the problems for which, according to contract, the partner company is responsible. Table I shows the types of problems with periods that the partner company according to contract has to solve and targeted time for troubleshooting.

TABLE I. TYPES OF PROBLEMS WITH PERIODS THAT THE PARTNER COMPANY ACCORDING TO CONTRACT HAS TO SOLVE

The level of proble ms	**Definition of the problem**	**The response time for problem solving**	**Targeted time for troubleshooting**
1.	A complete system crash - in the event that the system does not function and there is no alternative to establishing a temporary system operation.	1 hour after receiving a verbal notification of the problem	24 hours after receiving a verbal notification of the problem
2.	The critical problem: if a crucial part of the system does not function and thus prevents the continuous operation of the basic functions of the system (there is no alternative solution for the temporary establishment of functionality).	1 hour after receiving a verbal notification of the problem	24 hours after receiving a verbal notification of the problem
3.	Uncritical problem - part of the system or component that is not critical is not functioning, but the basic functions of the system are operating.	24 hours after receiving notification of the problem	7 work days after receiving notice of the problem

The level of proble ms	**Definition of the problem**	**The response time for problem solving**	**Targeted time for troubleshooting**
4.	Minor anomalies in the system which do not affect the operation or basic functionality of the IPTV system.	24 hours after receiving notification of the problem	7 work days after receiving notice of the problem
5.	The Telecom Operator's request for additional functions in the IPTV system.	By agreement with the management of the partner firm	In agreement with the management of the Telecom Operator

The company partner is the carrier of liability for the following stages of system development during the design and installation of the IPTV system:

1. The construction of MiddleWare (MW) system that has to monitor the whole system and consists of two servers that work in parallel mode.
2. Design and implementation Oracle data base into Real Application Clusters (RAC) environment and its correlation with the MiddleWare system.
3. The design and implementation of cryptosystems which implicit the formation of two independent clusters: the first cluster consists of Real Time Encryption Servers (RTES) which are used for encrypting live TV channels and the second cluster which consists of Verimatrix Servers (VCAS) which is used for encrypting movie with Video On Demand (VOD) servers and connecting clusters with the MiddleWare system.
4. The design and implementation of video systems for storing recordings and movies which consists from 2 Video On Demand (VOD) servers and their connection with the MiddleWare system.
5. The design and implementation of a Load Balancing (LB) system which is responsible for transferring the entire work of a certain system to another server in the same system in the event that one of the server from MW, RAC, VOD, RTES, VCAS fails.
6. The design and implementation of a monitoring system that needs to monitor the work of live TV channels in all regions of the Telecom Operator, monitoring the work of all the Set Tops Boxes that are registered on the system, to inspect the whole IPTV system, and to display an alarm in the event of a malfunction as well as connecting a monitoring system with the MiddleWare system.
7. The design and implementation of a Network Video Server system which is responsible for emitting video signals directly from the IPTV system through to the net and its connections with MiddleWare system.
8. The design and implementation of a an initial Headend system which consists of receivers and encoders for emitting live TV channels and realising into implementation 50 live TV test channels.

According to the contract the Telecom Operator is responsible for the following activities:

1. Planning, system analysis and specifications of requirements, with all the necessary activities for the development of the project (final project and term plan, defining the framework of the system, defining business processes and the specification requirements that the system must satisfy, as well as the preferences of the technical architecture of the system for hardware-systematic platform needed for the development of the project).
2. Testing the implemented IPTV system by the partner company.
3. Defining the IPTV package in combination with VOIP and Internet services that need to be offered to end users.
4. The integration of IPTV systems with central information system Telecom Operator in order to add users, terminate users, delete users, change system setting, add Set Top Boxes to the system, delete Set Top Boxes from the system, add channel packets, change channel packets, erase channel packets for a user.
5. The construction of a charging system for the IPTV system depending on the IPTV package purchased and the formation of categories for charging movies.
6. The construction of user manuals with a detailed description of the use of IPTV services.
7. Training users on helpdesk that will work in direct contact with the end users.
8. The expansion of the initial Headend system which consists of receivers and encoders that broadcast live TV channels and realizing into implementation 80 production live TV channels at the end of the mentioned project.

V. THE ANALYSIS OF IMPLEMENTATION OF SUPPLIER MANAGEMENT THROUGH THE IPTV SYSTEM OF BH TELECOM WITH RECOMMENDATIONS OF IMPLEMENTATION OF THE SAME ACCORDING TO ITIL V3

Table II provides a description of ITIL V3 suggestions for Supplier Management, description of realizing the recommendations through the IPTV system of the leading Telecom Operator in Bosnia and Herzegovina taking into consideration the examination of realized contract between BH Telecom and the partner company for the implementation and installation of the IPTV system, performance management of BH Telecom over the performance of the partner company in realization and maintenance of the IPTV system, and an overview of the percentages of implementation of ITIL V3 suggestion for Supplier Management process in the IPTV system of BH Telecom. For the comparative analysis six basic ITIL V3 recommendations for the Supplier Management have been chosen. Measurements of the implementation of ITIL V3 recommendations for Supplier Management have been done according to the recommendation of standardized methods for measuring the realization of ITIL V3 recommendations which is called the Balanced Scorecard [7]. Balanced Scorecard, for each recommendation finds its relation to: the end user, internet processes, finances and growth potential of the system based on all four parameters gives the final results for implementation of one of the recommendations in one particular system.

TABLE II. DESCRIPTION OF REALIZATION AND PROCENTAGE OF REALIZATION OF THE ITIL V3 RECOMMENDATIONS FOR SUPPLIER MANAGEMENT IN THE IPTV SYSTEM OF BH TELECOM

ITIL V3 recommendations for Supplier Management	**Realization of ITIL V3 recommendations for Supplier Management in BH Telecom's IPTV system**	**Percentage of ITIL V3 implementation from recommendations for Supplier Management in BH Telecom's IPTV system**
Recommendation 1: **Identification of business requirements and preparation of business**	In order to achieve financial gains and expansion of its services, BH Telecom has decided to introduce IPTV service by engaging foreign firms that should have a very high reference, the same as the previous 3 realized project in other Telecom Operators and own at least 20 certified IPTV consultants.	This recommendation has been fully realized (100%)
Recommendation 2: **Assignment of new agreements with partner companies**	There are clearly defined criteria for the prequalification stage and the final selection phase of IPTV consultants as the minimum number of consultants who must be at the stage of final selection for the entire procedure to successfully be completed. However, there is no defined criterion for a new business partner of choice in case of the termination of a partnership with a firm partner for the duration of the project. This conclusion is made after examining all the agreements that BH Telecom has with all his suppliers.	This recommendation has been partially realized (50%)
Recommendation 3: **Performance management of partner companies and contracts with them**	There is a complete performance management vendor, which is reflected through the work flow of support for which a firm partner is in charge and the types of problem solving in time by a firm partner under the contract must be resolved on the side where the firm partners in problem solving include all categories of technical staff from technicians to system engineers.	This recommendation has been fully realized (100%)
Recommendation 4: **Defining the end requests with an existing firm partner**	BH Telecom has not defined clear criteria by which to carry out re-election of the firm partners in the event of failure to fulfil all its obligations during the project implementation and installation of IPTV solutions as well as whether in this case the re-election of the firm partners take place through the public process through the stages of prequalification and the final selection of the IPTV consultant.	This recommendation is not implemented (0%)
Recommendation 5: **Categorization of partner companies and contracts with them**	There is no clearly defined procedure for the categorization of suppliers carried out in two phases: the phase in which the prequalification choose only those firms which meet the legal, financial, technical or professional abilities and the final stage of selecting a consultant, where the IPTV consultant will install and deploy IPTV solution based on criteria that include: the lowest price with the percentage share of 50% in the	This recommendation is fully applied (100%)

ITIL V3 recommendations for Supplier Management	**Realization of ITIL V3 recommendations for Supplier Management in BH Telecom's IPTV system**	**Percentage of ITIL V3 implementation from recommendations for Supplier Management in BH Telecom's IPTV system**
	selection, the quality of their solutions with the percentage share of 40% in the selection and the time of delivery with the percentage share of 10% in the selection of IPTV consultant.	
Recommendation 6: **Evaluation and mediation of new contracts with partner companies**	There is a clearly defined procedure of selection of other firms that may participate in the implementation of solutions where a firm aside from the IPTV consultant can come alone to implement and install a system database, the crypto system, video system, Load Balancing system, monitoring system, Network Video system or Headend system. Choosing this firm can happen at a later realization of the project when the contract expires with the initial IPTV consultant.	This recommendation is fully implemented (100%)

The comparative analysis showed that Supplier Management in the IPTV system of BH Telecom has been completely applied in four ITIL V3 recommendations (the identification of business needs and the preparation of businesses, controlling the performance of partner companies and contracts with them, categorizing partner companies and the contract with them, evaluation and intermediation of new contracts with partner firms), partially in one ITIL V3 recommendations (intermediation of new contracts with partner firms), and that it isn’t at all applied in one ITIL V3 recommendation (defining the end request with the existing partner company).

If you find the arithmetic mean of all percentages of recommendations from Supplier Management you come to the conclusion that 75% of the recommendations have been successfully applied in BH Telecom's IPTV system (Figure 1). Figure 1 shows the ratio of successfully and unsuccessfully implemented ITIL V3 recommendations for Supplier Management in BH Telecom's IPTV system.

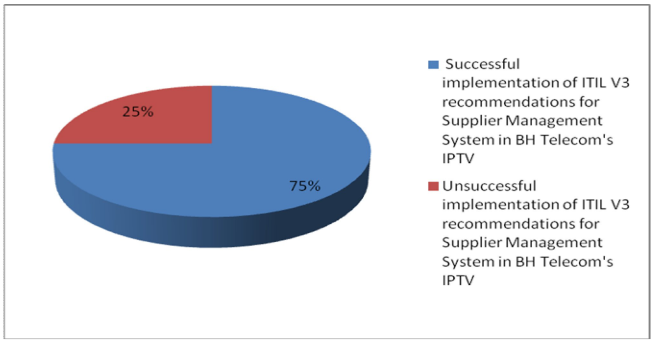


Figure 1. Ratio of successful and unsuccessful part of implementation of ITIL V3 recommendations for BH Telecom's IPTV system

VI. Conclusion

There are two main criterions that have to be taken into consideration upon defining Supplier Management during the implementation of IPTV system of Telecom Operators: criteria for granting contracts with partner companies for the implementation and installation of an IPTV system and criterions for controlling the performance of the partner company by the Telecom Operator [4], [12], [13], [16]. The criterion for granting contracts to partner firms for the implementation and installation of the IPTV system has to satisfy all legal, economical, technical and professional conditions that are set on the public process of choosing the partner company. Granting the contract to the partner company is usually managed in two phases: the phase of prequalification and the phase of final selection of the partner company. The criterion for controlling the performance of the partner company must define the stages of maintenance of the IPTV system by the partner company and phases of the development of the system that need to be defined by the Telecom Operator and the partner company.

The comparative analysis has shown the applicability of ITIL V3 recommendations for Supplier Management in the IPTV system of the Telecom Operator is 75%. For complete realization of all ITIL V3 recommendations it is necessary to define the criteria of choice of a new partner company for the duration of the project as well as the method of process after which the new partner company is chosen and defining clear criteria in the event that a partnership is terminated with the existing company. The criteria of selecting a new partner company in the event of a termination of the contract with the current partner company has to offer a contract for accepting the best positioned team that on the public process didn’t acquire the contract for running the business of implementation, installation and maintenance of the IPTV system. Such a procedure would last only a few days and this way in a very short period a new partner company would be chosen. In the event that none of the companies accept the contract then the public process is reopened, which according to the legislation lasts a few months.

Further investigations in this field should give a similar comparative analysis of the process of Change Management, Incident Management and Problem Management which occurred six months 6 months after the initial installation of the system which would include hardware and software changes of the IPTV system by the Telecom Operator and also conflicts with real incidents generated by the end users and real problems that may occur because of the interruption of one of the essential components of the IPTV system [10], [16], [18].

ACKNOWLEDGEMENT

The authors wish to thank the management of BH Telecom and Smart Com for the advice during the implementation of the IPTV system in BH Telecom as the leading Telecom Operator in Bosnia and Herzegovina, and expert advice from Netvisor, Motorola, Kasenna and Verimatrix based on the practical experience in the implementation of individual components of the IPTV system in a large number of Telecom Operators.

REFERENCES

[1] J. van Bon, A. de Jong, A. Kolthof, M.Pieper, R. Tjassing, A. van der Veen, and T. Verheijen, ”Foundations of IT Service Management Based on ITIL V3”, The Office of Government Commerce, September 2007.

[2] J. van Bon, A. de Jong, A. Kolthof, M.Pieper, R. Tjassing, A. van der Veen, and T. Verheijen, ”Service Design based on ITIL V3”, The Office of Government Commerce, June 2008.

[3] S.Taylor, M. Iqbal, and M. Nieves, “ITIL Version 3 Service Strategy”, The Office of Government Commerce, May 2007.

[4] S.Taylor, V. Lloyd, and C. Rudd, “ITIL Version 3 Service Design”, The Office of Government Commerce, May 2007.

[5] S.Taylor, S. Lacy, and I. Macfarlane, “ITIL Version 3 Service Transition”, The Office of Government Commerce, May 2007.

[6] S.Taylor, D. Cannon, and D. Wheeldon, “ITIL Version 3 Service Operation”, The Office of Government Commerce, May 2007.

[7] S.Taylor, G.Case, and G.Spalding, “ITIL Version 3 Continual Service Improvement”, The Office of Government Commerce, May 2007.

[8] P. Brooks, J. van Bon, and T. Verheijen, “Metrics for IT Service Management”, The Office of Government Commerce, April 2006.

[9] T. Metller and P. Rohner, “Supplier Relationship Management: A Case Study in the Context of Health Care”, Journal of Theoretical and Applied Electronic Commerce Research, Universidad de Talca – Chile, vol. 4, December 2009, pp. 58-71.

[10] K. Nakashima and S. M. Gupta, “Performance evaluation of a supplier management system with stochastic variability”, International Journal of Manufacturing Technology and Management, vol. 5, July 2003, pp. 28-37.

[11] J.M.Proth, G. Mauroy, Y. Wardi, C. Chu, and X. L. Xie, “Supply management for cost minimization in assembly systems with random component yield times”, Journal of Intelligent Manufacturing, vol. 8, September 1997, pp. 385-403.

[12] H. Gurnani, R. Akella, and J. Lehoczky, ”Supply management in assembly systems with random yield and random demand”, Journal of IIE Transactions, vol. 32, August 2000, pp. 701-714.

[13] D. A. Riggs and S. L. Robbins, “The Executive’s Guide to Supply Management Strategies: Building Supply Chain Thinking Into All Business Processes”, Amacom, March 2009.

[14] S. Graupner, S. Basu, and S. Singhal, „Collaboration environment for ITIL“, Integrated Network Management-Workshops, 2009. IM ’09. IFIP/IEEE International Symposium, vol. 5, June 2009, pp. 44-47.

[15] M. Aazadnia and M. Fasanghari, “Improving the Information Technology Service Management with Six Sigma”, IJCSNS International Journal of Computer Science and Network Security, vol. 8, March 2008, pp. 144-150.

[16] A. Tanovic and F. Orucevic, “Implementation of the Information System of the Telecom Operators Using the ITIL V3 Methodology for the Service Design Phase”, in Proceedings of the 2nd International Conferences on Advanced Service Computing, Lisbon, Portugal, November 2010.

[17] W. Hommel and S. Knittl, “An Access Control Solution For The Inter-Organizational Use Of ITIL Federated Configuration Managemen Databases”, Published in Proceedings of the 15 Annual Workshop of HP Software University Association (HP-SUA), May 2008.

[18] G. Blokdijk and I. Menken, “Supplier Management Best Practise Handbook: Evaulating, Sourcing, Managing and Delivering Supplier Excellence In Relationships. Quality and Costs”, Emereo Pty Ltd, August 2008.

An Approach to Service Deployment to the Service Cloud

Juha Puttonen, Andrei Lobov, José L. Martinez Lastra
Department of Production Engineering
Tampere University of Technology
Tampere, Finland
{juha.puttonen,andrei.lobov,jose.lastra}@tut.fi

Abstract—**Computing clouds facilitate rapid and effortless resource allocation. In particular, Infrastructure-as-a-Service clouds allow clients to dynamically lease virtual machines that behave similarly to physical servers. However, executing an application by directly using computing cloud resources is complicated and typically involves similar steps as installing and executing an application on a physical machine. Moreover, starting numerous application instances on a single virtual machine may result in poor performance. Thus, we propose developing a web service that acts as a mediator between the leased cloud resources and the cloud users and facilitates the use of the resources. When the mediator web service is used, an application can be started in a computing cloud effortlessly by invoking simple web service operations. Furthermore, in the case of several applications, the workload can automatically be distributed between several virtual machines, resulting in higher performance.**

Keywords- cloud computing; web services

I. Introduction

Computing processes require hardware resources, such as processing power and data storage capacity. Traditionally, the resources have existed on physical server machines. Hence, organizations have had to invest in the purchase of the hardware as well as allocate resources in installing and maintaining the systems. Moreover, the need for computing resources tends to considerably fluctuate, causing the expensive systems to be frequently idle. Adjusting the computing resources to match the current needs is typically expensive using traditional methods. Cloud computing provides a solution to the problem by allowing organizations to lease computing resources and only pay for the amount that they actually use [1].

A. Previous Work

There are different types of cloud computing. In Infrastructure-as-a-Service, IaaS, the leased resources are complete virtualised systems [2]. More specifically, the resource units leased from IaaS clouds are virtual machines [3], which behave identically to actual servers connected to the internet. However, they are created through virtualisation from actual servers. Other types of cloud computing include Software-as-a-Service, SaaS, and Platform-as-a-Service, PaaS. In SaaS, software vendors make their applications accessible over the Internet, while in PaaS the cloud systems provide platforms that allow software vendors to implement their applications. Then, the end users can access the applications over the Internet similarly to SaaS [4].

Public IaaS clouds are typically commercial enterprises from which virtual machines can be leased at certain prices [3]. The Amazon Elastic Compute Cloud, Amazon EC2 [5], is a notable example of public IaaS clouds. However, organizations can also create private clouds that are used internally and non-commercially [3]. The main purpose of private clouds is to share existing resources, rather than provide additional resources. On the other hand, private clouds may also use the resources of public clouds, and the combinations are called hybrid clouds [3].

Cloud computing toolkits, such as Eucalyptus [6] allow the creation of private clouds. While there are no standard computing cloud interfaces, the private clouds created using the Eucalyptus software framework conform to the Amazon EC2 cloud interface and can be used with the same client tools [6].

For example, companies consisting of several departments can benefit from the effortless resource allocation enabled by private IaaS clouds. If each department were allocated physical servers, those servers would be idle for considerable periods of time. While reallocating physical servers to different departments might cause considerable amount of additional work, it is straightforward to dynamically start virtual machines and attach virtual storage volumes to the virtual machines with all the necessary software and data.

In our work on semantic web services orchestration [7], we have proposed a set of web services providing a web service orchestration framework. The orchestrated web services can be hosted in resource-constrained embedded devices. In the orchestration framework, the performance issues related to memory and CPU resources can be overcome by outsourcing some of the resource-demanding functions to the cloud. Considering service oriented architecture (SOA), it would be natural for the applications deployed in the cloud to provide web service interfaces. Fortunately, computing clouds can facilitate the dynamic deployment of web services, such as those forming our web service orchestration framework. Some web services may be needed only for a limited time, after which the computing resources reserved by them should be released. Moreover, deploying the services on physical server machines might require considerable effort in configuring and installing the hardware and software. The use of cloud computing is a more feasible approach, as it allows the dynamic creation of virtual machines for hosting the web services, thus reducing the number of actual computer systems required and the amount of idle resources.

B. *Problem Formulation*

While IaaS clouds offer more flexibility and portability than SaaS and PaaS solutions, the workload in starting an application using an IaaS cloud is considerable [8]. Although many IaaS clouds support similar interfaces, starting an application in an IaaS computing cloud may be difficult due to the low-level nature of IaaS clouds. Indeed, a client must first select the appropriate virtual machine images to use, and communication with a virtual machine instance is typically performed by logging in to the instance with a terminal program. Hence, automating the use of the virtual machine in, for example, starting new applications, would be difficult. Therefore, we propose developing a web service that is deployed on a virtual machine in an IaaS cloud and facilitates the use of the leased computing resources. The web service interface provides operations that allow starting and terminating applications.

C. *Outline of the Paper*

The structure of this paper is as follows. Section II introduces the proposed new approach on IaaS cloud resource utilisation. Section III demonstrates the application of the approach and evaluates its performance aspects. Finally, Section IV contains conclusions and issues to be targeted in future research.

II. Main Results

We have developed a web service that facilitates using computing cloud resources. We have named the web service Cloud Gateway because it acts as a mediator between a computing cloud and the cloud users. Specifically, one instance of the Cloud Gateway service is started on each virtual machine leased from an IaaS cloud. The service instances then enable a user to effortlessly execute applications on the virtual machines. Moreover, the Cloud Gateway services can form networks spanning several virtual machines that may reside in separate computing clouds. Thus, when a Cloud Gateway is low on computing resources, it can delegate a request for starting a new application to another Cloud Gateway instance hosted by a less burdened virtual machine.

A. *Adding and Executing Applications*

Cloud Gateway provides operations for adding and removing applications to and from its application library as well as starting and terminating instances of the applications. Cloud Gateway assigns a unique string identifier to each application and to each started application instance. The most important operations in the Cloud Gateway service interface are described in Table I.

We have particularly considered the case where each application, when executed, creates and starts a web service compliant with the DPWS specification [9]. Thus, in the sequel, these types of applications are called server applications.

To facilitate the effortless transfer and execution of the server applications, they must be packaged into executable Java archive (JAR) files. Hence, Cloud Gateway can download the applications as single files. Furthermore, the applications can be executed on any platform that has a sufficiently new Java runtime environment installed.

An application can be executed by invoking the *StartApplication* operation and passing the application identifier as an input. Optionally, a list of command-line arguments may be specified to override the default arguments. As a response, *StartApplication* returns the identifier assigned to the new application instance or 'FAILURE' if starting the application failed.

Command-line arguments may contain keywords that Cloud Gateway expands before executing the corresponding application. Keywords are identified by enclosing them between '$#' and '#$'. For example, Cloud Gateway replaces each occurrence of the string '$#HOST#$' with the host machine network address.

A typical server application needs at least several seconds to deploy a set of web services. Web services compliant to the WS-Discovery specification [10] send *Hello* messages when they enter a network. Hence, Cloud Gateway listens to *Hello* messages originating from the host machine. Whenever Cloud Gateway receives such a message, it considers sends a *ServiceStarted* notification to all subscribed clients. The notifications allow the clients to determine server application start-up times.

Cloud Gateway allows starting multiple instances of each application. A running application can be terminated by executing the *TerminateApplication* operation. Since Cloud Gateway is able to terminate applications only in a forcible manner, the terminated applications must prepare for the abrupt termination of the underlying Java virtual machine and perform the necessary activities at such an event. For example, DPWS-compliant web services should broadcast WS-Discovery Bye messages when leaving a network. The Bye messages allow clients to automatically detect when the web services become unavailable.

B. *Resource Consumption*

Because the amount of applications that can be started with Cloud Gateway depend on the amount of hardware resources available to the virtual machine, Cloud Gateway must use some metrics for determining the amount of free resources. Furthermore, it must compare the determined resource levels to threshold values indicating the maximum allowed resource utilisation. Cloud Gateway accepts a request to start an application only if the determined resource utilisation levels are below the maximum allowed levels.

The metrics that most clearly define the resource utilisation of a virtual machine are the random access memory (RAM) and central processing unit (CPU) usage. In Linux systems, the percentage of RAM used can be measured fairly effortlessly by examining the contents of the virtual */proc* file system. The CPU usage level is more problematic to determine, but it can be derived from the system load average, which can also be determined from the

/proc file system. The load average represents the number of processes that are either in execution or queuing for CPU time. Hence, the higher the value, the more burdened the CPU is. If the load average is equal to the number of CPUs, CPU utilisation is optimal [11]. To calculate a value for the CPU utilisation level, Cloud Gateway divides the system load average with the number of CPUs.

If either the determined RAM or CPU usage value is higher than the corresponding threshold value, Cloud Gateway rejects any requests to start a new application. The threshold values can be specified by invoking the *SetThreshold* operation.

C. Cloud Gateway Networks

The system resources of a virtual machine will inevitably be exhausted if several application instances are executed on the machine. Therefore, Cloud Gateways residing on separate machines can form networks to balance the load between several machines. For this purpose, the Cloud Gateway service interface includes the operations *RegisterCloudGateway* and *DeregisterCloudGateway*, which allow registering and deregistering partner Cloud Gateways that will be used in workload balancing. The *StartApplicationInNet* operation will execute the application locally on the host machine only if the resource utilisation is within allowed boundaries. Otherwise, the *StartApplicationInNet* operation is recursively invoked on the partner Cloud Gateways to find one that is able to service the request. Similarly, the *SetThresholdInNet* is a recursive version of the *SetThreshold* operation.

The sequence diagram in Figure 1 represents a typical use scenario of Cloud Gateway. The *Client* object in Figure 1 can be an autonomous software agent or a software tool operated by an end user. In the beginning of the example sequence, the client registers *Cloud Gateway 2* is registered to *Cloud Gateway 1* to form a Cloud Gateway network. Then, the client registers a new server application to *Cloud Gateway 1*. Once *Cloud Gateway 1* has downloaded the application, the client executes it in the cloud by invoking the *StartApplicationInNet* operation. Because *Cloud Gateway 1* is low on computing resources, it delegates the request to *Cloud Gateway* 2, which then executes the application, effectively deploying a new web service. The response to the original *StartApplicationInNet* request includes the endpoint URI of the selected Cloud Gateway instance. Finally, the client requests *Cloud Gateway 2* to terminate the server application to release the computing resources for future use.

If a Cloud Gateway selects another service instance in the network to execute an application, it must first ensure that the other instance possesses a copy of the application and obtain the application identifier by invoking its *AddApplication* operation. This is illustrated by point 10 in the sequence diagram.

III. An Application Example

We have tested our approach both with a private cloud created using the Eucalyptus [6] software framework and with the Amazon EC2. This section will first present the experiment setup, and then describe the test results.

A. The Experiment Setup

We have set up a private cloud consisting of only one computing cluster composed of a single desktop running a

TABLE I. The Cloud Gateway service includes operations for managing the available applications as well as executing and terminating them.

Operation	Inputs	Outputs
AddApplication	**location** – the URL from which the JAR file can be read **parameters** – the default command-line arguments	The identifier assigned to the application or 'FAILURE' if reading the JAR file from the specified URL fails.
RemoveApplication	**id** – the application identifier	'SUCCESS' or 'FAILURE' if no application with the specified identifier exists, or if a running instance of the application exists.
StartApplicationInNet	**id** – the identifier of the application to start **parameters** – the list of command-line arguments, if empty, the default arguments will be used	The identifier assigned to the new application instance or 'FAILURE' if starting the application failed. The endpoint URI of the Cloud Gateway that started the application.
TerminateApplication	**id** – the identifier of the application instance to terminate	'SUCCESS' if the application was running, otherwise 'FAILURE'.
ListApplications	**-**	The list of uploaded applications. The identifier, JAR file name and default arguments are listed for each application.
ListAll	**-**	The list of all application instances. The instance identifier, application identifier, command-line arguments and state (running or terminated) are listed for each instance.
SetThresholdInNet	**MemoryThreshold** – a floating point value between 0 and 1 **CPUThreshold** – a non-negative floating-point value	'FAILURE' if the threshold values are outside the allowed ranges, otherwise 'SUCCESS'.
RegisterCloudGateway	**URI** – the endpoint URI of the Cloud Gateway instance to register	'FAILURE' if the Cloud Gateway service had already been registered, otherwise 'SUCCESS'.
DeregisterCloudGateway	**URI** – the endpoint URI of the Cloud Gateway instance to deregister	'SUCCESS' if the Cloud Gateway service had been registered, otherwise 'FAILURE'.
GetResourceUsage	-	Numeric values indicating the amount of free memory, total memory, the number of CPUs, the system load average as well as the current memory and CPU utilisation thresholds.

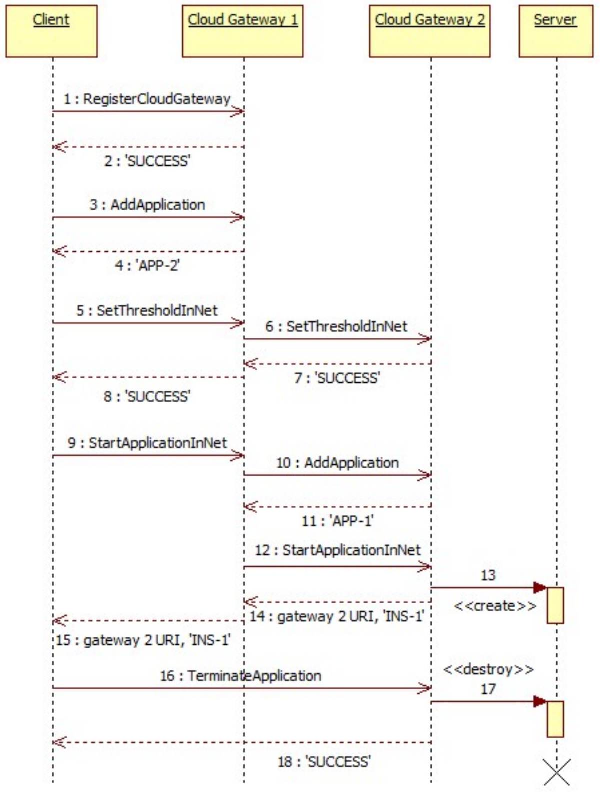


Figure 1. A typical use scenario of Cloud Gateway includes starting a web service and terminating it after use to conserve resources.

Linux operating system and Eucalyptus version 1.6.1. The restricted computing cloud limits, for example, the number of virtual machines that may be created; we have experimented with a maximum of two parallel virtual machine instances. However, even such a limited setting suffices for testing the proposed approach.

We have modified a virtual machine image so that it includes all the necessary software components for starting the Cloud Gateway server application. To upload the image to the cloud and to create virtual machines from it, we have used the Euca2ools command line utilities.

For interacting with the Cloud Gateway services, we have used our own application called Service Explorer. It includes a simple graphical user interface that allows, for example, inspecting web services and invoking their operations. In our experiments, we have executed Service Explorer on a laptop connected to the same local network as the desktop hosting the private computing cloud. Thus, Service Explorer is able to automatically detect the web services started on the virtual machines.

Each virtual machine executes a separate copy of the Cloud Gateway server application. We have deployed only two virtual machine instances in the private cloud. The experiment topology is depicted in Figure 2.

B. *Performance Measurement*

To measure the performance of the Cloud Gateway service, we have developed a test application with a simple user interface. The test application assumes the place of Service Explorer in Figure 2. The purpose of the test application is to measure the time required for deploying several independent web services in a computing cloud. The test application first invokes the *AddApplication* method to register an application and then sequentially invokes the *StartApplicationInNet* operation to execute the application a number of times specified by the user. After each *StartApplicationInNet* request, the test application waits for the Cloud Gateway to send a *ServiceStarted* notification before sending the next request. The user interface includes text fields for specifying the JAR file URL and the number of times to execute the JAR with Cloud Gateway. In addition, the performance test application allows specifying threshold values, which it requests Cloud Gateway to use by invoking the *SetThresholdInNet* operation.

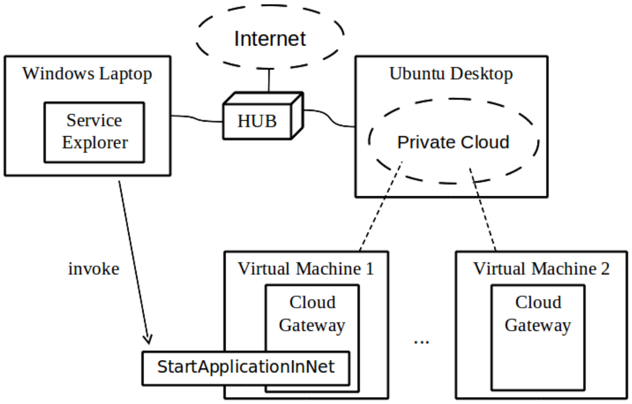


Figure 2. The test arrangement includes two physical machines, one of which hosts a private cloud containing virtual machines (VMs).

While experimenting with different threshold values, we have noticed that if only a RAM usage threshold were used, it should be set to at most 0.98 because the operating system never appears to let the RAM utilisation reach 99 percent but retains a small amount of memory as work space and compensates the missing memory with swap file usage. For example, with one gigabyte of RAM, the memory utilization typically reaches 98 percent after Cloud Gateway has started 28 conveyor service server applications, after which the proportion of used RAM fluctuates only marginally. However, the increased page file usage burdens the CPU, resulting in very poor performance. To prevent the CPU load from excessively increasing, a threshold value for CPU utilisation should be specified.

In one of our test runs, setting the RAM threshold to 0.99, and CPU utilisation threshold to 6.0 resulted in 42 applications being started before the CPU threshold was exceeded. Cloud Gateway started the applications in 217 seconds. However, the maximum number of applications and the start-up delay vary between different test runs. Given that the virtual machine hosting Cloud Gateway comprises only one (virtual) CPU, the load factor of 6.0 indicates that, on average, only five processes are queuing for CPU time. However, the web services, including Cloud Gateway, running on the machine seemed unable to respond to requests within the communication time out durations. Logging in to the virtual machine revealed that the load average had exceeded 60. Afterwards, the virtual machine became unreachable. Apparently, since the load average is computed over the previous minute [11], it is difficult to use it as a measure of the workload of the machine at a specific instant. On the other hand, the applications may temporarily have to queue for processing time at start-up, while later they will require less computation power.

On a virtual machine with only 256 megabytes of RAM, the memory utilisation exceeds 98 percent after Cloud Gateway has only started six server applications, and the overall delay is 22 seconds. When the memory threshold is set to 99 percent and CPU threshold is set to 6.0, Cloud Gateway starts nine server applications, but finally the virtual machine becomes unreachable.

The application start-up delay begins to increase steeply after Cloud Gateway has started a certain number of applications. This is obviously caused by the virtual machine having to compensate the lack of physical memory with page file usage. Moreover, the responsiveness of the applications running on a virtual machine is very poor when the machine is executing several applications simultaneously.

We have also experimented running the Cloud Gateway service on remote virtual machines leased from the Amazon EC2 cloud. In the experiments, each virtual machine hosting a Cloud Gateway service has been allocated 1.7 gigabytes of RAM. Table II shows the test results using a single virtual machine in the EC2 cloud. The table shows the memory threshold, number of started instances and overall start-up times. It also lists the reasons why Cloud Gateway stopped starting new server applications.

The last row in Table II represents a test scenario where the CPU threshold was set to 100. In this case, the client connection to the virtual machine abruptly terminated while starting the 94th application instance, apparently due to the excessive workload on the virtual machine.

C. Performance Measurement in a Network Setting

We have performed performance tests also in a setting of two Cloud Gateways running on separate virtual machines in our private computing cloud. Each of the virtual machines is allocated one gigabyte of RAM and five gigabytes of disk space. The Cloud Gateway Performance Test application communicates directly only with the main instance. However, it invokes the *RegisterCloudGateway* operation on the main instance to add the auxiliary instance to the Cloud Gateway network. The memory and CPU thresholds set in the user interface are submitted to each Cloud Gateway in the network.

In the scenario of two Cloud Gateway instances, the main instance will serve the first application requests. However, once it exceeds the memory threshold, the main instance starts delegating incoming application start requests to the auxiliary instance.

TABLE II. THE NUMBER OF CONVEYOR SERVICE APPLICATIONS THAT CAN BE STARTED ON A VIRTUAL MACHINE WITH 1.7 GB OF RAM.

Memory threshold	Reason for termination	Number of instances	Total duration (s)
0.9	memory	28	138
0.98	memory	33	173
1	CPU > 50	85	652
1	failure	93	1379

For example, in one of the test runs, the memory threshold of the two Cloud Gateways was set to 98 percent, while the CPU threshold was set to five. Finally, the performance test application started requesting the main Cloud Gateway instance to start instances of the conveyor service server application. The main instance exceeded the memory threshold after starting the 28th server application and started delegating the requests to the auxiliary instance on the other virtual machine. The auxiliary instance was able to start 27 applications before exceeding the memory threshold. Hence, a total of 55 application instances were started, and the total duration was approximately 210 seconds.

D. Inter-Cloud Experiment Scenario

To experiment web service orchestration across different computing clouds, we have performed an experiment with two remote virtual machines and one local virtual machine. The remote virtual machines are leased from the Amazon EC2 cloud, while the local virtual machine is running in our private computing cloud. Each virtual machine hosts one Orchestration Engine web service and three virtual conveyor web services.

The experiment consists of a cycle that starts when the Orchestration Engine on virtual machine 1 is requested to execute a BPEL process orchestrating the three conveyor web services. In the end of the process, the Orchestration Engine on virtual machine 1 requests the Orchestration Engine on virtual machine 2 to execute a similar process, which is represented by step 4 in Figure 3. Then, the Orchestration Engine on virtual machine 3 executes a similar BPEL process, which finally requests the Orchestration Engine on virtual machine 1 to again execute the BPEL process (step 12). Hence, the cycle continues indefinitely, so that only one Orchestration Engine is executing a BPEL process at a time.

To measure cycle durations, a client application monitors the Orchestration Engine service on the local virtual machine. Each time the Orchestration Engine sends a

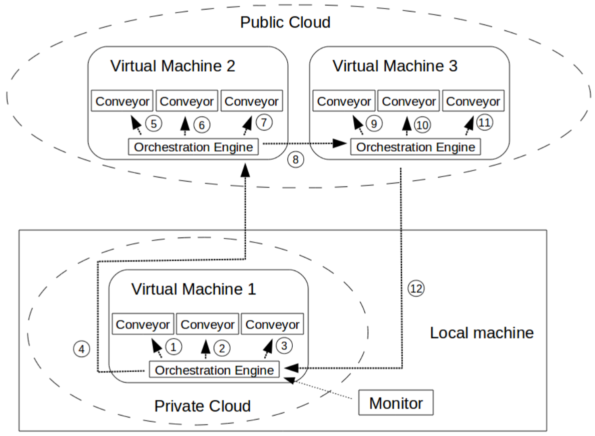


Figure 3. The private cloud is hosted on a local machine. Each virtual machine hosts four web services.

notification signalling that it has begun executing a BPEL process, the client application records the duration of the elapsed interval since the previous notification. It also determines the minimum, maximum and average interval length. The experiment topology is depicted in Figure 3. We have studied the use of BPEL in web service orchestration in [12] and described the Orchestration Engine web service in [7] and [13].

Table III contains the experienced minimum, maximum and average intervals in an experiment consisting of 20 cycles. To obtain a reference point, we repeated the experiment so that all of the web services were running on the local machine. As Table III shows, the average cycle duration is approximately two seconds longer when using computing clouds. This constitutes less than five percent of the average cycle time. The minor performance degradation is presumably caused by the network traffic between the web services on different virtual machines. However, network traffic is unavoidable when the Orchestration Engine services execute on different machines.

IV. Conclusion And Future Work

In this paper, we have proposed a web service that facilitates the use computing cloud resources. The method allows the use of computing cloud resources without knowledge on the cloud interface or internal composition. In particular, we have shown that a client can use the cloud resources by invoking simple web service operations without having to directly interact with the leased virtual machines.

A current limitation of the proposed approach is that cloud resources are somewhat inefficiently used. While it is possible to create a network of Cloud Gateway services running on separate virtual machines, the machines must be leased in a static manner, before launching the corresponding Cloud Gateways. Cloud Gateway could be enhanced so that it dynamically created new virtual machines as the utilisation of the existing ones reached a certain level.

We have carried out experiments to evaluate the performance of the proposed approach. While the automated execution of applications is effortless, the resource limits of the underlying virtual machine are eventually reached as the number of executed applications increases. Moreover, exhausting the resources over a certain point tends to considerably decrease application responsiveness. On the other hand, we have shown that forming networks of several Cloud Gateway services allows automatically balancing workload between several virtual machines.

Cloud Gateway measures the percentage of used memory and the system load average to avoid the overuse of computing resources. The method appears effective in preventing severe performance degradation when starting several applications. However, currently, Cloud Gateway is unable to estimate the amount of resources an application will consume once it has been started. Therefore, future research should target the implementation of a mechanism for evaluating the runtime resource consumption of the started applications.

In addition, we have experimented with web service Orchestration spanning separate computing clouds. The results suggest that using computing cloud resources causes no considerable performance drawbacks. However, deploying several web services on separate virtual machines requires manual work. Future research should investigate the use of Cloud Gateway in automating this task.

TABLE III. The duration of 20 cycles is quite similar regardless of whether cloud resources are used instead of local resources.

	Minimum (ms)	**Maximum (ms)**	**Average (ms)**
1 local VM, 2 remote VMs	47471	49106	47961
Only local web services	45519	46787	45682

References

[1] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, “A Break in the Clouds: Towards a Cloud Definition,“ ACM SIGCOMM Computer Communication Review, Vol. 39, Issue 1, pp. 50–55 (2009).

[2] K. Keahey, M. Tsugawa, A. Matsunaga, and J. Fortes, “Sky Computing,” Internet Computing, IEEE, Vol. 13, Issue 5, pp. 43–51 (2009).

[3] B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster, “Virtual Infrastructure Management in Private and Hybrid Clouds,” Internet Computing, IEEE, Vol. 13, Issue 5, pp. 14–22 (2009).

[4] G. Lawton, “Developing Software Online with Platform-as-a-Service Technology,” Computer, IEEE, Vol. 41, Issue 6, pp. 13–15 (2008).

[5] Amazon Elastic Compute Cloud, http://aws.amazon.com/ec2/, Referenced on 23.08.2010.

[6] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, Y. Lamia, and D. Zagorodnov, “The Eucalyptus Open-Source Cloud-Computing System,” Proc. 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pp. 124–131 (2009).

[7] J. Puttonen, A. Lobov, M. A. Cavia Soto, and J. L. Martinez Lastra, “A Semantic Web Services-Based Approach for Production Systems Control,” The Cognitive Factory special issue of the journal of Advanced Engineering Informatics, September 2010, in press.

[8] A. Sheth and A. Ranabahu, “Semantic Modeling for Cloud Computing, Part I,” Internet Computing, IEEE, Vol. 14, Issue 3 (2010).

[9] Devices Profile for Web Services Version 1.1, http://docs.oasis-open.org/ws-dd/dpws/1.1/os/wsdd-dpws-1.1-spec-os.html, Referenced on 23.08.2010.

[10] Web Services Dynamic Discovery (WS-Discovery), http://schemas.xmlsoap.org/ws/2005/04/discovery/, Referenced on 23.08.2010.

[11] R. Walker, “Examining Load Average,” Linux Journal, Issue 152, Dec. 2006, http://www.linuxjournal.com/article/9001, Referenced on 20.09.2010.

[12] J. Puttonen, A. Lobov, and J.L. Martinez Lastra, “An Application of BPEL for Service Orchestration in an Industrial Environment”, IEEE International Conference on Emerging Technologies and Factory Automation, pp. 530–547 (2008).

[13] A. Lobov, F. Ubis Lopez, V. Villaseñor Herrera, J. Puttonen, and J. L. Martinez Lastra, “Semantic Web Services Framework for Manufacturing Industries”, IEEE International Conference on Robotics and Biomimetics, pp. 2104–2108 (2009).

Qualitative Assessment Dynamics
For Trust Management in e-Business Environments

Denis Trček
Faculty of Computer and Information Science
University of Ljubljana
Tržaška cesta 25, 1000 Ljubljana, Slovenia - EU
Faculty of Mathematics, Natural Sciences and Inf. Technologies
University of Primorska
Glagoljaška 4, 6000 Koper, Slovenia - EU
denis.trcek@fri.uni-lj.si

Eva Zupančič
Faculty of Computer and Information Science
University of Ljubljana
Tržaška cesta 25, 1000 Ljubljana, Slovenia - EU
eva.zupancic@fri.uni-lj.si

***Abstract*—Trust is a core issue when it comes to acceptance of contemporary e-services. It was first addressed almost thirty years ago in Trusted Computer System Evaluation Criteria standard by the US DoD. But this and other proposed approaches of that period were actually addressing security. Roughly some ten years ago, methodologies followed that addressed trust phenomenon at its core, and they were based on Bayesian statistics and its derivatives, while some approaches were based on game theory. However, trust is a manifestation of judgment and reasoning processes. It has to be dealt with in accordance with this fact and adequately supported in e-environments. On the basis of the results in the field of psychology and our own research, a methodology called qualitative assessment dynamics (QAD) has been developed, which deals with so far overlooked elements of trust phenomenon. It complements existing methodologies and provides a basis for comprehensive trust management in e-environments.**

***Keywords*-distributed e-services; trust management; reasoning and judgment; modeling and simulation**

I. Introduction

Trust is an important phenomenon that forms the basis for many of our everydays decisions. Cyber space is no exception - the more sensitive an interaction in terms of security, privacy or safety is, the more trust there has to exist for an entity is to engage into an interaction. Some researchers even claim that trust is such essential resource that it is the main social virtue for the prosperity of societies [6]. Trust certainly has economic implications: In a trusted society business processes may run smoother and cheaper, because there is a reduced need for many checks (e.g., business reports), and acquisition of various means of insurance (e.g., bank guarantees, letters of credit). To ordinary users this may not appear familiar, but considering e-business environments like e-Bay, it becomes clear that trust in e-environments has significant business implications. Last but not least, the importance of trust is evident also to the highest ranking officials in the EU Commission that are stating that ”there is not yet enough trust in the Net” [19].

Before going into methodological details it is necessary to give the basic definitions first. According to the Cambridge Advanced Learner’s Dictionary, *trust is a belief or confidence in the honesty, goodness, skill or safety of a person, organization or thing*. For trust management in e-environments, this definition is not sufficient. A better definition is the one provided by Denning at the beginning of the nineties [4], when trust started to be more and more exposed in relation to security in information systems (IS). She vividly concluded that *trust is not a property of an entity or a system, but is an assessment. Such assessment is driven by experience, it is shared through a network of people interactions and it is continually remade each time the system is used*. And what is reputation? According to the Cambridge Advanced Learner’s Dictionary, *reputation is the opinion that people in general have about someone or something, or how much respect or admiration someone or something receives, based on past behaviour or character*. This enables us to treat reputation as an aggregated trust on the level of a certain society. Consequently, trust presents the basic building block, and we will concentrate on it in the rest of the paper.

The paper is structured as follows. In the second section an overview of existing methodologies for computerized trust management is given. In the third section a new, complementary methodology, called qualitative assessment dynamics (aka qualitative algebra) is presented that takes into account also research done in the field of psychology. There is a brief description of a technological solution for computerized trust management in the fourth section, while conclusions are given in the fifth section. The paper ends with the references in the last section.

II. A Brief Overview of the Field

A large number of initiatives in the field of trust management in e-environments came from the security research area. The main reason is probably that security and trust are closely related. These terms were used interchangeably as if they were expressing largely overlapping notions, which can be seen in early technical solutions. Although these were trust focused solutions, they were in fact security solutions. The first example is from 1996 when the World Wide Web Consortium standardized a Platform for Internet Content Selection (PICS) [13]. This technology was about access control, more precisely web-sites filtering. Web pages were rated by using defined labels and browsers could be set to exclude pages with a particular PICS rating or pages without this rating. The second example also dates back to 1996 when AT&T developed PolicyMaker, which was aimed at addressing trust management problems in network services [2]. Again, this was primarily a security solution that bounded access rights to the owner of a public key, whose identity was bound to this key through a certificate. The third example is from the year 2000, when IBM entered the area with the Trust Establishment Module [7]. This module was a Java based solution with appropriate language, similar to PolicyMaker. It enabled trusting relationships between unknown entities by using public key certificates and security policy.

At the turn of the century, EU funded projects followed that targeted trust. These attempts were already closer to addressing user behavior and the essence of trust, but many can be still characterized as largely security related technologies - some of them follow next. ITrust was a forum for cross-disciplinary investigation of the application of trust as a means of establishing security and confidence in the global computing infrastructure, where trust was recognized to be a crucial enabler for meaningful and mutually beneficial interactions [10]. TrustCOM was a framework for trust, security and contract management in dynamic virtual organizations. It was intended to be an open source reference implementation that builds on public specifications [5]. And finally INSPIRED was aimed at developing the next generation of security technologies needed for trusted access of users to e-services in a mobile or fixed environment. It was focused on smart-cards [12].

An interesting research from a non-security domain is described in the work of Cassell and Bickmore [3]. This approach addresses the essence of trust by deploying small talk to model social language and developing a collaborative relationship with users in agents based applications. Another interesting approach is taken in TRUSTe project [22] that is intended for promoting on-line business. TRUSTe services allow companies to communicate their commitment to privacy, and let consumers know which businesses they can trust. A similar approach is given in [17], where trust is supposed to be a matter of accreditation and certification of IT technology, which certainly makes sense within specific contexts.

Getting now to the theoretical basis, trust in computing environments is most often treated on the basis of Bayes theorem as the starting point. The theorem states that the posterior probability of a hypothesis H after observing datum D is given by $P(H \mid D) = P(D \mid H) * P(H)/P(D)$, where $P(H)$ is the prior probability of hypothesis H before datum D is observed, $P(D \mid H)$ is the probability that D will be observed when H is true, while $P(D)$ is the unconditional probability of datum D. This theorem has been used mainly for so called naïve trust management implementations [23].

A generalized Bayes theorem, the Dempster Shaffer theory of evidence, extends the classical concept of probability, where a probability p of stochastic event x, i.e. $p(x)$, and probability p of its complement $\overline{x}$, i.e. $p(\overline{x})$, sum up to 1. It does this by introducing uncertainty, meaning that $p(x) + p(\overline{x}) < 1$. The theory serves as a basis for subjective algebra, developed by Jøsang that is also used in computational trust management [9]. This algebra defines a set of possible states, a frame of discernment Θ. Within Θ, exactly one state is assumed to be true at any time. So if a frame of discernment is given by atomic states x_1 and x_2, and a compound state $x_3 = \{x_1, x_2\}$, which means that $\Theta = \{x_1, x_2, \{x_1, x_2\}\}$. Then, the belief mass is assigned to every state and in case of, e.g., x_3 it is interpreted as the belief that either x_1 or x_2 is true (an observer cannot determine the exact sub state that is true). Belief mass serves as a basis for belief function, which is interpreted as a total belief that a particular state is true, be it atomic or compound. This gives a possibility for rigorous formal treatment on a mathematically sound basis, where subjective algebra, in addition to traditional logical operators, introduces new operators like recommendation and consensus, and where trust is modeled with a triplet (b, d, u): b stands for belief, d for disbelief and u for uncertainty. Each of those elements obtains its values from the interval [0, 1], such that $b + d + u = 1$.

Finally, among main-stream methodologies that have been developed for computational trust management also game theoretic based ones should be mentioned - one typical representative is [1].

III. Qualitative Assessment Dynamics

The basis for methodology presented in this section is the research done in the area of psychology that provides an additional useful perspective on trust as a kind of reasoning and judgment process [18], [14], [15], and our own research [21]. Taking these works into account, the main factors that have to be considered are the following ones (for additional explanations of the above factors and their use for a formalized model that supports trust in computing environments, a reader is referred to [20]):

- Temporal dynamics - agent's relation towards the object / subject being trusted is certainly a dynamic relation that changes with time.

- Rationality and irrationality - an agent's trust can be driven by rational or irrational factors.
- Feed-back dependence - trust is not a result of an independent mind, but is influenced by environment.
- Action binding - trust can be a basis for agent's deeds.
- Trust differentiation - trust evolves into various forms because of the linguistic abilities of an entity expressing trust, or its intentions, and because of perception capabilities of a targeting entity.

The above works provide the main guidelines. However, additional reasons that suggest the need for a new, qualitative methodology, are the following (these address the shortcomings of the existing methodologies that are described in the previous section):

1) As to Bayesian statistics based methodologies, subjects have to understand basic concepts. However, many research results show that users often have problems with basic mathematical concepts like probability (see, e.g., [16]). Now even if subjects understand these basic mathematical concepts, very few of them understand advanced concepts that are required by, e.g., theory of evidence.
2) Methodologies that are based on game theory cannot be generally used for trust because of problems with preferences. In case of trust, preferences need not to exist, while in case of their existence, they are not necessarily transitive. So the two basic tenets of game theory are not fulfilled.
3) Our research indicates that users prefer qualitative expressions over quantitative ones when trust is in question. The qualitative ordinal scale is likely to consist of five ranks (qualitative descriptions) [21].

These facts call for a complementary method, which will be defined in the rest of this section.

Definition 1. *Trust is a relationship between agents A and B that can be totally trusted, partially trusted, undecided, partially distrusted, and distrusted; it is denoted by $\omega_{A,B}$, which means agent's A attitude towards agent B.*

The below figure illustrates the definition. There are four trust relationships, two of them addressing judgments of entities A and B towards themselves ($\omega_{A,A}$ and $\omega_{B,B}$), and two of them addressing judgment of one entity towards another entity ($\omega_{A,B}$ and $\omega_{B,A}$).

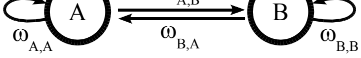


Figure 1. The definition of trust relationships

Next, the general nature of trust is that it is not reflexive (in certain contexts one may trust himself / herself, in others not), not symmetric (if agent A trusts agent B in a certain context, this gives no basis for automatic conclusion that agent B also trusts agent A), and not transitive (entity A may trust entity B, which in turn may trust entity C, but the latter may not be trusted by A).

This already suggests that trust is not an easy problem. Moreover, it can be proved that it is computationally hard problem - a proof outline follows: Suppose entity A assigns trust value for herself in a certain context, while entity B assigns another value to himself in the same context. When these entities are treated as a new compound entity AB (a team), the trust of this compound entity towards itself often differs from both trust values mentioned earlier (a typical example are sports games where an additional player in a team presents advantage for the whole team and changes judgments about its capabilities at all members of the team). The above fact implies that all relationships have to be considered among all possible entities, be it atomic or compound. As the number of compound entities can be obtained by computing the number of combinations that can be formed from the set of atomic entities, the total number of trust relationships N in a society with n atomic entities is given by the following equation:

$$N = (\sum_{m=1}^{n} \binom{n}{m})^2 \qquad (1)$$

Suppose we have a society with $n = 3$ atomic entities A, B and C. This means that the number of atomic entities is three, the number of compound entities with two atomic elements is three, and the number of compound entities with three atomic elements is one. So the total number of atomic and compound entities is $k = 7$, and there are $(k-1)*k$ relationships between them, where k relationships have to be added, because trust is not reflexive. Thus the total number of trust relationships is $N = 49$.

To enable the analysis and modeling of trust dynamics in social environments trust graphs are introduced. The links of trust graphs are directed and weighted accordingly. If a link denotes trust attitude of agent A towards agent B, the link is directed from A to B. Because graphs can be equivalently presented with matrices, this second definition can be given.

Definition 2. *In a given context Γ, trust in social interactions is represented by trust matrix $\mathbf{M}_\Gamma$, where elements $\omega_{i,j}$ denote trust relationships of i-th agent towards j-th agent, and where its values taken from the set* $\{1, 1/2, 0, -1/2, -1, -\}$. *These values denote trusted, partially trusted, undecided, partially distrusted and distrusted relationships. The last symbol, "-", denotes an undefined relation (an agent is either not aware of existence of another agent, or does not want to disclose its trust).*

A general form of trust matrix $\mathbf{\Omega}_\Gamma$ of a certain society with n agents in a given context Γ is as follows:

$$\begin{bmatrix} \omega_{1,1} & \omega_{1,2} & \cdots & \omega_{1,n} \\ \omega_{2,1} & \omega_{2,2} & \cdots & \omega_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n,1} & \omega_{n,2} & \cdots & \omega_{n,n} \end{bmatrix}_{\Gamma}$$

An example of a certain society with trust relationships and qualitative weights is given in Fig. 3:

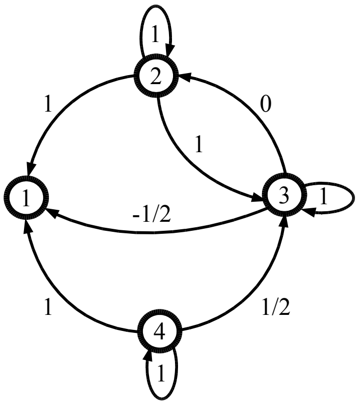


Figure 2. An example society that includes a dumb agent

The corresponding matrix is as follows:

$$\begin{bmatrix} - & - & - & - \\ 1 & 1 & 1 & - \\ -1/2 & 0 & 1 & 1/2 \\ 1 & - & 1/2 & 1 \end{bmatrix}$$

Trust matrices operations differ from those in ordinary linear algebra. Rows represent certain agents trust towards other agents, while columns represent trust of community related to a particular agent (columns are referred to as trust vectors). Further, technological components or services are treated as dumb agents. They can be recognized in a trust matrix by rows that consist exclusively of ”–” values.

It is a fact that certain entity may not equally treat all judgments from various entities, therefore there has to exist a possibility for pondering values. This is achieved by introduction of a ponder matrix $\mathbf{\Pi}$:

$$\begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,n} \end{bmatrix}_{\Gamma}$$

Above, $p_{i,j}$ states a weight (from the interval [0,1]) that an entity i is assigning to judgments of entity j. Therefore, rows represent ponders that a certain entity is assigning to judgments of all other entities in a society. To keep things simple, this matrix will be left out the rest of the paper.

Now qualitative operators can be introduced; they are taken from the set $\{\Uparrow, \Downarrow, \leadsto, \leftrightarrow, \uparrow, \downarrow, \odot\}$, and defined in detail in table 1, and described below:

- Extreme-optimistic judgment, which results in the most positive judgment in a society; it is denoted by ”$\Uparrow$”.
- Extreme-pessimistic judgment, which outputs the most negative judgment in a society; it's denoted by ”$\Downarrow$”.
- Centralistic consensus seeker judgment, which results in a towards zero ”rounded average”; its symbol is ”$\leadsto$”.
- Non-centralistic consensus-seeker judgment, which results in a value that is ”an average” rounded away from 0; it's denoted by ”$\leftrightarrow$”.
- Moderate optimistic judgment, which means the expressed judgment is ”strengthened” to the next higher level, narrowing the gap towards the aggregated judgment of the rest of community if this is more optimistic than the agent's trust is; it is denoted by ”$\uparrow$”.
- Moderate pessimistic judgment, which means the expressed judgment is weakened to the next lower level, narrowing the gap towards the aggregated judgment of the rest of community if this is more pessimistic than the agents trust is (the value changes one level downwards); it is denoted by symbol ”$\downarrow$”.
- Self-confident judgment, which preserves the same value after changes are calculated; its symbol is ”$\odot$”.

For the calculation of new trust values (and new trust matrix) the following algorithm is defined:

1) Take the first value in a trust matrix.
2) If the value is ”-”, write again ”-”, and go to step 6.
3) Calculate the average of a trust vector by excluding agents own opinion and values marked with ”-”.
4) Round the obtained average to the nearest possible judgment value from the set of judgment increments $\{1, 1/2, 0, -1/2, -1\}$.
5) Compute the result $\omega^{+}_{i,k}$ according to table 1 by treating the value from step 4 as $\omega^{-}_{j,k}$, and agents own opinion as $\omega^{-}_{i,k}$.
6) If there still exist unprocessed values, take the next value from the trust matrix and go to step 2, else stop.

Now suppose that in the example society in Fig. 3 agent 2 conforms to the optimistic operator, agent 3 to pessimistic operator, while agent 4 is a centralistic consensus seeker, the calculated simulation would be as follows:

$$\begin{array}{c} \\ \Uparrow \\ \Downarrow \\ \leadsto \end{array} \begin{bmatrix} - & - & - & - \\ 1 & 1 & 1 & - \\ -1/2 & 0 & 1 & - \\ 1 & - & 1/2 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} - & - & - & - \\ 1 & 1 & 1 & - \\ -1/2 & 0 & 1/2 & - \\ 1/2 & - & 1/2 & 1/2 \end{bmatrix}$$

Note that matrices $\mathbf{M}_{\Gamma}$ contain non-calculated values, but only ”pure judgments” entered by entities. They constitute, so to say, raw data for our calculations that are used by our algebra to support decision making. Now some important decision making questions are as follows:

- By running the simulation on a given society, is the society likely to reach an equilibrium?
- If it does reach an equilibrium, which entities will be most likely trusted by the society, and which not?

- How long will it take for the society to reach the most likely state and what state will this be?
- On which part of the society makes most sense to put most efforts to drive the community into a desired state?

$\omega^-_{i,k}$	$\omega^-_{j,k}$	$\omega^+_{i,k}, \Uparrow_i$	$\omega^+_{i,k}, \Downarrow_i$	$\omega^+_{i,k}, \leadsto_i$	$\omega^+_{i,k}, \leftrightarrow_i$	$\omega^+_{i,k}, \uparrow_i$	$\omega^+_{i,k}, \downarrow_i$	$\omega^+_{i,k}, \odot_i$
-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-½	-½	-1	-½	-1	-½	-1	-1
-1	0	0	-1	-½	-½	-½	-1	-1
-1	½	½	-1	0	-½	-½	-1	-1
-1	1	1	-1	0	0	-½	-1	-1
-1	–	–1	-1	-1	-1	-1	-1	-1
-½	-1	-½	-1	-½	-1	-½	-1	-½
-½	-½	-½	-½	-½	-½	-½	-½	-½
-½	0	0	-½	0	-½	0	-½	-½
-½	½	½	-½	0	0	0	-½	-½
-½	1	1	-½	0	½	0	-½	-½
-½	–	-½	-½	-½	-½	-½	-½	-½
0	-1	0	-1	-½	-½	0	-½	0
0	-½	0	-½	0	-½	0	-½	0
0	0	0	0	0	0	0	0	0
0	½	½	0	0	½	½	0	0
0	1	1	0	½	½	½	0	0
0	–	0	0	0	0	0	0	0
½	-1	½	-1	0	-½	½	0	½
½	-½	½	-½	0	0	½	0	½
½	0	½	0	0	½	½	0	½
½	½	½	½	½	½	½	½	½
½	1	1	½	½	1	1	½	½
½	–	½	½	½	½	½	½	½
1	-1	1	-1	0	0	1	½	1
1	-½	1	-½	0	½	1	½	1
1	0	1	0	½	½	1	½	1
1	½	1	½	½	1	1	½	1
1	1	1	1	1	1	1	1	1
1	–	1	1	1	1	1	1	1
–	*	–	–	–	–	–	–	–

Figure 3. The definition table for qualitative operators ($*$ means any value)

GUI of trustGuard component that is used for QAD simulations is given in Fig. 4. The parameters were set as follows: The complete society consisted of ten agents, of which 40% behaved according to optimistic operator, 20% according to pessimistic operator, and there were 20% opponents and 20% centralists. Further, the initial distribution of trust values in the trust matrix was 20% of values denoted by 1, 20% denoted by $1/2$, 20% denoted by 0, 20% denoted by $-1/2$, and 20% denoted by -1 (there was no dumb agent). In addition, 30% of agents were allowed to randomly change their operators, and there were 5 simulation steps between these random changes. After running the situation for a sufficiently rong time (for approx. 970 steps), we reach an equilibrium, where 10% of values in the trust matrix 0 (i.e. undecided), while 90% of values were -0.5 (partially distrusted). Finally, an agent with a fat line around it is partially distrusted by the society in the end.

Despite the fact that more detailed discussion of the simulation processes exceeds the scope of the paper, an experienced reader can see that this component enables sound simulations by providing, e.g., expected values for variables in question, their distribution, etc. To conclude this section - it clearly follows that we are dealing with a non-linear dynamic system. Therefore analytic solutions will be mere exceptions and we will have to rely on simulations (to search for various heuristics and solutions for typical, reference scenarios, etc.). Despite this, various interesting theoretical questions can be addressed [21]).

IV. Trust Management Implementation

Our solution for trust management is called trustGuard. It consists of two basic building blocks: the distributed database where trust values (matrices) are stored, and the user interface that accesses this database, performs insertion and retrieval of these values, and does QAD calculations. The distributed database is implemented on SOA standards, so user interface interacts with these databases through SOAP protocol. For this to happen, the following two primitives are needed. The first one is *trustQuery*, and the second one is *trustReply*. These primitives are defined with XML schema. But for clarity and conciseness, XML DTD is chosen to present the syntax of *trustReply* primitive:

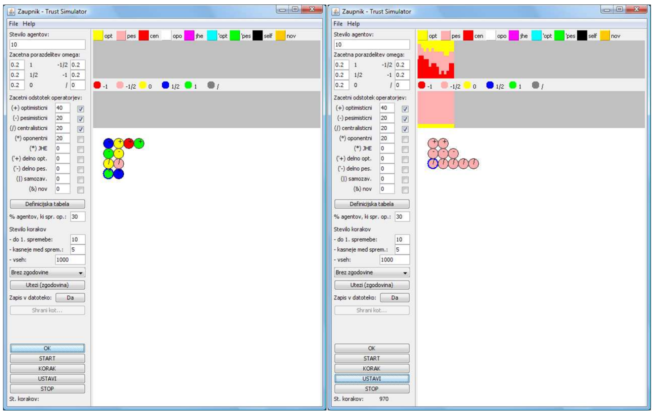
```
<!ELEMENT   trustResponse (timeStamp, trustMatrix,
            function?, extension?) >
<!ELEMENT   timeStamp (#PCDATA) >
<!ATTLIST   timeStamp zulu
            CDATA #REQUIRED >
<!ELEMENT   trustMatrix (omega+) >
<!ELEMENT   omega (id1, id2, trustAssessment) >
<!ELEMENT   id1 (#PCDATA) >
<!ATTLIST   id1 URI1
            CDATA #REQUIRED >
<!ELEMENT   id2 (#PCDATA) >
<!ATTLIST   id2 URI2
            CDATA #REQUIRED >
<!ELEMENT   trustAssessment EMPTY >
<!ATTLIST   trustAssessment
            value (−1| − 0.5|0|0.5|1|−) ” − ” >
<!ELEMENT   function (#PCDATA) >
<!ATTLIST   function OID
            CDATA #REQUIRED >
<!ELEMENT   extension (#PCDATA) >
```

The generalized time is expressed as Greenwich Mean Time (Zulu) in the form YYYYMMDDHHMMSS, while trust assessment functions are uniquely identified through OIDs [8]. The syntax of *trustQuery* is similar to the syntax of *trustReply*, except that there are no *trustMatrix* elements. The *extension* element is included and is added in both primitives for future extensions.

Current trustGuard implementation supports not only qualitative algebra, but also, e.g., Jøsang' s subjective algebra. As further implementation details exceed the scope of this paper, a reader can find more information in [11].

V. Conclusion

In the medieval era, Shakespeare advised us to love all, trust a few, and do wrong to none. Later, the famous German poet Goethe, with a strong sense for deep analyses claimed that as soon as one trusted himself (herself), one knew how to live. And recently, prof. H. Smead vividly noted: ”When we were young, we didn’t trust anyone over thirty. Now that we’re over thirty, we don’t trust anyone at all”.



Stevilo agentov: 10
Zacetna porazdelitev omega:
0.2 1 -1/2 0.2
0.2 1/2 -1 0.2
0.2 0 / 0
Zacetni odstotek operatorjev:
(+) optimisticni 40 ☑
(-) pesimisticni 20 ☑
(/) centralisticni 20 ☑
(°) oponentni 20 □
(*) JHE 0 □
('+) delno opt. 0 □
('-) delno pes. 0 □
(|) samozav. 0 □
(&) nov 0 □
% agentov, ki spr. op.: 30
Stevilo korakov
- do 1. spremebe: 10
- kasneje med sprem.: 5
- vseh: 1000
Zapis v datoteko: Da
St. korakov:

Stevilo agentov: 10
Zacetna porazdelitev omega:
0.2 1 -1/2 0.2
0.2 1/2 -1 0.2
0.2 0 / 0
Zacetni odstotek operatorjev:
(+) optimisticni 40 ☑
(-) pesimisticni 20 ☑
(/) centralisticni 20 ☑
(°) oponentni 20 □
(*) JHE 0 □
('+) delno opt. 0 □
('-) delno pes. 0 □
(|) samozav. 0 □
(&) nov 0 □
% agentov, ki spr. op.: 30
Stevilo korakov
- do 1. spremebe: 10
- kasneje med sprem.: 5
- vseh: 1000
Zapis v datoteko: Da
St. korakov: 970

Figure 4. An example of a simulation run with the trustGuard component

It follows that trust is a sensitive and scarce resource. This especially holds true for e-business environments, where competition is only a few mouse clicks away, while the medium by its nature is not able to provide communication details that are available in face to face contacts; therefore new mechanisms have to be developed and deployed. Further, if users are to be adequately supported when trust management is an issue, the solutions have to be aligned with mental models. These issues have led to the development of qualitative algebra, and they were also the basis for theoretical views, as well as for the practical implementation.

Qualitative algebra complements existing approaches that depend on rational mechanisms like Bayesian statistics and game theory. It is based on research in the field of psychology and addresses irrational elements, feed-back dependence, and context dependence. It provides basic means also for more advanced problems through simulations like ”How can a society be guided in order to achieve (with a certain probability) a trusted atmosphere?”. Clearly, with questions like this simulation is one of the most suitable approaches, because trust related problems belong to the area of complex, non-linear dynamics. Thus computationally supported trust management is not only a must because of the nature of e-media, but because of the trust phenomenon itself.

ACKNOWLEDGMENT

Authors thank to Slovene Research Agency ARRS for support of this research with grants J2-9649 and P2-0359 (Pervasive computing).

REFERENCES

[1] Aberer K. and Despotovic Z., *On Reputation in Game Thory - Application to Online Settings*, Working Paper, Swiss Federal Institute of Technology (EPFL), Zurich, 2004.

[2] Blaze M. and Feigenbaum J., Lacy J., Decentralized Trust Management, *Proc. of the ’96 IEEE Symposium on Security and Privacy*, Oakland, pp. 164–173, 1996.

[3] Cassel J., Bickmore T., Negotiated Collusion - Modeling Social Language and its Relationship Effects in Int. Agents, *User Model. & User-Adapted Int.*, 13(4), pp. 89–132, 2003.

[4] Denning D., A new Paradigm for Trusted Systems, *Proc. of ACM SIGSAC New Security Paradigms Workshop*, ACM, New York, pp. 36-41, 1993.

[5] Dimitrakos T., Wilson M., Ristol S., TrustCoM - A Trust and Contract Management Framework enabling Secure Collaborations in Dynamic Virtual Organisations, *ERCIM News*, 2004(59), pp. 59–60, 2004.

[6] Fukuyama F., *Trust: The Social Virtues and the Creation of Prosperity*, Free Press, New York, 1995.

[7] Herzberg A. et al, Access Control Meets Public Key Infrastructure, *Proc. of the IEEE Conf. on Security and Privacy*, Oakland, pp. 2-14, 2000.

[8] ITU-T, *Specification of Abstract Syntax Notation One (ASN.1)*, Recommendation X.208, Geneva, 1988.

[9] Jøsang A., A Logic for Uncertain Probabilities, Int. J. of Uncertainty, *Fuzziness and Knowledge-Based Systems*, Vol. 9, Issue 3, pp. 279–311, World Scient. Publishing, London, 2001.

[10] Klyne G., Survey of Papers from the iTrust 2003 and 2004 Trust Management Conferences, 2004, http://www.ninebynine.org/iTrust/iTrust-survey.html.

[11] Kovač D. and Trček D., Qualitative trust modeling in SOA, *J. Syst. Archit.*, Vol. 55, No. 4, pp. 255–263, Elsevier, 2009.

[12] Linke A. and Manteau L., Report on the EU Research Project Inspired: The future of Smart Cards, 2005, http://www.inspiredproject.com/documents/20050728-paper-isse2005-final.pdf.

[13] Miller J., Resnick P., Singer D., PICS Rating Services and Systems, 1996, http://www.w3c.org/TR/REC-PICS-services.

[14] Muir B.M., Trust in automatition-I - Theoretical issues in the study of trust and human intervention in automated systems, *Ergonomics*, 37(11), pp. 1905–1922, 1994.

[15] Muir B.M. and Moray N., Trust in automatiotion-II - Experimental studies of trust and human intervention in a process control simulation, *Ergonomics*, Special Issue, Cognitive ergonomics, 39(3), pp. 429–460, 1996.

[16] Nisbett R.E., Krantz D.H., Jepson C., Fong T.G., Improving inductive inference, in Kahneman D., Slovic P., Tversky A. (Eds.), Judgement under uncertainty: Heuristic and Biases, pp. 445–459, Cambridge University Press, Cambridge, 1982.

[17] Osterwalder D., Trust Through Evaluation and Certification?, *Soc. Sci. Comp. Review*, 19(1), pp. 32–46, 2001.

[18] Piaget J., *Judgment and Reasoning in the Child*, Routledge, London, 1999.

[19] Reding V., Safety on the Net. Int. High Level Research Seminar on Trust in the Net, Vienna, 2006, http://ec.europa.eu/comm/commission_barroso/reding/docs/speeches/viennaq_20060209.pdf.

[20] Trček D., A formal apparatus for modeling trust in computing environments, *Math. and Comp. Modelling*, 49(2009), pp. 226–233, Elsevier, 2009.

[21] Trček D., Ergonomic Trust Management in Pervasive Computing Environnets (invited talk), *Proc. of ICPCA ’10*, Maribor, pp. 1–6, 2010.

[22] TRUSTe, Security Guidelines. TRUSTe, San Francisco, 2005, http://www.truste.org/pdf/SecurityGuidelines.pdf.

[23] Wang Y. and Vassileva J., Trust and Reputation Model in P2P Networks, *Proc. of the 3.rd Int. Conf. on Peer-to-Peer Computing*, pp. 150–1554, 2003.

An Experimentation System for Bus Route Planning and Testing Metaheuristics Algorithms

Krzysztof Golonka, Leszek Koszalka, Iwona Pozniak-Koszalka, Andrzej Kasprzak
Dept. of Systems and Computer Networks, Wroclaw University of Technology
Wroclaw, Poland
e-mail: leszek.koszalka@pwr.wroc.pl

***Abstract*—In this paper, we present an experimentation system for school bus route planning and testing various algorithms to solve such an optimization problem. It is a crucial social issue that concerns faster and more comfortable transport. Moreover, the route optimization allows decreasing the ticket price by maximizing the profit of the provider. Since the problem belongs to hard optimization problems, thus, we considered four meta-heuristic algorithms: three adapted, including Tabu Search, Simulated Annealing, Genetic Algorithm, and algorithm invented by the authors called Constructor. The efficiency of algorithms was tested and compared to that found by Complete Overview using the designed and implemented experimentation system. The investigations made on various problem instances, allowed to emerge the most efficient algorithm.**

Keywords-experimentation system; metaheuristic algorithms; route planning; optimization; efficiency

I. INTRODUCTION

Since banking crisis from 2008 many companies were forced to cut expenses and look for more savings. Moreover, nowadays a lot of pressure is put on being green – environmentally - friendly, especially when it comes to industry or transport e.g. [1]. A lot of effort must be put in analysis and planning process to come across these challenges. This paper focuses on optimization of a school bus route and proposes the direction of searching optimal solutions. The main goal is to find the most profitable route (e.g. the shortest path).

This optimization belongs to non-polynomial problem and has a huge solution space, meaning we can not find the best solution in polynomial time. For small instances it is easy to search through the whole solution space but when instance begins to grow, the required time may become unacceptable. The only one reasonable way to solve this is to use meta-heuristic algorithms that were invented to struggle with such problems. An idea to consider such algorithms based on artificial intelligence like Tabu Search (e.g. described in [2], [3], and [4]), Simulated Annealing (e.g. explained in [5], [6], and [7]) and Genetic Algorithm (e.g. illustrated in [8], [9], and [10]) seems to be promising.

We decided to adapt all three mentioned ideas for implementing algorithms to find an efficient solution to bus route problem. In addition, we tried to invent on our own a new algorithm and we created the algorithm called Constructor which is described in this paper. To determine the shortest path from the starting point to the ending point through all possible bus routes we implemented Bellman-Ford Algorithm (e.g. [11]).

Moreover, we implemented Complete Overview (CO) algorithm to be able to calculate differences between maximum and optimum found using meta-heuristics (unfortunately, CO may be used only for instances smaller than 20 bus routes because of its complexity and the required time). Finally, we had to introduce an evaluating function as the measure of efficiency for the considered algorithms.

To assess algorithms' efficiency we implemented an experimentation system that allows user to perform series of tests along with multistage experiment design ideas presented in [12].

The rest of paper is organized as follows: Section II defines the problem to be solved. Section III describes the considered algorithms and their roles in optimization process. Section IV shortly presents the implemented experimentation system. The results of the research appear in Section V. Section VI provides some final remarks and conclusion.

II. TYPE STYLE AND FONTS

There is given a certain urban area (Fig. 1), that consists of links and Bus Stops (BSs). Lines represent links, numbers next to them represent their lengths and red dots symbolize Bus Stops. The beginning of the route is marked by a green rectangle, but blue ending point (rectangle) represents the location of the school.

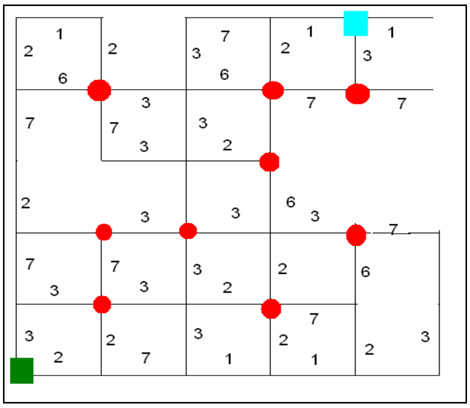


Figure 1. An example of an instance of bus route planning problem.

It is necessary to determine the most profitable route of a school bus to maximize profits of a bus provider.

ISBN:978-1-61208-114-4

This means that the route may consist only BSs at which the number of students (pupils) waiting for a bus is sufficient not to make loses. Once the route is planned the bus may omit some BSs which are not on this specified route.
The basis for making decision on which BS should stop is the observation of statistics that deliver the number of students (pupils) waiting at a BS on a given time. These values are represented by a matrix, in which each row corresponds to the next hour in bus transit and the columns show the number of pupils (Table 1). The problem parameters are listed in Table 2.

TABLE 1. Students (pupils) statistics - an example.

Bus Stop Time	#1	#2	#3	#4
8:00am	5	5	16	9
8:45am	6	1	3	11
9:30am	10	22	4	9
10:15am	0	2	0	0

TABLE 2. Input parameters of the problem.

Sign	Parameter
SBS	Set of potential BSs
(x_i,y_i)	BS_i coordinates
$L_{i,j}$	Link length between i and *j* BS
$P_{k,j}$	Pupils at k BS at j transit
T	Ticket price
DC	Driver's cost
BC	Bus exploitation cost
J	Number of transits

A. Basic Terms

Bus Stop (BS) is a point on a map where pupils wait for a bus to school. Each BS has its coordinates *x* and *y* on a map.

$$BS = [x, y] \tag{1}$$

Set of Potential BSs (SBS) is a collection of all BSs on a map, where SBS(*i*) may be defined by (2).

$$SBS(i) = \begin{bmatrix} BS_1 \\ BS_2 \\ \mathrm{M} \\ BS_i \end{bmatrix} = \begin{bmatrix} x_1, y_1 \\ x_2, y_2 \\ \mathrm{M} \\ x_i, y_i \end{bmatrix} \tag{2}$$

Link (L) is a matrix defined by (3) that describes lengths of links between BS_i and BS_j. Some BSs may not be directly linked to others but each BS must have at least one link.

$$L(i,j) = \begin{bmatrix} l_{1,1} & \Lambda & l_{1,j} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ l_{i,1} & \Lambda & l_{i,j} \end{bmatrix} \tag{3}$$

$l_{i,j}$ is defined as:

$$l_{i,j} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \tag{4}$$

Route (R) is a path consisting starting point, ending point and going through BSs chosen from SBS. A Route may contain smaller amount of BSs than SBS.

$$R = SBS(k) \qquad \text{where } k \leq i \tag{5}$$

Ticket Price (T) informs how much each pupil must pay for taking a bus.

Driver's Cost (DC) equals money paid to a driver for driving one unit of a length of a Route.

Bus Exploitation Cost (BC) equals expenses for fuel used by a bus after driving one unit of a length of a Route.

Number of Transits (J) says how many times the bus is going a route per a day.

Route Length (RL) is a length of a shortest path.

B. Evaluating Function

The evaluating function introduced by us is called the Balance (6) interpreted as a daily balance – obtained after one day of work. If *Q(R)* is less than 0 that the provider gets loses, if greater than 0 that the provider gets profits.

$$Q(R) = \sum_{j=1}^{J} \left(\sum_{k=1}^{K} P_{k,j} * T - RL * (DC + BC) \right) \tag{6}$$

Moreover, to make sure that the bus will not go from starting point right to the destination point, we introduce a constraint. The constraint describes the minimal percentage number of all pupils from the statistics (Table 1) that should be delivered to the school (7).

$$100\,\% * \frac{\sum_{j=1}^{J} \sum_{k=1}^{K} P_{k,j}}{\sum_{j=1}^{J} \sum_{i=1}^{I} P_{k,i}} \geq P_{\%} \tag{7}$$

III. THE ALGORITHMS

A. Basic Ideas

In the paper, we consider four meta-heuristic algorithms as the main algorithms, including three known algorithms (but specially adopted) : TS - Tabu Search, SA - Simulated Annealing, GA - Genetic Algorithm, and the Constructor (originally proposed by the authors of the paper). The first two of them are described, e.g. in [13]. Our adaptations of GA as well as the Constructor are explained below. The

Route Length is calculated by using Bellman-Ford algorithm ([11]). TS, SA, and GA perform calculations for a certain number of iterations, processing on a solution and its neighborhood. A solution is a RL and the neighborhood is a set of Routes with one different BS, without one BS or with additional one BS.

The performance of each meta-heuristic algorithm is affected by a few factors such as an instance parameters (the size of SBS) and algorithms inner parameters.

B. *Simulated Annealing (SA)*

In any iteration the solution is replaced by a new one randomly chosen from the neighborhood if the new one is better. If the new solution is worse it has 50% of chances to replace the previous one. In this particular implementation we do not have such thing as temperature that changes the probability of replacing solutions. Here probability is constant. This is the only one difference between our SA and the one described in [7].

C. *Tabu Search (TS)*

The implemented Tabu Search is more complex algorithm than SA because of applying the searching procedure through the whole neighborhood of a recent solution and choosing the best one unlike SA. Moreover, TS is more resistant to loops thanks to the taboo list. The best found solution may replace the previous one only if it differs from all the records in a taboo list more than of a certain percentage value. The length of the taboo list is limited and when the list is full, the old records are overwritten.

D. *Genetic Algorithm (GA)*

This algorithm is based on evolutionary mechanisms. The main idea is to create a population of a constant size and observe its evolution meanwhile registering the best ones. The most interesting here is a cross-over process. It requires two individuals and eventually gives two children. DNA chain is represented in this situation as a single solution. The crossover is described below on an example:

parent no1: 1001|1101
parent no2: 1100|0111
child no1: 1001|0111
child no2: 1100|0111

All the solutions have a chance to hand over gens but the higher the Price function value of the solution, the higher possibility of being picked as a parent. Moreover, each child may mutate with probability 50% that leads to changing only one randomly picked chromosome. As the population size is constant, the newborns must replace the old solutions regardless of their breeding history.

The algorithm implemented in our specified problem may be described as follows:

Step 0. Initial population is picked randomly.
Step 1. Pick parents randomly from existing population.
Step 2. Perform breeding process.
Step 3. Choose solutions to extinct.
Step 4. Add children solutions to the rest of solutions in existing population.
Step 5. Check whether the evaluating function for each solution in current population is not the best global optimum from already explored solution space.
Step 6. Go back to step 1 if stop condition (defined by parameter Iterations) is not fulfilled.

E. *The Constructor*

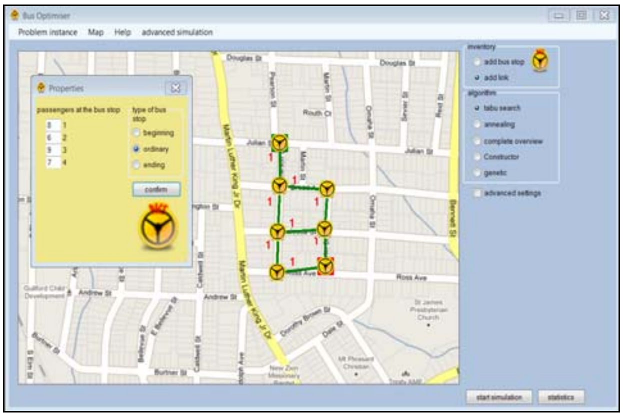
The algorithm named Constructor was invented by us – it is based on the decomposition concept - we assumed that splitting a big instance to smaller ones, solving them separately and joining all together may be effective.

At the beginning the Constructor splits the whole instance - containing two BSs, the beginning BS and the last BS. Next, it searches through a neighborhood of each small instance and modify them. After that, the algorithm combines in pairs small instances integrating them as obtaining the initially considered instance.

Step 1. For specified short route that consists of beginning BS, ending BS and temporary amount of BSs find new possible routs picked from the short route's neighborhood.
Step 2. Check the optimization function for solutions, which include new route. Find the best neighbor.
Step 3. Modify current solution by including that best neighbor as a part of the solution.
Step 4. Pick next temporary amount of BSs and go back to step 1, unless all of BSs has already been picked.
Step 5. Double the temporary amount of BSs and go back to step 1, unless the temporary amount of BSs exceeded amount of all BSs.

IV. EXPERIMENTATION SYSTEM

The application was designed, mainly in order to visualize the tested algorithms. It was created using Visual Studio 2008. The implementation language was C#. In Fig. 2, a screenshot of the application window is shown.



passengers at the bus stop
8 1
6 2
9 3
7 4

type of bus stop
○ beginning
● ordinary
○ ending

○ add bus stop
● add link

● tabu search
○ annealing
○ complete overview
○ Constructor
○ genetic

□ advanced settings

Fig. 2. Application window.

In the beginning the user defines an instance of problem by putting BSs on the map and creating links between them.

Next, the beginning and the ending points of the potential route are precised and pupil statistics (by clicking and selecting properties) is determined. Finally, the user selects the considered algorithm and fixes its parameters. After clicking on “start simulation” button the application is searching for an optimal solution. There is a possibility to see an animation of a transit, and some statistics presented also on plots (e.g. served BSs, Balance, the percentage of served pupils) that allows observing current results.

V. INVESTIGATIONS.

A. Calibrating Algorithms - Concept

The first part of research refers to finding the values of the best inner parameters for two algorithms:

- Tabu Search,
- Genetic Algorithm

For each new set of parameters, a new simulation was made. The parameters of the problem for the considered instances are shown in Table 3. A single series of experiment meant that 10 different instances were tested; moreover, each instance was repeated 10 times. In Table 4 and Table 5 are the averages values over a given set of series of experiment.

TABLE 3. Parameters – set 1.

Bus Stops	15
Test iterations	10
Instances to test	10
[%] passengers	50
Driver's cost	1
Bus exploitation cost	1
Ticket price	3
Number of transits	4

B. Adjusting Parameters for Genetic Algorithm

Tests showed that increasing size of the population as well as increasing number of parents does not have a remarkable influence on improvement of solution (only about 5%). But the number of iterations has vast impact on results (Table 4).

TABLE 4. Results given by GA.

Test no.	GA	CO	ΔGA [%]	Population	Parents	Iterations
1	1360	1990	46,32	10	6	20
2	1384	1971	42,41	20	6	20
3	1292	1765	36,61	20	10	20
4	1243	1726	38,86	20	14	20
5	1215	1728	42,22	20	18	20
6	1256	1610	28,18	20	18	30
7	1307	1573	20,35	20	18	40
8	1259	1489	18,27	20	18	60

It may be observed, that for 60 iterations the results differ from the maximum just of approx. 18%. Because of slight differences in results between 60 and 40 iterations, the optimal value of this parameter was taken as equal to 40.

C. Adjusting Parameters for Tabu Search

The obtained results of research are shown in Table 5.

TABLE 5. Results given by TS.

Test no.	Tabu Search	CO	ΔTS [%]	Tabu list Length	% Precision	Iterations
1	1328	2255	69,80	4	10	20
2	1361	2074	52,39	4	5	20
3	1263	1571	24,39	4	2	20
4	1416	1592	12,43	4	1	20
5	1489	1595	7,12	4	1	30
6	1329	1352	1,73	4	1	50
7	1185	1213	2,36	8	1	50
8	1445	1505	4,15	16	1	50

As shown in Table 5, changing only two parameters make noticeable difference in precision and Tabu (taboo) list length.

Decreasing precision to 1% causes improvement of results in comparison to parameters from the first test. Apparently, smaller precision allows TS to make smaller but more frequent steps. This means, that TS explores larger solution space and it is obvious that in this situation the probability of encountering better result is higher. The vital parameter is the number of iterations - along with the same as in genetic algorithm property: the more iterations, the better solution.

D. Comparison of Metaheuristic Algorithms

The next part of research was to compare to CO all three other algorithms with the best inner parameters. Instances of parameters were the same as in previous part (Table 5) apart from minimal percentage passengers served (% passengers) which is variable in this test (Table 6).

TABLE 6. Parameters – set 2.

TS		SA		GA	
Iteration	50	Iteration	50	Iteration	40
Tabu length	4			Population	20
Precision	1			Parents	18

According to Fig. 3 the smallest inaccuracy was found for TS - from 10% to less than 1% for 90% passengers. The second efficient was GA - from more than 50% to about 10%. SA performed as third and the last place for Constructor. Constructor presents the biggest inaccuracy for 50% passengers (almost 250% inaccuracy) but its performance improves when constraint becomes more strict - 90 % passengers. Despite this surprising result, the rest of results leave a lot to wish, either.

ISBN:978-1-61208-114-4

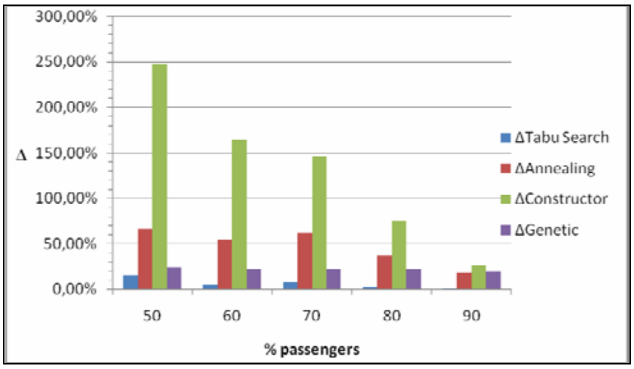


Figure 3. The average inaccuracy.

E. *Comparison – TS vs SA*

The influence of the number of iterations is shown in Fig. 4. The results of this series of experiments justify an observation (rather obvious) that the number of iterations has significant impact on the obtained results. The more iterations, the better results is given by the algorithm. The main useful observation is that TS gives better results than SA regardless of the number of iterations, thus TS algorithm may be recommended for searching the optimal route.

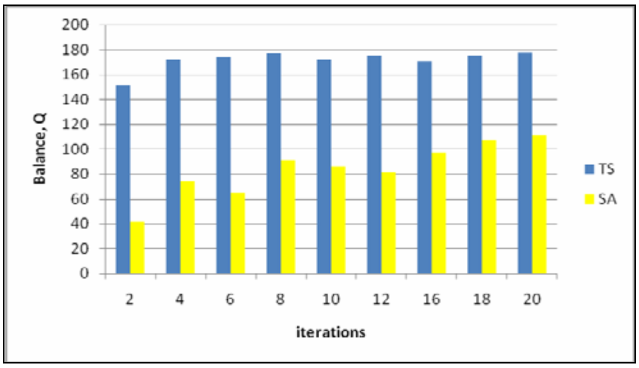


Figure 4. Comparison of alghoritms: TS vs SA.

F. *Comparison – TS vs GA*

The impact of the number of iterations is shown in Fig. 5.

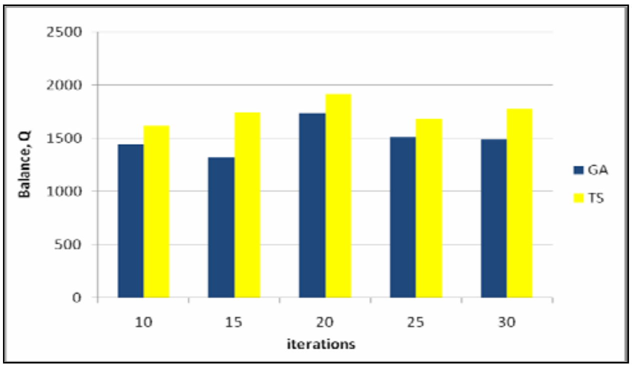


Figure 5. Comparison of algoritms: GA vs TS.

This complex experiment was designed in order to observe differences between the two metaheuristic algorithms: TS and GA, with their best inner parameters apart from parameter - *Iterations* which was a variable. The problem parameters used were such as specified in Table 3.

Similarly to the previous test, TS defeats competitor - GA regardless of iterations number. Although TS wins this competition, the genetic algorithm GA kept pace of TA and the results given by GA were not that bad as in SA case (see Sub-section *E*).

VI. FINAL REMARKS

To sum up, performed research justified the conclusion that TS algorithm gives much better results than SA regardless of defined advanced settings for searching the best solution. SA may give quite good results but much more iterations are needed. The only one algorithm that can compete with TA is GA but the average results of tests show that it would rather never come up with better results than TA. The Constructor algorithm turns out to be the worst and certain improvements are needed to make it somehow useful. Choosing the best algorithm is half the success, however, setting the most appropriate parameters of such algorithm is a vital issue.

According to research presented in this paper, the proposed and recommended algorithm for route planning is Tabu Search (TS).

REFERENCES

[1] J. Skladzien, “Ecological aspects of vehicle transport development”, Opole, 2008 /in Polish/.

[2] M. Gendreau, “An Introduction to Tabu Search”, Universite de Montreal, 2003.K. Elissa, “Title of paper if known,” unpublished.

[3] F. Glover, “Tabu Search – part I”, ORSA Journal on Computing, vol. 1, no. 3, 1997.

[4] F. Glover and G. A. Kochenberger, “Handbook of Metaheuristics”, Springer, Heidelberg, New York, 2002.

[5] V. Granville, M. Krivanek, and J. P. Rasson, “Simulated Annealing: a proof of convergence”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, 1994, pp. 652-656.

[6] S. Kirkpatrick, C. D. Gelatti, and M. P. Vecchi, “Optimization by Simulated Annealing”, Science, vol. 220, 1983, pp. 671-680.

[7] J. M. Laarhoven, H. Emile, and L. Aarts, ”Simulated Annealing: Theory and Applications”, Springer, Berlin, 1987.

[8] L. D. Davies, “Genetic Algorithms and Simulated Annealing”, Morgan Kaufmann Publ., 1987.

[9] H. Youssef and S. M. Sait, “Iterative Computer Algorithms with Applications in Engineering”, Washington., 1997.

[10] D. Ohia, L. Koszalka, and A. Kasprzak, “Evolutionary Algorithm for Congestion Problem in Computer Networks”, Springer, Lecture Notes in Artificial Intelligence, vol. 5711, 2009, pp. 113-122.

[11] A. Kasprzak, “Packet Switching Wide Area Networks”, WPWR, Wroclaw, 1997 /in Polish/.

[12] L. Koszalka, D. Lisowski and I. Pozniak-Koszalka, “Comparison of Allocation Algorithms with Multistage Experiments”, Lecture Notes in Computer Science, vol. 3984, Springer, 2006, pp. 58-67.

[13] P. Wroblewski, “Algorithms: data structure and programming technologies”, WNT, Warsaw, 2003 /in Polish/.

Reconstruction Quality of Congested Freeway Traffic Patterns from Probe Vehicles Based on Kerner's Three-Phase Traffic Theory

Jochen Palmer
IT-Designers GmbH
Entennest 2
D-73730 Esslingen, Germany
Email: jochen.palmer@it-designers.de

Hubert Rehborn
Daimler AG
GR/PTF - HPC: 050-G021
D-71059 Sindelfingen, Germany
Email: hubert.rehborn@daimler.com

***Abstract*—This paper discusses the reconstruction quality of spatio-temporal congested freeway traffic patterns depending on the information provided by different equipment rates of probe vehicles. Vehicles in spatio-temporal congested traffic patterns experience a sequence of accelerations and decelerations. To enable vehicles and vehicular assistance applications to react on the traffic conditions, high quality traffic information is required, i.e. a high quality reconstruction of spatio-temporal congested traffic patterns. The paper uses Kerner's three-phase traffic theory which distinguishes two different phases in congested traffic: synchronized flow and wide moving jam. This theory explains empirical traffic breakdown and resulting spatio-temporal congested traffic patterns. In the presented approach spatio-temporal congested traffic patterns are reconstructed from intelligent probe vehicle information generated by an onboard traffic state detection, identifying traffic states along a vehicle's trajectory at any time. With a data fusion algorithm combining the data of several probe vehicles a detailed picture of spatio-temporal congested traffic patterns is revealed. The quality of the reconstructed congested traffic patterns is assessed by introducing quality indices for (i) travel time, (ii) regions of synchronized flow and wide moving jams as well as (iii) fronts of synchronized flow and wide moving jams. The indices are evaluated by examining a congested traffic pattern with varying probe vehicle equipment rates. Comparing ground truth with the reconstructed traffic pattern shows that a reconstruction quality sufficient for some ITS applications is achievable with probe vehicle equipment rates of about 0.50 %.**

***Keywords*-Traffic monitoring; Traffic state detection; Traffic data fusion; Three-phase traffic theory; Traffic data quality**

I. Introduction

Nowadays congested traffic on highways is still a major problem with severe implications for personal life and the economy. In recent years congested traffic data was mostly gathered with stationary loop detectors. It is expensive to equip a road network with detectors of high quality and small detector distance. Recent progress in mobile communication technology, like WLAN and 3G/UMTS, allows traffic data to be gathered by probe vehicles. This paper tries to answer the question how many probe vehicles are needed to deliver a quality of traffic data comparable to a dense, high quality detector network.

At first Kerner's Three Phase Traffic Theory and its impact on vehicles is described. Then the proposed method for traffic pattern reconstruction and quality indices are introduced. The paper closes with results and conclusion.

II. Kerner's Three Phase Traffic Theory

A. Elements of Three Phase Traffic Theory

Based on extensive traffic data analyses of available stationary detector measurements spanning several years Kerner discovered that in addition to free flow (F) two different traffic phases must be differentiated in congested freeway traffic: synchronized flow (S) and wide moving jam (J) ([1], [2]). Empirical macroscopic spatio-temporal objective criteria for traffic phases as elements of Kerner's three-phase traffic theory ([1], [2]) are as follows:

1) A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic state or freeway bottleneck.
2) In contrast, the downstream front of the synchronized flow phase is often fixed at a freeway bottleneck and does not show the characteristic features of wide moving jams.

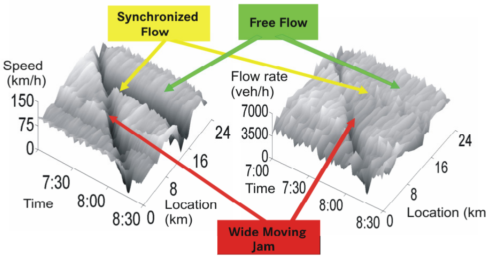


Figure 1. Explanation of traffic phase definitions from empirical data: Spatio-temporal overview of speed (left) and flow rate of traffic (right) on a selected freeway section

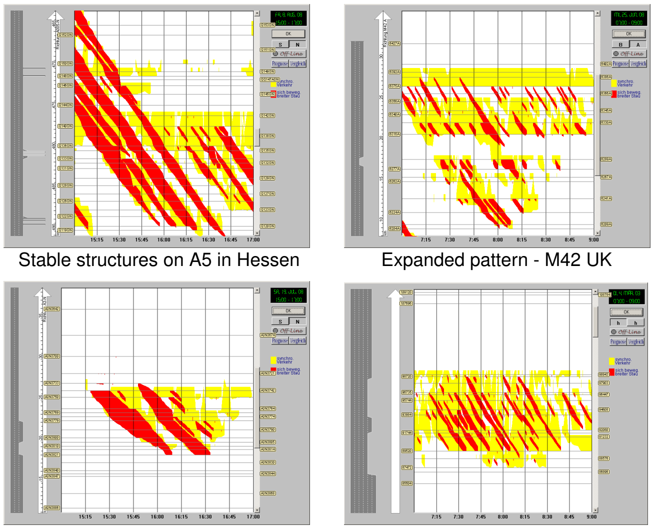
However, neither the observation of speed synchronization in congested traffic nor other relationships and features of congested traffic measured at specific freeway locations (e.g., in the flow-density plane) are a criterion for the phase differentiation. The clear differentiation between the synchronized flow and wide moving jam phases can be made on the above objective criteria 1) and 2) only.

Figure 1 illustrates a vehicle speed and flow profile over time and space based upon real measured traffic data. A wide moving jam propagates upstream as a *low speed valley* through the freeway stretch. In contrast, a second speed valley is fixed at the bottleneck location: this congested traffic phase belongs to the synchronized flow phase.

B. Spatio-Temporal Congested Traffic Patterns

The distribution of traffic phases over time and space on a road represents a spatio-temporal congested traffic pattern. Kerner's three-phase traffic theory is able to explain all empirically measured traffic patterns on various roads in many different countries. For recognition, tracking and prediction of the spatio-temporal congested traffic patterns, based on stationary loop detectors, the models ASDA (Automatische Staudynamikanalyse; automatic congestion analysis) and FOTO (Forecasting of Traffic Objects) ([1], [2]) have been proposed by Kerner based on the key elements of the theory.

Nowadays the models ASDA and FOTO are deployed in the federal state of Hessen and in the free state of Bavaria where they perform online processing of data. In addition they have been successfully used in a laboratory environment on the M42 near Birmingham, UK and the I-405 in California, USA (Figure 2).



Stable structures on A5 in Hessen

Expanded pattern - M42 UK

A3 in Northern Bavaria

I405 in Orange County (LA) - USA

Figure 2. Resulting empirical spatio-temporal traffic patterns when applying the models ASDA and FOTO to traffic data measured in different countries [6]

III. Impacts for Vehicles and Vehicular Assistance Application

Vehicles driving through a spatio-temporal congested traffic pattern experience a number of traffic state changes. A traffic state change represents a unique and exact position in time and space where the traffic phase changes, e.g., from free flow to wide moving jam. It is experienced at any position, where a vehicle hits the upstream or downstream front of a region of a traffic phase. In contrast a traffic phase transition represents the start point or the end point, respectively, of a traffic phase and occurs only twice for each region of a traffic phase (see figure 3).

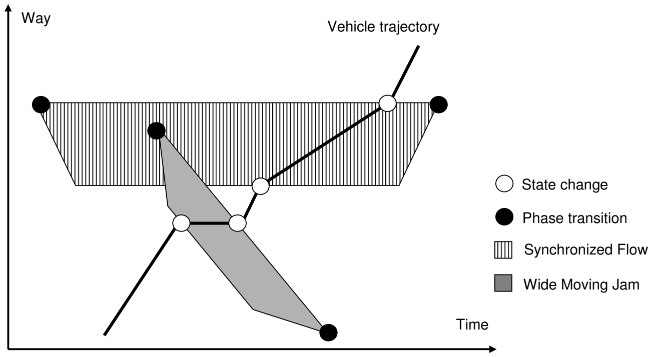


Figure 3. Qualitative explanation of traffic phase transitions and traffic state changes a vehicle experiences on its way through a spatio-temporal congested traffic pattern

Traffic state changes between the different traffic phases have impacts of different strength on the vehicle and vehicular assistance applications. Two of the most distinguishing parameters of different traffic phases from a vehicle's perspective are the vehicle speed v and the vehicle density ρ. Both parameters influence the ability of the vehicles to choose their driving speed as well as the possibility for them to overtake other vehicles and to freely choose their driving lane. Different traffic state changes have a different effect on the value of these parameters. Table I and table II show the expected changes for the possible speed v and the vehicle density ρ, respectively.

State	to F	to S	to J
F	-	Deceleration	Strong deceleration
S	Acceleration	-	Deceleration
J	Strong acceleration	Acceleration	-

Table I
VEHICLE SPEED: CHANGE DEPENDING ON SPECIFIC TRAFFIC STATE TRANSITIONS

State	to F	to S	to J
F	-	Increase	Strong increase
S	Decrease	-	Increase
J	Strong decrease	Decrease	-

Table II
VEHICLE DENSITY: CHANGE DEPENDING ON SPECIFIC TRAFFIC STATE TRANSITIONS

Vehicular assistance applications, like adaptive cruise control or hybrid engine control depend on and benefit from the knowledge of the current and in some cases future values of these parameters, as each traffic state change represents a control and parameter adaption point for these applications.

IV. Test and Simulation Environment

The Kerner-Klenov microscopic three-phase traffic model ([1], [2]) has been used for the generation of a large number of single vehicle trajectories. As input data for the model a description of the simulated track as well as initial starting conditions of speed and flow at the most upstream border are required. All other areas in space and time are governed by the Kerner-Klenov model. The microscopic model's output can be regarded as a realization of *ground truth* which is qualitatively comparable to spatio-temporal congested traffic patterns measured on highways ([1], [2]).

Ground truth means that the model output represents the reference information or the *reality* which should be reconstructed by the vehicle trajectories. Quality is therefore measured as the difference between ground truth and the cooperative reconstruction of spatio-temporal congested traffic patterns based on the generated trajectories.

First a traffic state detection is performed in each virtual vehicle, in order to detect all traffic state changes this virtual vehicle experiences. After that all traffic state changes are combined to a reconstructed congested traffic pattern by applying a clustering algorithm. It combines the autonomously detected traffic state changes of several probe vehicles to a collective and cooperatively reconstructed traffic pattern (see figure 4).

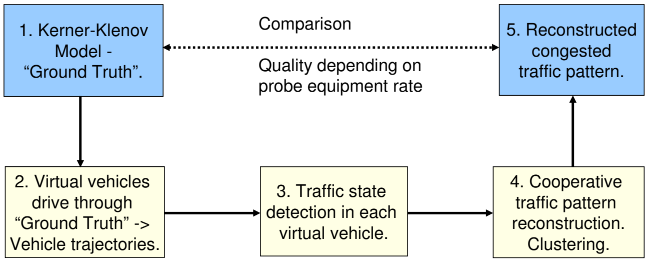

1. Kerner-Klenov Model - "Ground Truth".
Comparison
Quality depending on probe equipment rate
5. Reconstructed congested traffic pattern.
2. Virtual vehicles drive through "Ground Truth" -> Vehicle trajectories.
3. Traffic state detection in each virtual vehicle.
4. Cooperative traffic pattern reconstruction. Clustering.

Figure 4. Steps necessary for probe equipment rate investigations

A. Traffic State Detection in Autonomous Vehicle

Instead of stationary loop detectors, probe vehicle data is used for the detection and reconstruction of spatio-temporal congested traffic patterns. Many systems using probes transmit only aggregated travel times for pre-defined road sections. Here we are not only interested in the travel time losses caused by spatio-temporal congested traffic patterns, but also in their detailed structure (figure 3, [1], [2]).

In Kerner's three-phase traffic theory there is one phase of free flow (F) and two phases of congested traffic, synchronized flow (S) and wide moving jam (J). Each traffic pattern consists of a unique formation and behavior of regions in time and space belonging to exactly one of these three phases. One of these three phases is assigned to each of the probe positions in time and space in an autonomous way ([2], [3]). A traffic state change is performed when the chosen measured values are above or below specific thresholds in speed and time, which are chosen according to microscopic traffic criteria [2].

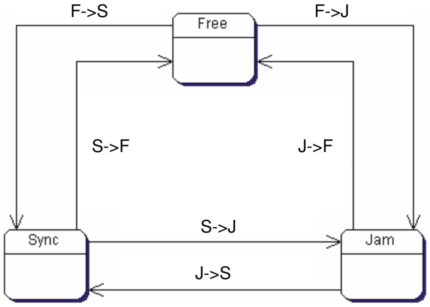


Figure 5. State diagram for three traffic phases [3]

B. Cooperative Reconstruction of Spatio-Temporal Congested Traffic Patterns

For the reconstruction of spatio-temporal congested traffic patterns a clustering algorithm is employed. First the traffic phase is identified, then depending on the identified traffic phase, an algorithm tailored for the specific characteristic features of the synchronized flow and wide moving jam traffic phases is applied as shown in figure 6.

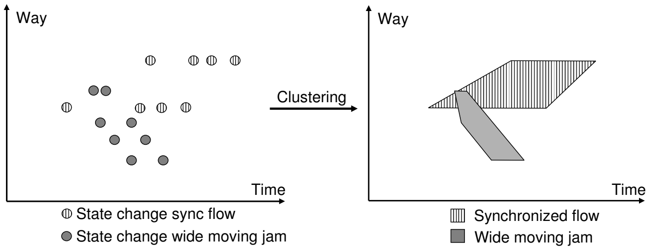


Figure 6. Using clustering to reconstruct traffic patterns from traffic state changes [6]

V. Reconstruction Quality Indices

In order to assess the quality of reconstructed spatio-temporal congested traffic patterns, the reconstructed pattern is compared with a ground truth version of this pattern, which represents the best known information about this particular situation. Ground truth traffic patterns are either directly measured with high quality detection systems or generated by a suitable simulation environment as described above. Common to both approaches is that the ground truth information with the highest possible quality is a continuous information in both time and distance.

A continuous traffic reality assigns a traffic state TS to each spatio-temporal position $P(x, t)$, whereas x represents a continuous value in distance and t a continuous value in time. The traffic reality R represents the combination of all

$$TS(x,t) \in \{F, S, J\} \tag{1}$$

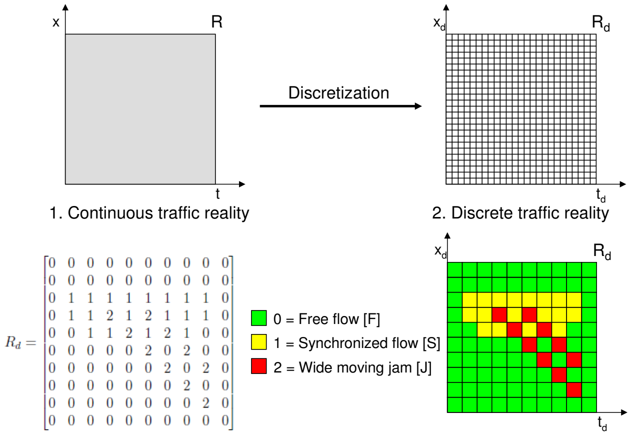
within the borders x and t, hence

$$R := \{TS(x,t) | x_s \leq x \leq x_e \wedge t_s \leq t \leq t_e\} \tag{2}$$

Compared to R a reconstruction of traffic patterns using a model M in most cases shows deviations in time and space. The quality Q of the reconstruction is given by the deviation D between M and R. Hence

$$M := \{TS(x,t) | x_s \leq x \leq x_e \wedge t_s \leq t \leq t_e\} \tag{3}$$

$$Q = D_{M \rightarrow R} \tag{4}$$



$$R_d = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 2 & 1 & 2 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 2 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 7. Discretization of spatio-temporal congested traffic patterns

Currently, it is difficult to measure, store and process continuous spatio-temporal information. Therefore, usually discrete values are used, which average time and distance in discrete intervals. Consequently x and t degrade to the discrete values x_d and t_d. Their resolution represents the quality of knowledge about the spatio-temporal information. Using x_d and t_d the continuous traffic reality R becomes the discrete traffic reality R_d as shown in figure 7.

$$\begin{aligned} R_d := \{TS_d(x,t) | x_s \leq x_0, x_1 ... x_n \leq x_e \\ \wedge t_s \leq t_0, t_1 ... t_n \leq t_e\} \end{aligned} \tag{5}$$

In order to determine the discrete quality Q_d a discrete model output M_d with the same spatial and temporal resolutions x_d and t_d is required. For Q_d this leads to

$$Q_d = D_{M_d \rightarrow R_d} \tag{6}$$

In the following a reconstruction is used synonymously with a model M_d while the known ground truth information used as a comparison reference is used synonymously with the reality R_d (see figure 8).

A. *Quality Index for Regions of Synchronized Flow and Wide Moving Jam*

The quality index for regions of a specific traffic phase assesses the general reconstruction quality between M_d and R_d. For each congested traffic phase the spatio-temporal areas they cover are compared against each other with regard to hits and false alarms. Figure 8 illustrates this concept, which is an adaption of the ROC (Receiver operating characteristic) analysis of traffic data presented in [7] and [8] to traffic data containing three distinct traffic phases.

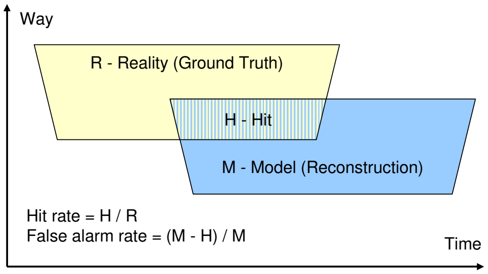

Way
R - Reality (Ground Truth)
H - Hit
M - Model (Reconstruction)
Hit rate = H / R
False alarm rate = (M - H) / M
Time

Figure 8. Definition of the Quality Index for Regions of Synchronized Flow and Wide Moving Jam

The definitions of M_d and R_d given in the previous section can directly be applied to yield a computationally efficient calculation of this quality index. For both M_d and R_d three binary matrices are calculated, one for each traffic phase, which contain a 1 at all spatio-temporal positions where a specific traffic phase occurs and a 0 at all other positions.

For R_d this leads to R_d^F, R_d^S and R_d^J and for M_d to M_d^F, M_d^S and M_d^J respectively. Using a row vector r_v and a column vector c_v with correct dimension for M_d and R_d with all $a_{ij} = 1$ allows to count the number c of elements a_{ij} within matrix A for which $a_{ij} = 1$.

$$c = r_v * A * c_v \tag{7}$$

Using this operation leads to the quality index for hits in regions of synchronized flow,

$$H_{Sync} = \frac{r_v * (M_d^S \wedge R_d^S) * c_v}{r_v * R_d^S * c_v} \tag{8}$$

the quality index for false alarms in regions of synchronized flow,

$$F_{Sync} = \frac{r_v * (M_d^S \wedge R_d^J + M_d^S \wedge R_d^F) * c_v}{r_v * M_d^S * c_v} \tag{9}$$

the quality index for hits in regions of wide moving jam,

$$H_{Jam} = \frac{r_v * (M_d^J \wedge R_d^J) * c_v}{r_v * R_d^J * c_v} \qquad (10)$$

and the quality index for false alarms in regions of wide moving jam.

$$F_{Jam} = \frac{r_v * (M_d^J \wedge R_d^S + M_d^J \wedge R_d^F) * c_v}{r_v * M_d^J * c_v} \qquad (11)$$

All of these quality indices yield a number between 0 and 1, whereas 0 represents the minimum value and 1 the maximum value.

B. Quality Index for Travel Time

For each point in time t the total travel time T_{total}, consisting of free flow travel time and congested travel time, can be calculated as shown in [4]

$$T_{total}(t) = \frac{L_F(t)}{v_F(t)} + \frac{L_J(t)}{v_J(t)} + \frac{L_S(t)}{v_S(t)} \qquad (12)$$

where $L_F(t)$, $L_S(t)$ and $L_J(t)$ represent the lengths of the traffic phases and $v_F(t)$, $v_S(t)$ and $v_J(t)$ represent the average speeds within these phases at time t. The relative average deviation ΔT between the travel time reported by the reconstruction T_M and the travel time T_R reported by the reference ground truth information is now given by [5].

$$\Delta T = \frac{1}{N} \sum_{t=1}^{N} \frac{|T_M(t) - T_R(t)|}{T_R(t)} \qquad (13)$$

As a result ΔT is a number ≥ 0, whereas a result of $\Delta T = 0$ represents the best possible quality.

C. Quality Index for Fronts of Synchronized Flow and Wide Moving Jam

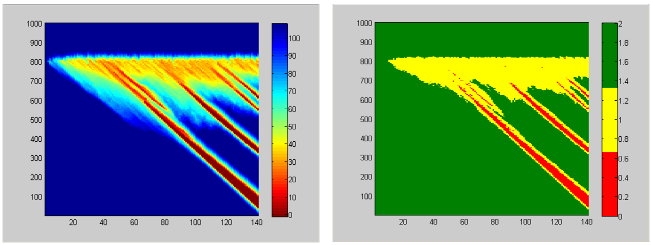
The quality index for fronts of synchronized flow and wide moving jam calculates the average deviation for each point in time t.

The deviation is calculated between the front position reported by the model M and the front position reported by the ground truth information R for both the upstream and downstream front. This results in Δup and $\Delta down$ respectively.

$$\Delta up = \frac{1}{N} \sum_{t=1}^{N} |pos_M^{up}(t) - pos_R^{up}(t)| \qquad (14)$$

$$\Delta down = \frac{1}{N} \sum_{t=1}^{N} |pos_M^{down}(t) - pos_R^{down}(t)| \qquad (15)$$

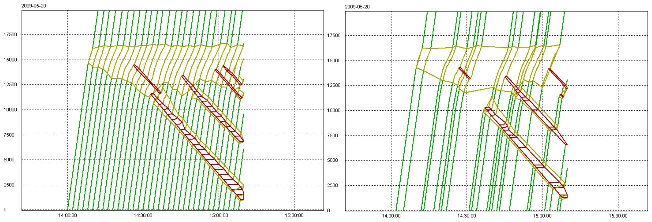
Again both Δup and $\Delta down$ give a result ≥ 0 m whereas a result of 0 m represents the best possible quality.



Speed

Traffic State

Figure 9. Kerner-Klenov simulation output of a general pattern - Ground Truth



0.75% probe density, uniformly distributed

0.50 % probe density, randomly distributed

Figure 10. Example results for uniform and random distributions of probe vehicles

VI. Results

A. Examined Situation

For evaluation the Kerner-Klenov three-phase microscopic model was used to simulate a spatio-temporal congested traffic pattern. The simulation was configured to simulate a 20 km highway stretch with 2 lanes and one on-ramp. By increasing the incoming flow an on-ramp bottleneck was established, leading to an $F \rightarrow S$ traffic breakdown and the formation of a region of synchronized flow. Within the synchronized flow region several wide moving jams emerged (figure 9).

Example results of the traffic state detection and clustering processing based on probe vehicle data are shown in figure 10. In the following the quality index results of this process are compared with results of the ASDA/FOTO models. ASDA/FOTO was used with a high quality stationary loop detector network having detector distance of 1-2 km.

B. Quality Index for Regions of Synchronized Flow and Wide Moving Jam

Method	Det./Veh.	H_{Sync}	F_{Sync}	H_{Jam}	F_{Jam}
ASDA/FOTO	1 km Det. dist.	0.83	0.12	0.81	0.24
ASDA/FOTO	2 km Det. dist.	0.77	0.16	0.65	0.25
Probe Veh.	1.50 % uniform	0.89	0.17	0.83	0.24
Probe Veh.	0.38 % uniform	0.82	0.24	0.73	0.21
Probe Veh.	0.50 % random	0.78	0.21	0.72	0.22

Table III
QUALITY INDEX RESULTS FOR DIFFERENT METHODS AND DETECTOR DISTANCES / PROBE EQUIPMENT RATES

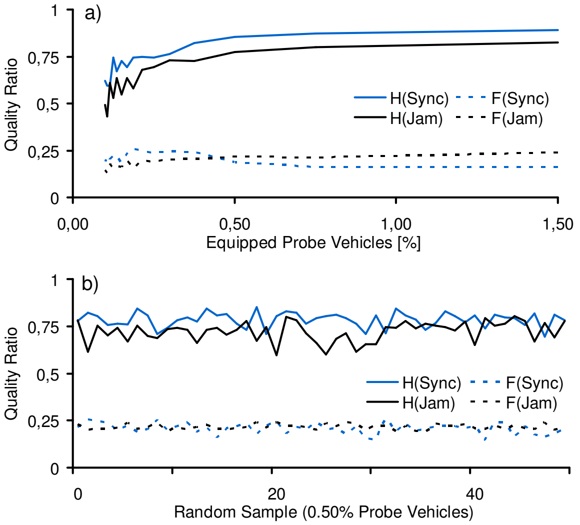


Figure 11. Quality index results for uniform (a) and random (b) probe vehicle distributions

Table III and figure 11 show that the results of the region quality index for a equipment rate of 0.38 % uniformly distributed as well as a equipment rate of 0.50 % randomly distributed are comparable to the established ASDA/FOTO method using detector distances between 1-2 km.

C. Quality Index for Travel Time

Method	Det./Veh.	ΔT
ASDA/FOTO	1 km Det. dist.	0.038
ASDA/FOTO	2 km Det. dist.	0.061
Probe Veh.	1.50 % uniform	0.033
Probe Veh.	0.38 % uniform	0.065
Probe Veh.	0.50 % random	0.053

Table IV
TRAVEL TIME QUALITY INDEX RESULTS FOR DIFFERENT METHODS AND DETECTOR DISTANCES / PROBE EQUIPMENT RATES

The travel time quality index confirms the results of the region quality index (see table IV and figure 12).

VII. CONCLUSION

Probe vehicle equipment rates of 0.38 % uniformly distributed and 0.50 % randomly distributed processed with the proposed method are comparable to detectors with distances of 1-2 km processed with the existing system ASDA/FOTO. In addition the proposed method promises the provision of high quality traffic data on all highways while existing systems rely on stationary loop detectors. Higher probe vehicle equipment rates promise an even higher quality traffic data suitable for future ITS and vehicular assistance applications. Subjects of further ongoing research are the evaluation of the quality index for fronts of synchronized flow and wide moving in combination new vehicular assistance applications as well as system communication and architecture alternatives.

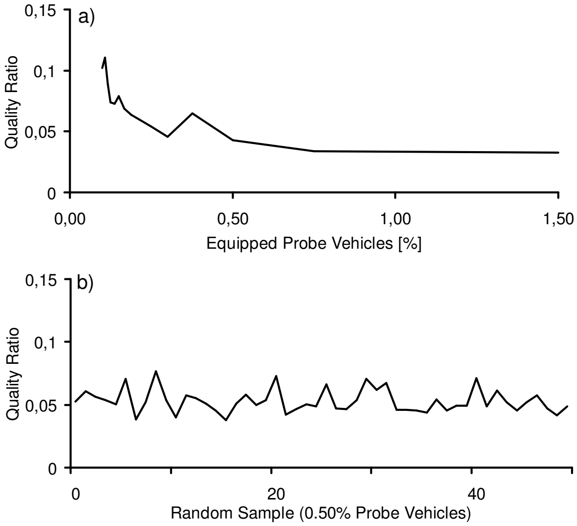


Figure 12. Travel time quality index results for uniform (a) and random (b) probe vehicle distributions

ACKNOWLEDGMENT

The authors would like to thank Boris S. Kerner and Sergey L. Klenov for support regarding the microscopic three-phase traffic model and many fruitful discussions.

REFERENCES

[1] B. S. Kerner, *The Physics of Traffic* Berlin, New York: Springer, 2004.

[2] B. S. Kerner, *Introduction to Modern Traffic Flow Theory and Control* Berlin, New York: Springer, 2009.

[3] B. S. Kerner, S. L. Klenov, J. Palmer, M. Prinn and H. Rehborn, *German Patent Publication DE 10 2008 003 039 A1, 2008.*

[4] H. Rehborn and J. Palmer, *ASDA/FOTO based on Kerner's Three-Phase Traffic Theory in North Rhine-Westphalia and its Integration into Vehicles* 2008 IEEE Intelligent Vehicles Symposium, Eindhoven, 2008.

[5] J. Palmer and H. Rehborn, *Reconstruction of Congested Traffic Patterns Using Traffic State Detection in Autonomous Vehicles Based on Kerner's Three Phase Traffic Theory* 16th World Congress on ITS, Stockholm, 2009.

[6] J. Palmer and H. Rehborn, *Vehicular Assistance Applications in the Scope of Kerner's Three-Phase Traffic Theory* Networks for Mobility - 5th International Symposium, Stuttgart, 2010.

[7] K. Bogenberger, *Qualitaet von Verkehrsinformationen* Strassenverkehrstechnik, Kirschbaum Verlag GmbH Bonn, 10:518 - 526, 2003.

[8] S. Lorkowski, *Fusion von Verkehrsdaten mit Mikromodellen am Beispiel von Autobahnen* PhD Thesis, Technical University of Berlin, 2009.

Innovation Process – Managing Complexity through Financial Aspects

Borut Likar
Faculty of Management
University of Primorska
Koper, Slovenia
borut.likar1@guest.arnes.si

Peter Fatur
Faculty of Management
University of Primorska
Koper, Slovenia
peter.fatur@fm-kp.si

Abstract - **The research investigates one of the most complex parts of the business system – innovation process. The relationship between the company's innovation inputs and its innovation performance was studied on a sample of 2503 companies from manufacturing and selected service sectors. The research was based on official Eurostat statistical data. As these data were essentially not adapted for analysis of companies' innovation performance, the methodology for companies' innovation-business performance was developed. The groups of non-innovators, innovation followers and innovation leaders were formed. In the group of leaders, ROE is 40% higher as regards the group of innovation followers. Each euro invested into innovations yields 13.9 Euros in the same group. But an increase in investments is related with the growth of productivity of invested assets only in the group of followers. Interestingly, increasing a portion of »breakthrough« innovations reduces the productivity of these assets in the group of leaders. On the basis of these findings recommendations are provided as to which policies of innovation investments should different types of innovation followers adopt so as to catch up with the group of innovation leaders.**

Keywords – innovation; R&D; technology; economic performance; productivity; complex systems

I. Introduction

The research investigates one of the most complex parts of the business system – the innovation process and its performance. Community Innovation Survey [1] represents the basic statistical instrument for innovation performance measurement in the EU countries. Its methodology is standardized and it is relatively well known among respondents/companies and a number of respondents is very high (as filling out the questionnaire is compulsory). However, the methodology was prepared for benchmarking primarily at the country level. Therefore, the data are not very constructive for studies focused on input-output innovation relations. One of the basic aims of our research was to develop a methodology which would enable the use of EU statistical data connected with innovation-business performance. Besides, our research is oriented towards identifying differences of influential factors as to the innovation/business performance.

The paper is organized as follows. Section II introduces the theoretical principles of innovation measurement and its limitations. Section III is the methodology. In section IV the results of the analysis are presented and findings are discussed. Section V is the conclusions, where findings are summarized and the innovation policy recommendations are given.

II. Literature survey and limitations

The survey providing the core data for our research is the most recent Community Innovation Survey [2] for Slovenia. Literature addresses several approaches to the monitoring of innovation. One of the fundamental methodological approaches is the analysis of input (investments), process and output (results) groups of indicators. The selection of indicators proves extremely diversified. Expenditure for research and development activities [3] or a number of days dedicated to education/training of employees [4] are used as input indicators representing "investments" in organizational system. Process indicators help us establishing the state of innovation process management (organization, planning, management, and supervision). Output indicators reflect the results of innovation processes, for example the number of patents and new market products, market share, revenues from sales of innovations/innovative products and suchlike [5].

Various researches discuss the relation between innovation strategy and economic successfulness of an organization. Many of them show that the connection is positive yet weak [6]. The researches highlight also the importance of strategic decision to innovate in achieving economic results [7, 8]. Besides, the proportion of intramural expenditure on R&D is supposed to have an extremely valuable influence, which is manifested in an improvement of product quality [9].

In the mentioned studies, we face the problem of defining process and output indicators. The role of the innovation process is often not clearly defined – it is a result of inputs (e.g., financial inputs) or it could be treated like the innovation input. Besides, outputs often represent an indirect output (e.g., number of patents, new products etc.) which do not obviously lead to improved business performance. Therefore, we took into consideration the inputs, which represent the basic step towards mastering the innovation process and those outputs which clearly present the financial situation.

III. Methodology

The research is focused on 2503 Slovenian companies from manufacturing and selected service sectors. The Slovenian contribution to the CIS 2006 survey includes data for the period from 2004 to 2006 on the enterprises' product (goods or service), organizational and process innovations, innovation activities and expenditures, co-operation in innovation and the effects of innovation. In addition, company's financial data (balance sheet, profit and loss account and some key financial ratios) was collected from the official statistical database on companies (the Agency of the Republic of Slovenia for Public Legal Records and Related Services), while the third statistical database (Statistical Register of Employment - SRDAP) provided data on the educational structure of employees. The first challenge was developing a methodology which could enable measuring the relation between company's innovation inputs and innovation performance.

The two key variables that represent a measurable output from the innovation process were defined as: RII ("Index of revenues from innovation"), i.e., a proportion of total turnover resulting from innovations (either new to the market or new to the company only), and RMI ("Index of revenues from market innovation"), i.e., a ratio of turnover from innovations new to the market to total innovation turnover.

In the following phase of the research, the companies were divided into 5 groups. The first one is a group of companies (group 0) having no revenues from innovations (RII=0). The groups recording any revenues arising from innovations (RII>0) were divided into four groups pursuant to the value of indices RII and RMI. As a limit of division the medians were set, thus ensuring equal representativeness of companies across all four groups. Dividing the RII/RMI matrix in points predefined with both medians, 4 quadrants are obtained.

After the development of the innovation performance matrix, the relationship between the innovation and business performance was explored. The 5 innovation groups were in pairs compared using nonparametric tests (Mann-Whitney and Kruskal Wallis Tests) so as to establish in which variables the groups significantly differ. The variables in comparison are the ones included in the CIS 2006 survey (innovation related variables), supplemented by the companies' business performance variables.

The same statistical method (Mann-Whitney nonparametric test) was applied to the innovation investments variables to compare the proportions of financial assets that the companies appertaining to a particular group invest into innovation and how efficiently do these companies turn such investments into revenues arising from innovations, i.e., to test the extent and productivity of innovation investments.

IV. Results and Discussion

A. *The innovation performance matrix development*

As demonstrated the values of the two primary output indices of the invention-innovation process (RII and RMI) served as criteria for grouping the companies. These values are indicated in Table I separately for non-innovative (0) and innovative (1 & 2) companies (these are further divided into 4 subgroups 1a, 1b, 1c and 2); see Figure 1. Accordingly, the innovation leaders are companies having a high portion of turnover from innovations and a high portion of turnover from "radical" innovation in total innovation turnover (high RII and high RMI). Non-innovators is a group (0) of companies having RII/RMI=0 – no turnover based on innovation.

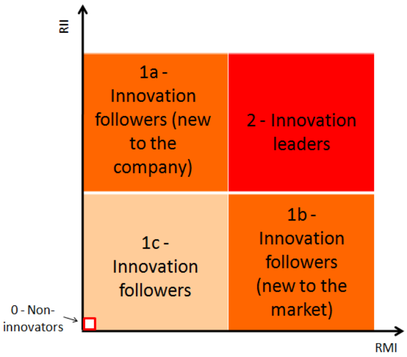

RII
1a - Innovation followers (new to the company)
2 - Innovation leaders
1c - Innovation followers
1b - Innovation followers (new to the market)
0 - Non-innovators
RMI

Figure 1. RII/RMI matrix

B. *Relationship between the innovation and business performance*

So far the method of dividing the companies active in the field of innovation into 5 groups on the basis of RII and RMI indicators has been demonstrated. Companies classified in different groups differ at least pursuant to the volume and structure of their revenues from innovation, i.e., direct results of innovation recorded on the market. However, differences in the revenues from innovation do not necessarily indicate also differences in the companies' business performance. Furthermore, revenues arising from innovation do not provide any conclusions as regards the organisation of innovation systems in the companies.

Let us therefore have a look as to whether there are any differences among the groups regarding their business performance. The Mann-Whitney rank sum test is applied to compare the financial ratios for the groups 0, 1 (a, b, c) and 2. As it compares medians, which are insensitive to outliers compared to means, the Mann–Whitney test is less likely to spuriously indicate significance than the t-test because of the presence of outliers – i.e., Mann–Whitney is more robust.

We would like to focus on most important performance indicators – on those, where the Mann-Whitney test showed significant differences (Table I); these were observed between groups 0-1 and simultaneously between 0-2. It is very interesting, that the Group 1 (even though with better innovation results concerning RII/RMI compared to Group0) is performing in total (as to ROE and average growth of net revenues) worse than the non-innovators (Group 0). On the other hand the company's financial performance of Group 2 is much better; ROE proves to be 40% higher in the group of innovation leaders that among followers and non-innovators, the average growth of net revenues is 33% higher compared to group 0 and even 41% better than the innovation followers (Group 1).

TABLE I. BASIC STATISTICS FOR INDICES RII AND RMI AS PER INNOVATION GROUPS.

Group	N	Variable	Median	Mean
Non-innovators (0)	1790	RII	0.00	0.00
		RMI	-	-
Innovators in total (1 & 2)	713	RII	20.00	29.31
		RMI	40.00	41.89
Innovation followers (1c)	206	RII	10.00	9.39
		RMI	0.00	6.60
Innovation followers (1b)	195	RII	10.00	10.59
		RMI	100.00	79.39
Innovation followers (1a)	163	RII	48.00	52.42
		RMI	1.01	11.83
Innovation leaders (2)	149	RII	50.00	56.05
		RMI	71.43	74.48

At the same time the revenues appertaining to the group 2 recorded between 2006 and 2007 increased by 7%, while the average revenues from 2003 to 2007 increased by 41%.

Similar relations may be observed between the groups 0 and 2; additionally, statistically significant differences in return on sales (ROS) and return on assets (ROA) may be observed. Companies appertaining to the group 2 pay out 6% higher salaries than the companies in the group 0.

On the basis of the aforementioned findings it may be concluded that the companies appertaining to the group 2, which in comparison with the groups 0 and 1 innovate more successfully (achieve higher values of RII and/or RMI), record also better company's performance assessed with the financial ratios.

C. *Extent and productivity of innovation investments*

What portions of financial assets do companies appertaining to a particular group invest into innovation and how efficiently do these companies turn such investments into revenues arising from innovations? To answer these two questions the group of followers (1) shall again be examined by dividing it into three subgroups as indicated in Figure 1, i.e., 1a, 1b and 1c. A company of type 1a may enter into the category of innovation leaders (2) by increasing its RMI; a company of type 1b by increasing its RII, and a company of type 1c by increasing both RII and RMI. In order to make the examination simpler let us suppose that simultaneously only one of both coefficients may be increased, i.e., a path from the subgroup 1c into the group 2 leads either through 1b or through 1a.

Using Mann-Whitney's test it shall be established in which variables the groups 1b and 1a significantly differ from the group 2 and between each other. A comparison with the group of non-innovators (0) is not possible since this group fails to record any costs of innovation. The efficiency of turning the invested assets into revenues arising from innovations – productivity of investments – shall be expressed with a CRIT variable (share of innovation expenditure to total revenues arising from innovations – see also Figure 2), representing a reciprocal value of the productivity of investments which enables its calculation also for the companies which recorded revenues arising from innovation in the period in question (as in the case of groups 1 and 2), yet did not record any innovation costs.

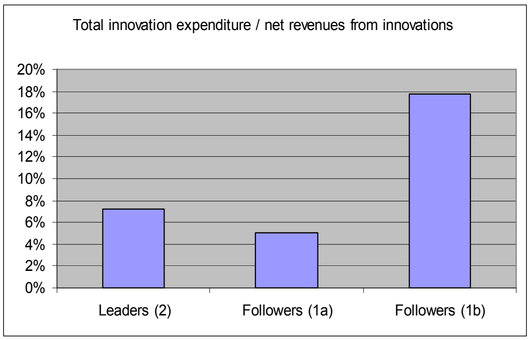


Figure 2. CRIT - Total innovation expenditure / Net revenues from innovations.

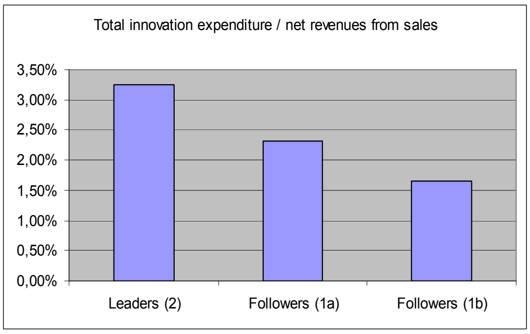


Figure 3. Share of innovation expenditure - Total innovation expenditure / Net revenues from sales.

Table II shows medians of innovation expenditure and productivity of investments (CRIT) for the groups 1a, 1b and 2 (see also Figure 3). Statistically significant correlations are indicated in bold. Comparison of the groups 1a and 2 indicates that the companies appertaining to the group 2 invest significantly more than the companies in the group 1a (innovation costs as a portion of revenues from sales in the group 2 are higher by 1.4).

TABLE II. COMPARISON OF MEDIANS OF THE FINANCIAL RATIOS AS PER GROUPS

Variable	Description	Median			Mann-Whitney test; Sig. (2-tailed)	
		Group 0	Group 1	Group 2	0vs.2	1vs.2
F_03_a	Operating efficiency ratio	1.02	1.02	1.02	0.67	0.90

F_03_b	Return on sales – ROS	0.02	0.03	0.04	**0.02**	0.28
F_03_c	Pre-tax return on sales	0.03	0.03	0.04	**0.04**	0.29
F_03_d	Total revenues per employee (€)	82183	85601	81794	0.90	0.41
F_03_e	fit per employee (€)	1906	2439	3079	0.17	0.53
F_03_f	Gross profit per employee (€)	2332	2905	3490	0.25	0.55
F_03_g	Operating profit per employee (€)	2963	3168	3976	0.37	0.57
F_03_h	Labour costs per employee (€)	17690	18396	18884	**0.01**	0.20
F_03_i	Average salary per employee (€)	12613	12962	13345	**0.02**	0.19
F_03_j	Return on equity – RO	0.10	0.10	0.14	**0.01**	**0.00**
F_03_k	Return on assets – ROA	0.03	0.04	0.05	**0.04**	0.17
F_03_m	Sales-to-Assets	1.33	1.27	1.26	0.21	0.66
F_03_n	Return on equity before taxes – ROEBT	0.13	0.12	0.17	**0.02**	**0.00**
F_04_d	Growth of net revenues 07/06 (%)	3.63	3.60	3.84	0.06	**0.01**
F_04_e	Average growth of net revenues 2003 to 2007 (%)	9.62	9.06	12.81	**0.01**	**0.00**

Regarding the groups 1b and 2, companies appertaining to the group 2 invest more than the group 1b (innovation expenditures in the group 2 are higher by 1.9). Productivity of innovation investments is significantly higher in the group 2 than in the group 1b. The companies in the group 2 succeed in making an average of 13.9 Euros from every Euro invested into innovation (or almost 2.5 times more than the companies in the group 1b), see Table III.

The groups 1a and 1b do not differ regarding the costs yet only as regards the productivity of innovation investments. Productivity of innovation investment is 3.5 times higher in the group 1a than in the group 1b. Each euro invested into innovations yields 19.6 Euros in the group 1a and 5.5 Euro in 1b. This result may be explained by the fact than investing similar amount of financial assets the group 1a on average generates higher revenues from innovation (RII).

Which type of innovation investment policies do companies appertaining to the groups of innovation followers (1a or 1b) need to adopt in order to reach the group of innovation leaders (2)? The company in the group 1a needs to increase investments into innovation in order to increase RMI and thus enter into the group 2 (by factor 1.4 on average at unchanged exploitation of these assets).

So as to increase RII and enter into the group 2 the company in group 1b needs to ensure simultaneous increase in investments and increase in the efficiency of their exploitation. Therefore, a transition through the intermediate level (1a) is in this case reasonable.

Transition from the group 1b into 1a shall not demand an increase in the invested financial assets yet it shall require a substantial increase in the efficiency of exploiting the existing assets (by factor 3.5 on average). The next step, transition from the group 1a into the group 2 shall – on the other hand – demand company's more intensive investments into innovation, at unchanged efficiency of their exploitation.

TABLE III. COMPARISON OF MEDIANS OF INNOVATION EXPENDITURE AND PRODUCTIVITY OF INVESTMENTS AS PER GROUPS

Variable	Description	Median			Mann-Whitney test; Sig. (2-tailed)		
		Group 1a	Group 1b	Group 2	1a vs. 2	1b vs. 2	1a vs. 1b
A_e2_i	Total innovation expenditure / Net revenues from sales	2.3%	1.7%	3.3%	**0.048**	**0.001**	0.205
A_e2_v	Total innovation expenditure / Number of employees (€)	1920	1391	2623	0.059	**0.001**	0.269
A_x1_b	CRIT (total innovation expenditure / revenues from innovations)	5.1%	17.8%	7.2%	0.112	**0.000**	**0.000**
1/A_x1_b	1/ CRIT = Productivity of investments = revenues from innovations/ total innovation expenditure	19.6 EUR	5.6 EUR	13.9 EUR			

V. CONCLUSION

The main findings are summarised hereunder.

A. The innovation performance matrix

As the statistical indicators (Eurostat) were basically not adapted for analysis of companies' innovation performance, the indicators cannot directly serve as a reference for the companies' performance improvement. However, based on the research we defined the innovation performance matrix consisting of five innovation groups (non-innovators, 3 groups of innovation followers and a group of innovation leaders). The matrix is based on two parameters – RII ("Index of revenues from innovation") and RMI ("Index of revenues from market innovation").

B. Crucial financial performance indicators

The return on equity (ROE - net income divided by the shareholder's equity) as a fundamental indicator from the investor's point of view was also proved as an important indicator regarding the research's aims.

Besides, the return on sales (ROS) and the return on assets (ROA) are the output financial indicators which were considered important.

Another important indicator is the average growth of net revenues, measured during the period of four years.

C. Financial results –innovation leaders and followers

The group of innovation leaders innovate more successfully (record higher values of RII and/or RMI) and achieve also better business performance, assessed with the financial ratios. Company's performance in terms of ROE is 40% higher in the group of innovation leaders than in the groups of followers and non-innovators.

Beside ROE, significant differences between the groups 0 and 2 as well as between 1 and 2 were observed regarding the growth of revenues from sales. Average growth of net revenues is 33% higher compared to group 0 and even 41% better than the innovation followers (Group 1).

If the profit is an economic category on which it is possible to influence by way of accountancy, e.g., with an objective to lower taxes, then the "growth of revenues from sales" (indicates entering new markets, increase in market shares on the existing markets or, last but not least, achieving higher selling prices) proves to be a more »factual« category.

Surprisingly, the group of followers (1) – even though with better innovation results concerning RII/RMI compared to non-innovators – is performing worse than the non-innovators (0) as to ROE and average growth of net revenues.

D. *Productivity of innovation investment*

The relationship between the financial investments into innovation and the revenues arising from innovation (productivity of investments) is confirmed only partially. Significant difference may be observed between the groups 1b and 2, yet not between the groups 1a and 2. While the groups 1a and 2 transform their innovation investments into revenues with a similar efficiency (measured as the CRIT variable), the efficiency of both are approx. 3 times higher than the one in the group 1b.

Each euro invested into innovations yields 19.6 Euros in the group 1a and 5.5 Euros in 1b while in group 2 (innovation leaders) this value counts 19.6 Euros.

We can conclude, the impact of innovation investments is quite high, but seems not to be linear; increased investments do not (necessarily) result in an improved innovation performance.

E. *Innovation policy recomendations*

Which policies of innovation investments are thus to be selected by the companies appertaining to the group of innovation followers (1a or 1b) so as to enter the group of innovation leaders (2)? A company appertaining to the group 1a needs to increase investments into innovation in order to increase RMI and thus enter the group 2 (at unchanged efficiency of exploiting these assets). So as to increase RII and enter the group 2 the company 1b needs to ensure simultaneous increase in investments and increase in the efficiency of their exploitation. Therefore, a transition through the intermediate level (1a) is in this case reasonable. Transition from the group 1b into 1a shall not demand an increase in the invested financial assets yet it shall require a substantial increase in the efficiency of exploiting the existing assets (by factor 3.5 on average). The next step, transition from the group 1a into the group 2 shall – on the other hand – demand company's more intensive investments into innovation, at unchanged efficiency of their exploitation.

Acknowledgment

Authors would like to thank the Slovene Research Agency ARRS for the support provided for this research with grants J5—0425 and J2—2039.

References

[1] European Commission, "The Community Innovation Survey 2006", http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/, 2009, Accessed on 25. 11. 2010.

[2] Eurostat, "Fourth Community Innovation Survey", More than 40% of EU27 enterprises are active in innovation Co-operation with customers, Office for Official Publications of the European Communities, Luxembourg, 2007.

[3] R. G. Cooper and E. J. Kleinschmidt, "Winning Businesses in Product Development: The Critical Success Factors," Research-Technology Management 50 (3), 2007, pp. 52-66.

[4] A. A. M. Leenders and B. Wierenga, "The effectiveness of different mechanisms for integrating marketing and R&D," The Journal of Product Innovation Management 19 (4), 2002, pp. 305-317.

[5] M. Michalisin, "Validity of annual report assertions about innovativeness: an empirical investigation," Journal of Business Research, Vol. 53, 2001, pp.151-161.

[6] B. Milfelner and A. Petejan, "Vpliv inovativnosti na uspešnost poslovanja v slovenskih podjetjih; Impact of Innovation on the Successfulness of the Slovenian Companies", Ekonomsko-poslovna fakulteta, Maribor, 2003.

[7] P. Fatur and B. Likar, "Statistical Analysis for Strategic Innovation Decisions in Slovenian Mechanical Industry," Journal of Mechanical Engineering 56, no. 7-8, 2010, pp. 489-496.

[8] V. Potocan and M. Mulej, "Development economics' view on growing entrepreneurship in Slovenia," International journal of entrepreneurship and innovation management 8 (3), 2008, pp. 305-319.

[9] B. Likar, "The influence of innovation, technological and research processes on the performance of Slovenia's woodworking industry," Wood research 53 (4), 2008, pp. 115-120.

A Service-Oriented Monitoring System of Pressurized Air in Industrial Systems for Energy Awareness

Piotr Rabeko, Andrei Lobov, Jose L. Martinez Lastra
Department of Production Engineering
Tampere University of Technology
Tampere, Finland
{piotr.rabeko, andrei.lobov, jose.lastra}@tut.fi

Abstract— **This paper focuses on development of an independent system dedicated to monitoring of compressed-air systems. Main focus has been put on discovery of anomalies which influence energy efficiency of compressed-air systems. System is able to recognize anomalies in an automatic manner and notify the user if these are discovered. These improprieties are leakages, pressure drops and increased humidity in the system. Designed system consists of hardware device which can be connected to any desired point of compressed-air line and is responsible for data acquisition and processing. System offers two Human Machine Interfaces. Additionally one of HMIs exploits Web Service technology enabling major decrease in communication bandwidth usage.**

Keywords - monitoring; compressed-air; leakage; HMI; SOA

I. Introduction

It is estimated that around 90% of manufacturers worldwide use compressed-air systems (CAS). In EU, CAS use about 10% of produced electricity. The particular countries energy usage of compressed-air is around 10-15 TWh annually [12]. Therefore CAS are major contributors to electric power consumption. Kyoto Protocol obliges industrialized countries to significant reduction of greenhouse gasses. Key policy to achieve these norms is energy efficiency. According to [14] only 19% of the power used by a compressor can be transformed to usable air energy. However compressed-air possesses appealing advantages like: safety and simplicity of usage, reliability, easiness of distribution and storage capabilities [12]. Unfortunately many CAS lack energy efficiency due to many factors. The major one is said to be the leakage in the system, being responsible for wasting 20-50% of a compressor's output. The cost of leaks is the cost of energy required to balance the volume of lost air. As can be seen there is a huge potential of energy saving and at the same time cost reduction in that particular area. However most of the research in the field of leak detection focuses on pipe networks, transporting media like petroleum, gases, steam or water. Whereas on the plant level there is evident lack of means to detect compressed-air losses in a quick, continuous and effective manner. Additionally due to the stereotype that compressed-air is cheap and harmless in operation, many compressed air properties are not investigated enough, what may lead to malfunction or destruction of compressed-air powered equipment. Water condensation or pressure drop at particular points of CAS may prove to be essential to monitor in order to extend equipments life cycle and reduce the maintenance costs. Considering that compressed-air is a second after electricity energy source in most of the plants, there is evident shortage of methods to closely monitor its parameters to boost its energy efficiency. Therefore, proposed solution will target area of monitoring systems which use compressed-air energy. Main focus will be put on discovery of three improprieties: leakage, increased humidity and pressure abnormalities.

II. Literature and Technology Review

During the CAS lifetime, operating energy costs may become five times higher than the initial investment [13]. The first step of saving process is to determine where the highest potential for saving is. As stated in [12], [13] in order to improve CAS performance, following measures might result in large savings: reduction of leaks, improved air treatment, avoiding over pressurizing. As can be seen there are many factors that can minimize the energy loss. Study in this paper will target three factors which can contribute to energy savings in CAS. System monitoring leaks, pressure deviations and dew point temperature will be developed.

A. *Losses due to leaks, humidity and pressure changes*

The biggest contributor to losses and the same time potential saving are leaks in the system. According to [12] leaks represent 42% of potential of energy savings in CAS. That figure makes it natural step to firstly tackle with leaks, whenever plans to decrease energy losses in CAS are undertaken. Based on equation presented in [3] losses due to leaks can be directly translated into money loss. For instance one leak of 6mm diameter will bring around 8000€ losses annually. Usually there are more leaks, which depicts a magnitude of possible losses. Apart from high energy losses and costs, leaks may contribute to fluctuations of system pressure - resulting in decreased efficiency or even destruction an air-operated equipment. Additionally the supply equipment may suffer due to unnecessary cycling and altered run time, resulting in shorten life cycle and extra maintenance.

Besides leaks, other functionality threatening factor is the humidity inside air distribution system. Depending on application of air, there is a need of diverse dryness. Air as general consists of mixture of nitrogen, oxygen and water vapour. Unlike nitrogen and oxygen, water vapour concentration is highly variable. According to the Dalton's law the total pressure of a gas is the sum of partial pressures of separate component gases. The maximum partial pressure of water vapour depends strictly on temperature. Adding more vapour will result in condensation of water. To determine the measure of how much water vapour is in the gas, dew point temperature is used. If the air temperature drops below the dew point temperature, it becomes saturated and some of water condenses. In CAS, changing gas pressure changes dew point temperature. The knowledge of dew point temperature has significant influence in CAS as any condensation of water may cause improprieties in the plant. Majority of compressed air distribution systems are made of steel, thus any contact with moisture results in corrosion, which may lead to creation of leaks. Moisture as well as rust blown with air may affect the air-powered equipment, making it sluggish or damaged. Any trace of moisture could be critical in the processes like: paint spraying, pharmacy, food industry causing adherence of hygroscopic products. Usually dew point is measured after the drying process, however possibility of measuring the humidity at different locations may be beneficial as it may happen that part of the distribution system is exposed to low outside temperatures, resulting in water condensation. Additionally accurate knowledge can improve energy efficiency of desiccant dryers which decrease the humidity level of compressed air. In that way savings up to 80% can be achieved in the process of air drying [4], [9].

Besides the previously mentioned issues also the pressure Affect the efficiency of the system. Pressure drops in the system occur mainly due to mechanical obstructions in distribution system or due to treatment of the air – in order to improve its quality. Usually the system should have a drop much smaller than 10% of compressor's output. Drops can result in performance decrease of the equipment and increased energy consumption [6]. While the drops due to mechanical constructions can be planned at the design stage, drops resulting from dirty and clogged filters may increase unnoticed. This can translate to energy and money loss. It is noted that pressure drop of 2 psi may result in 1% increase of energy costs [11]. As an example, annual cost of drop of 0.5 bar with compressor of 75kW performance can reach around 1300 €. More information regarding pressure drop and electricity dependency can be found in [11].

B. *Condition monitoring*

Main monitoring paradigm incorporated in that study is based on concept of condition monitoring. In industry, maintenance is considered to be an essential tool to provide reliable functioning. Maintenance can be divided into several categories, based on the methodologies [15]. These approaches are: breakdown, scheduled, condition monitoring, predictive. Condition monitoring fills the shortcomings of breakdown and scheduled approaches. Condition monitoring is a supplementary approach filling the drawbacks of scheduled approach and offering constant monitoring in order to avoid failures. It employs collection of data about certain part of equipment or system in order to assess if its state or performance is within or likely to remain in predefined limits. It mainly concerns with failures which may influence any kind of physical property and have evolutionary nature. That maintenance regime is existent in many fields ranging from machinery to fluid applications. According to International Foundation for Research in Maintenance it applies to more than 80% of maintenance [1].

C. *Leak detection methodologies*

The variety of mediums distributed in pipelines makes the discovery of improprieties like leakage, existent in a wide range of fields. Most of the focus in research is usually put in areas where any leakage may pose a threat to life to environment. These include big scale piping systems transporting petroleum or gas. In a smaller scale there are industrial systems which may use equipment powered by steam or compressed air. However methodologies dealing with CAS leaks usually limit to simple and not always fully reliable solutions. Although following methods deal with varied distributed mediums, the detection principles may be applied to all of them in most of cases. The division of leak detection methods can be found in [17]. Amongst the used methods are: negative pressure, acoustic, mass balance, flow or pressure change observation. Advantages of negative pressure are: useful for long distance distribution systems; offers possibility of localizing the leak as described in [8]. However that method requires computationally demanding processing. Acoustic method is usually conducted with the use of handheld, ultrasonic acoustic detectors [2]. Although widely used, it has significant drawbacks like: device needs to be in a close range from the piping system; checks are performed periodically which does give the opportunity for leak development between periods. Widely used leak detection method is based on mass balance equation. Despite simple principle, it requires constant measurements of other parameters which affect the gas mass. Interesting comparison of negative pressure and mass balance are shown in [8]. Flow or pressure change observation methods rely on principle that high change of flow or pressure signifies a leakage. It offers easiness of implementation. Based on one point of measurement the assessment of flow abnormalities can be performed. Research work presented in [10] utilizes that concept. Dynamic model methods are based on fundamental understanding of the underlying physics of the process. Extensive explanation can be found in [16]. Although some of the presented concepts have been used for monitoring leaks in CAS, most of them are utilized in applications which deal in transportation of fluids over long distances. That shows the need of a system monitoring compressed-air applications.

III. METHODOLOGY

A. *Sensor selection*

Discovery of leak can be performed having the knowledge of actual flow of air in the CAS. Therefore appropriate flow sensors had to be selected. It has been quite challenging task due to varied operational principles. Each of them has their pros and cons depending on the application. The first division results from measured value type. There are mainly two types: mass and volumetric flow sensors. Besides indicating different units, their operating principles are different. Under the volumetric group, there can be found sensors based on differential pressure, positive displacement and velocity principle. The mass flow meter group differentiates two operating principles: Coriolis and thermal. The advantage of volume flow meters is its low cost, big versatility on the market, applicability varied fluids. Although the volume flow devices are a common choice in industry, they possess unwanted property - temperature and pressure dependency. Fluctuation of these parameters results in changing gas density and viscosity. When that occurs, the previously set calibration point will no longer be valid and the measurement will not be accurate. As the designed system is supposed to be applicable in varied environments, selected sensor is based on mass flow measurement principle. Mass flow meters in contrast to volume flow meters are relatively immune to changes in inlet temperature and pressure – which is the biggest advantage. Mass flow meters based on thermal principle are more available on the market and at cheaper cost. Sensor has been chosen from Festo offer. Important factors that influenced the selection were: wide selection of ranges, price and freedom of installation. Quite often flow sensors have a restriction of place of installation. They had to be placed in laminar-flow regions, making installation demanding. Festo sensors due to their construction let the flow go through the bypass channel that generates laminar flow, where the sensing element is located. Considering required applicability of designed system in different scenarios, sensor should cover widest range possible. Festo sensors discover flow from 0.1 up to 5000 l/min. Best solution would be to choose one sensor that covers big range. However each of offered sensor covers some portion of 0-5000 l/min range. In order to cover widest range possible, two sensors have been selected: model SFAB covering range from 0.5–50 l/min and SFAM with range of 50-5000 l/min. Due to the big difference in sensors channels diameter, they were connected in a parallel connection. Series connection would result in significant obstruction of flow. However that decision resulted in additional signal processing. In parallel connection total flow is sum of two sensors outputs. Due to varied diameters there are differences in proportion of flow that is coming through sensors channels. Thus when either one of the sensors falls out of the range, other one needs a measurement correlation factor to compensate particular constant proportion of flow that is coming through the other channel. Otherwise, when both sensors cover the range, total flow is a sum of SFAB and SFAM sensors. Correlation factor has been found with help of Matlab tool – curve fitting. Quadratic equations appeared to be accurate enough.

Main objectives while choosing humidity sensor were: measurement range, accuracy, long-term stability, operating pressure and cost. In terms of precision, sensors based on chilled mirror principle are considered to be most accurate. However the cost is several times higher than sensors based on other principles. Therefore sensor based on capacitive measurement principle has been chosen. Although capacitive sensors cannot be compared in their performance to chilled mirror ones, they offer measurements accurate enough for industrial needs. Selected sensor, SF52 model is offered by Michell Instruments. It measures relative humidity. In order to obtain dew point temperature the values of relative humidity are correlated to reference dew points confirmed with a NIST-traceable (National Institute of Standards and Technology) chilled mirror hygrometer. The biggest advantage of this method is that the cost of such a sensor is several times lower than the one based on the chilled mirror methodology.

The selection of pressure sensor has been mainly motivated by cost and accuracy. There is very rich offer of pressure sensors on the market. Decision has been taken to find the pressure sensor from the same vendor that flow sensors were obtained, i.e. Festo. Selected sensor is a SDET type and offers maximum range measurement up to 10 bars.

B. Data acquisition – event driven model

Traditional approach in industrial monitoring applications is based on the pull paradigm. This communication scheme is based on the periodical polling of relevant data from hardware I/O devices. Drawback of that approach is high bandwidth usage as unnecessary requests are made in a short time interval in order to keep high accuracy. On the other hand, increasing the interval to relieve the communication link load, results in increase of update latency. Polling approach is thus inefficient as it may happen that observed data is at constant level for long period but data is polled anyway. Solution for decreasing the bandwidth usage could be using a contrasting approach - push paradigm. In that approach information is sent in form of event from the controller to other endpoints (for instance HMI). Event can be defined as occurrence of a happening of interest, which could be change of input value. Event-driven paradigm is much more efficient in terms of bandwidth usage as only relevant data is transmitted. Additional benefit of using event model is that neither the event nor the subscriptions to them are dedicated for specific endpoints. Such a design increases the interoperability, making the separation of communication from computational part more evident. Due to that, integration of low level devices into higher level system is easier and no additional drivers are needed to communicate. Eventing functionality is part of Service Oriented Architecture (SOA) and particularly its implementation – Web Services (WS). Used controller enables usage of WS through DPWS (Device Profile for Web Services) specification.

C. Discovering compressed-air anomalies

Following the main concept of condition monitoring, system is supposed to perform constant observation and announce any abnormalities in CAS, before their magnitude grows to level generating major energy losses or endangering equipment or process. After reviewing common practices in leak detection most of them include usage of distributed sensors in order to perform mass balance calculation. Additionally quite many of them incorporate computationally demanding algorithms. Due to requirement of mobility, system designed for this thesis is able to give one point of measurement thus distributed sensor approach cannot be considered. All data processing is planned to be done at the controller level thus any computationally demanding algorithms are restricted. Therefore the leakage discovery methods implemented in that project will base on flow observation, relying on principle that change of flow value from previously obtained reference band will signify a leakage.

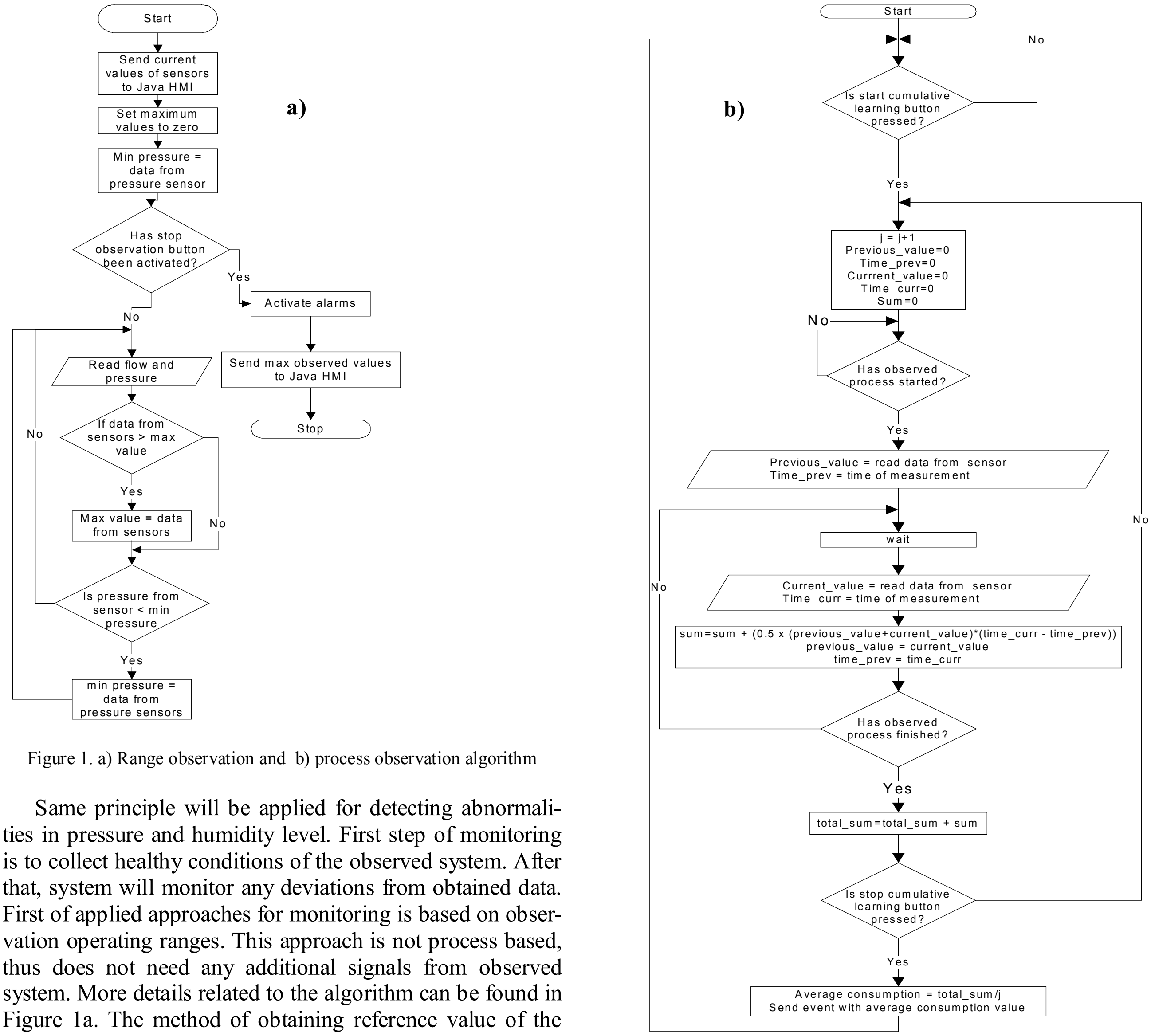

a)
Start
Send current values of sensors to Java HMI
Set maximum values to zero
Min pressure = data from pressure sensor
Has stop observation button been activated?
Yes
Activate alarms
Send max observed values to Java HMI
Stop
No
Read flow and pressure
If data from sensors > max value
Yes
Max value = data from sensors
No
No
Is pressure from sensor < min pressure
Yes
min pressure = data from pressure sensors
b)
Start
No
Is start cumulative learning button pressed?
Yes
j = j+1
Previous_value=0
Time_prev=0
Currrent_value=0
Time_curr=0
Sum=0
No
Has observed process started?
Yes
Previous_value = read data from sensor
Time_prev = time of measurement
wait
Current_value = read data from sensor
Time_curr = time of measurement
sum=sum + (0.5 x (previous_value+current_value)*(time_curr - time_prev))
previous_value = current_value
time_prev = time_curr
No
Has observed process finished?
Yes
total_sum=total_sum + sum
Is stop cumulative learning button pressed?
No
Yes
Average consumption = total_sum/j
Send event with average consumption value

Figure 1. a) Range observation and b) process observation algorithm

Same principle will be applied for detecting abnormalities in pressure and humidity level. First step of monitoring is to collect healthy conditions of the observed system. After that, system will monitor any deviations from obtained data. First of applied approaches for monitoring is based on observation operating ranges. This approach is not process based, thus does not need any additional signals from observed system. More details related to the algorithm can be found in Figure 1a. The method of obtaining reference value of the dew point temperature has been decided to be put manually by the user. Motivation of that decision was fact that in order to discover dangerous level of humidity in CAS, there is a need of measurement of ambient temperature. System at current stage of development does not have a temperature sensor included, thus value of surrounding temperature is put by the user. If dew point temperature of compressed-air will approach set reference point, alarm will be set. Based on obtained normal operating conditions, alarms are set. System has two types of adjustable alarms: low and high. Algorithm defined for alarms is a loop constantly comparing current data from sensors with calculated thresholds. Each alarm indicates one of the anomalies. If flow alarm is triggered it indicates higher than usual consumption of air - which announces leakage in observed system. Any deviation from pressure normal working conditions notifies either about pressure drop - which could result in air-operated device malfunction - or notify about over pressurizing equipment - which could result in devices destruction. In case when dew point temperature reaches set point, alarm will inform about serious possibility of condensation of water in the air distribution system. Second monitoring approach is process oriented and mainly focuses on observation of flow. In industry, many operations are repetitive due to automation of the processes. It could be pick and place performed by robot or work performed by cylinder. Taking under account repetitiveness of certain operations, particular amount of air is consumed during such an operation. This value can be a ground for monitoring possible leakages in the CAS. Any deviation would indicate higher consumption of air during the process which is result of leakage. After observation is started from Java HMI level, system waits for trigger indicating beginning of the observed process. After that, two consecutive measurements of flow are done with a predefined delay between them. By knowing the values of flow at those particular instances, integration can be performed in order to obtain consumption value in liters. Due to controller limitation, trapezoidal rule has been used to approximate definite integral. It is particularly accurate when integrating periodic functions over their periods, which is the case. Algorithm on which process oriented observation is based can be found in Figure 1b. Monitoring system is informed when the observed process starts and ends by means of messages send from controller responsible for observed process. More details on that operation will be presented in implementation section. After the average consumption value for process is calculated, observation is performed. Leakage is discovered by comparing the previously calculated average consumption with current process consumption. User similarly to normal operating range method specifies the thresholds of two alarms. Any deviation from usual consumption reaching the alarm zone will be notified to the user and considered as leakage in CAS. Additional features, helpful in monitoring process are total consumption information and leakage diameter estimation value. Former one accumulates each, calculated process consumption together. Leakage diameter approximation is based on transformation of equation for volumetric flow escaping through the leaks described in [7].

IV. Implementation

A. Hardware implementation

All the components i.e. four sensors, controller, in order to constitute one unified, independent system, have been placed in a cabinet. From the user perspective such a system should be in a sense a "black box" with an interface to connect to it.

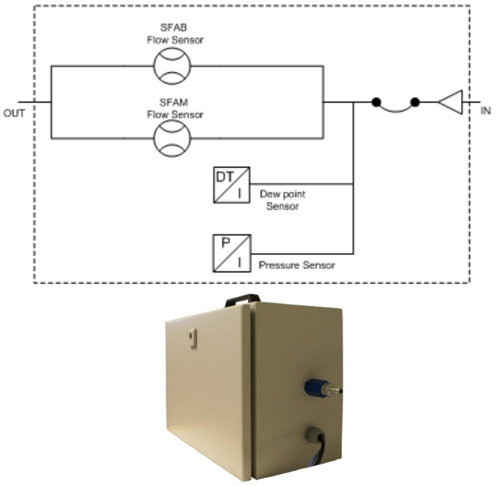


Figure 2. Hardware part of the system

As can be seen in Figure 2 that aim has been achieved. System can become part of compressed distribution line in an easy manner thanks to two push-in 10mm pneumatic connections .Thanks to its Ethernet based communication it can also very easily become either part of underlying communication layer or be connected directly to PC to gather process information in a straight forward manner.

B. Software implementation

All processing of the data gathered from the sensors is done at the controller level. Due to that and the implementation of Web Service (WS) technology, bandwidth usage can be reduced. S1000 (by Inicotech) controller which has been selected for that project offers Web browser interface, as well as WS capabilities. Web based access enables implementation of the logic as well as all necessary configurations independent of any additional software. There were two stages related to programming and configuration of the controller. First one included logic programming in Structured Text (IEC 61131-3) and is responsible for processing sensors data. Second one focused on configuration of the WS in order to properly communicate with external applications – in this case Java HMI. Configuration required definition of eight events and six input messages. Events are generated asynchronously, depending on the predefined rules and send to interested parties. Inputs are dedicated for HMI to communicate with controller. Each message is actually an XML message including specific data and appended as SOAP Body content when transferred to interested subscribers. Some events have been grouped together in order to decrease the bandwidth usage. For instance, alarm values and sensor measurement is encapsulated in one event instead of two separate ones. Next step included creation of a WSDL (Web Services Description Language) file. WSDL file is a language providing description on how to interact with particular service. Next and final step included implementation of HMIs. Decision has been taken to develop two independent HMIs. First one would be controller based, accessed through Web browser. Motivation for that approach has been: firstly exploiting HMI creation capabilities of the controller, secondly allowing easy access to monitoring system through Web browser. Second HMI has been developed as a Java standalone application. Its aim was to investigate SOA approach. Moreover it enabled development of additional functions which would not be possible at the controller HMI. As depicted in Figure 3, controller-based HMI consists of several graphical elements. There are three gauge components enabling live preview of current readout from sensors. Next to them there are alarm lights indicating two types of alarm: green - normal working condition and red -alarm. After specified period of observation has elapsed, normal working conditions in terms of operating ranges as well as proper process oriented consumption of air, are calculated. User can see those limits in form of two tables. They represents normal condition range observation results, values obtained in process oriented observation. Average value depicts average consumption after observation stage. Current value informs about consumption in latest process. Total value represents total consumption of air start-up monitoring system. In case any leak is discovered its approximated size will be displayed in the leak parameter.

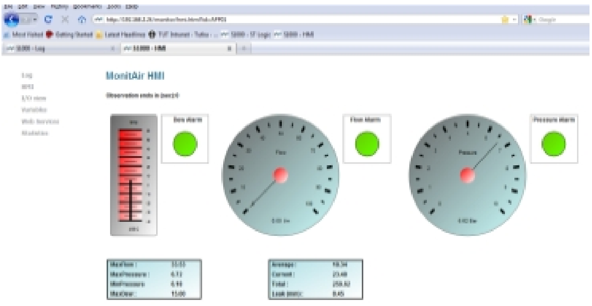


Figure 3. Controller based HMI

The main core of Java HMI is based on open source Java API for developing HMIs. However it has been highly modified in order to implement SOA and specific functions. Controller has already defined service thus in order to use it, HMI needs to be able to comply with Devices Profile for Web Services (DPWS). Thus the HMI has been built on top of Java Multi Edition DPWS Stack (ver.0.9.7) which implements the DPWS specification [5]. JMEDS enables creation of the client which can discover and use services available in the network. Application discovers specific services available on the network on the start-up. Application offers several functions which are accessed through the tabs. First of all, two types of observation and learning normal working conditions: range and process oriented approach.

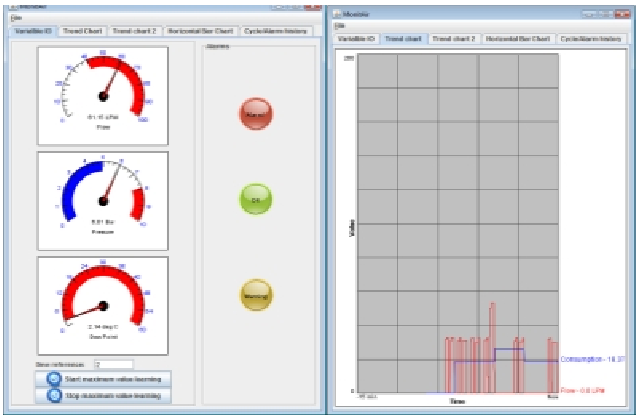


Figure 4. Java HMI

Two alarm states: low and high. Two real-time visualization of the parameters: gauge and bar chart types. Trend charts offering historic preview of parameters like: flow, pressure dew point and consumption. Parameters like average, last cycle and total consumption as well as leak diameter estimation. Additionally HMI offers alarm log as well as data recording to spreadsheets. All of information is exchanged by means of events. Both presented HMIs are independent.

V. RESULTS

Monitoring system has been tested at automatic manufacturing line. Main operation performed on one of cells, was loading and unloading products on the pallets by a robot, using suction grippers. This gave the opportunity to monitor transient conditions of the compressed air.

A. Event-based communication

System uses eight different event types. Some events are more frequent than the others. These are particularly events including real-time information. Events are generated by sampling two consecutive values with predefined delay. If the deviation at these time instances is bigger than predefined threshold, event is published. During the tests it appeared that changes of delay and threshold affect the amount of generated events quite significantly. First approach assumed short sampling interval and small deviation. However what occurred during tests was high amount of unnecessarily published events increasing the load of the network. Secondly this approach did not work with slow, long term changes of parameter where deviation was not rapid. In order not to overload network with high event rate and at the same time make system sensitive for long term changes (where the function of change over time is not steep) additional program has been added which samples the sensors inputs with much longer time interval. Manipulation of the deviation threshold between two time samples in both programs highly affects the number of events as well as sensitivity. In case of short sampling program, threshold has been set to higher values – in this case rapid and significant changes will be discovered. In case of long term sampling, small deviation threshold has been set to ensure that less significant changes are reported to the HMI. Proposed approach reduces amount of events, increasing the bandwidth and at the same time notifies user of system status in a reliable manner. Table 2 presents results related to bandwidth.

Table 2. Bandwidth usage results

	Download	Upload
Max rate	108.8 KB/s	25.2KB/s
Average	3KB/s	700 B/s
One event	20.4 KB/s	4.6 KB/s

B. Compressed-air anomalies discovery

First test phase focused on how well method for obtaining normal working conditions by assessment of operational ranges, will succeed in case of simulated anomalies.

Primary test concentrated on discovery of the leakage in CAS. Leak has been simulated by installing a valve with 4mm tubing at several different locations. Leakage discovery resolution is dependent on value of set thresholds for alarms – and especially the low alarm. Tests have shown that leakage of less than 1 mm in diameter will be detected with threshold of 5 % deviation from obtained reference range. Setting the value of warning alarm less than 5% may result in creation of sporadic false alarms. Repeatability of detection was very high. As leaks in industrial environments are never completely unavoidable it may be reasonable to set higher alarm thresholds – when repair due to magnitude of leak will be profitable. In case of pressure drop discovery this method's approach also proved to be successful. In case of tested environment, robot's pick and place operation created drop of around 0.6 bars. Thus the normal range has been at level from 6.2-7.4 bars. After obtaining the range, pressure has been decreased at the cell compressed-air inlet. Simulated decrease has been around 0.1 bars from usual conditions. Setting the low alarm threshold to 2%, reported the loss of pressure accurately. Due to the nature of compressed air, pressure drops may vary with time thus at least 5% threshold should be set for low alarm in order to avoid false alarms. Both sudden and long term pressure drops were discovered. For instance, high deviations have been discovered over a week observation in pressure level, which resulted in improper functioning of one of robot grippers. Similar successful results have been obtained with notification of high humidity level of the compressed-air. Due to the difficulty of adding extra humidity in to the compressed air or moving the testbed into much colder environment other approach have been conducted. Firstly the dew point reference value has been set when the compressed-air pressure was low. After that line pressure has been steadily increased – causing the dew point to steadily rise, according to Dalton's law. Alarms have been correctly triggered whenever humidity has reached alarm threshold.

Process oriented method observed pick and place robot operation using suction gripper. To notify monitoring system where the process starts and stops, controller responsible for observed process sends SOAP messages. After many process cycles, average consumption standard deviation has been at level of 0.178, which proves that measurement accuracy is high. System detected leaks of diameter around 0.5mm. This approach gives more accurate results than the previously described method, mainly due to the fact that is based on accumulated consumption during period of time and not on direct reading of flow measurement. Previous approach might periodically produce false alarms for instance when the thresholds are inappropriately set.

VI. CONCLUSIONS

Created system proved to be reliable tool in monitoring of compressed air anomalies. System announces deterioration before it affects functionality. Due to that, CAS is kept at high energy efficiency level. As the system monitors in an autonomous manner, supervision can be performed from remote location. System exhibits high accuracy, thus even small improprieties can be detected. Monitoring methods give possibility to monitor with and without knowledge of any process. Due to its design system is easy to connect to existing compressed-air and communication network – making it independent from underlying infrastructures. SOA model reduces amount of transferred data.

REFERENCES

[1] R. Beebe, Condition monitoring, Elsevier , 2004.

[2] B. Capehart, Encyclopedia of energy engineering and technology, CRC Press, 2007.

[3] Energy Tips – Compressed Air, Compressed Air Tip Sheet #3, U.S. Department of Energy, 2004.

[4] Dewpoint Measurement in Compressed Air Systems, Vaisala, 2001.

[5] http://www.ws4d.org (accessed: November 2010)

[6] Improving Compressed Air System Performance, a Sourcebook for Industry, U.S.Department of Energy, 1998.

[7] D. Kaya and P. Phelan. "Energy conservation in compressed-air systems." International journal of energy research, 2002, pp. 837-849.

[8] J.C. Martins and P. Seleghim, "Assessment of the Performance of Acoustic and Mass Balance Methods for Leak Detection in Pipelines for Transporting Liquids." Journal of Fluids Engineering , vol. 132, iss. 1, 2010

[9] B. McDuffee, Increase Compressed Air Plant Efficiency, Vaisala, 2006.

[10] Z. Nakutis, "An approach to pneumatic cylinder on-line conditions monitoring." MECHANIKA. Nr.4(72), 2008, pp. 41-47.

[11] Pennington, Jason, and Madhukar Puniani. "It pays to give energy management a thought." 2009.

[12] P. Radgen and E. Blaustein, Compresses Air Systems in the European Union, LOG_X Verlag GmbH, 2001.

[13] R. Saidur, "A review on compressed-air energy use and savings." Renewable and Sustainable Energy Reviews, 2010, pp. 1135-1153 .

[14] H. Shanghai, "Improving Energy Efficiency of Compressed Air System Based on System Audit". Berkeley Laboratory, 2008.

[15] I. Sherrington and T. Spring "Condition monitoring as a tool to aid compliance with ISO 14000." Tribology and Interface Engineering Series 2005, pp. 295-304.

[16] V. Venkatasubramanian, "A review of process fault detection and diagnosis." Computers and Chemical Engineering, 2002, pp. 293-346

[17] J. Zhang, Designing a Cost Effective and Reliable Pipeline Leak Detection System. REL Instrumentation Limited, 1996.

Contextualization of Learning Objects for Self-Paced Learning Environments

Freimut Bodendorf
Department of Information Systems
University of Erlangen-Nuremberg
Nuremberg, Germany
bodendorf@wiso.uni-erlangen.de

Kai-Uwe Götzelt
Department of Information Systems
University of Erlangen-Nuremberg
Nuremberg, Germany
sekretariat.wi2@wiso.uni-erlangen.de

Abstract— **The use of learning objects in constructivist learning environments causes a dilemma between reusability and context representation. The extension of current metadata standards with XML-based context resources offers a broad, transparent and efficient representation of the context of learning objects. They are the basis for context aware knowledge acquisition in self-paced learning environments.**

E-learning; hypermedia; learning context; context model; XML; LOM

I. Introduction

In contrast to traditional face-to-face learning, e-learning applications are available independent of time and space. They have in particular the potential to support self-paced and problem-oriented learning from a constructivist point of view. Learning processes are performed individually through active construction based on existing mental representations. Those individual learning processes can hardly be supported in face-to-face arrangements with a multitude of participants.

In addition to self-paced learning, constructivist learning environments are mainly characterized through representation of multiple contexts and perspectives [14] Transparent context representation of learning content is crucial for flexible knowledge acquisition and to apply the knowledge later on. Hypermedia learning environments meet these requirements. Furthermore, they also support various information processing abilities of learners through multifaceted encoding of learning contents [14, 15]. Due to variation and choice, they also motivate the learner during his learning process [8].

However, while using hyper-structured learning contents the modularization of existing learning resources is necessary. Thereby learning material is divided into small self-contained units. The main reason for using learning objects during the modularization process is the reusability of learning content for several learning scenarios and the effective development of digital learning material accordingly. Learning objects are described as "any digital resource that can be reused to support learning" [16]. These building blocks for learning mainly represent de-contextualized knowledge. The less specific context a learning object contains, the sooner it can be used for different learning scenarios and the higher is its "reusability" value. This modularization is often associated with disorientation and cognitive overload problems on the learners' side [3]. Only a transparent context representation of learning content allows active construction of knowledge and guarantees learning success [6]. This dilemma is also called Reusable Object and Instruction Paradox [1].

This paper points out how contextualization of learning objects independent of their granularity can be realized for hypermedia environments that support self-paced learning. First, the context that is relevant for self-paced learning environments is characterized. In addition, the deficiencies of existing metadata standards for a comprehensive, transparent and efficient description of the context of learning objects are illustrated. Subsequently, the approach and the solution of XML-based context specifications, which are used to describe the context of learning objects, are presented.

II. CONTEXTUALIZATION OF LEARNING OBJECTS

A. *Access to Learning Objects in Hypermedia Learning Environments*

In order to define the relevant context of learning objects it is crucial to know how learners access learning material in self-paced learning environments. The access to learning contents in hypermedia learning environments mainly happens topic-oriented. Based on desired knowledge the learner selects relevant nodes from a network. It is necessary to illustrate the relationships between learning contents of the current node and other contents as well as previous knowledge. On the one hand this is essential for learners in order to integrate the semantically impact of the information into the own knowledge structure and existing mental representations [12]. On the other hand, it must be transparent for the learner, which learning paths he can choose according to his desired and previous knowledge. For the representation of the context generic terms, specific terms and synonyms of the current topic as well as additional qualified relationships are being used (see fig. 1). Appropriate elements for navigation allow the retrieval of several contents and mapping techniques support the transparent presentation of the context [13].

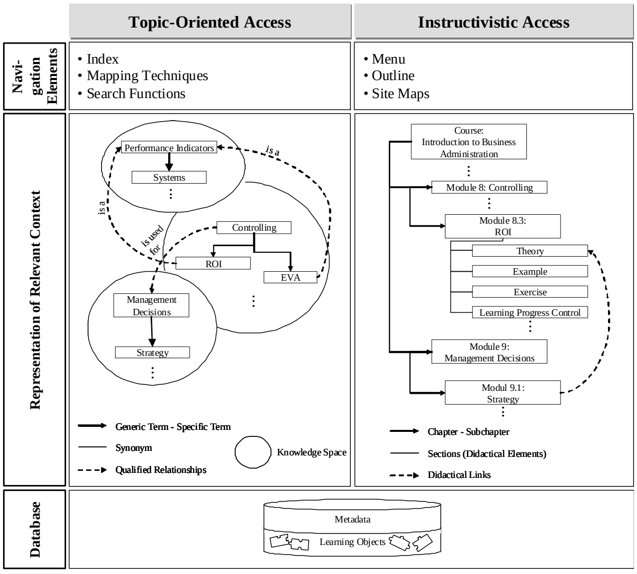


Figure 1. Access to Learning Objects

It is also necessary to organize learning material hierarchically, i. e. to provide content refining paths. The learner can determine his current learning status and is guided through individually selected contents by an instructivistic approach. Here the learning context is classified into several chapters which can be divided into subchapters and sections [2]. A standard learning path which reflects the curriculum structure appears for the learner. Besides this hierarchy of chapters and sections there are often didactical links [2]. Those cross-references occur i. e. for a comparison. The learner navigates through the contextualized content by menus, outlines, and site maps.

B. Metadata

The context of learning objects is described by metadata, which on the one hand characterize the learning objects themselves and on the other hand the relationship between each other [10]. Describing learning objects with metadata is a part of learning technology standards, which were developed due to compatibility and reusability of learning resources. Standards, which provide elements for describing relationships between learning objects, are of special interest to contextualization. This is widely realized in the IEEE LTSC Learning Object Metadata (LOM)-Standard [9]. The category “LOM.Relation” plays an important role in specifying the context. This category allows the definition of directed relationships, starting from the metadataset in which they are formulated to various target metadatasets [7].

However, while using the metadata scheme for a detailed description of the context three fundamental problems appear. In order to realize a bi-directional connection between learning objects always two opposing uni-directional relations have to be defined. Those pair wise connections demand a high effort when first specifying the context, as all relevant learning objects have to be found first which are to be referred to. In case of extension and updating as well as removal and exchange of learning contents metadata of both learning objects have to be changed.

As the authors of the learning objects mainly carry out the creation of metadata sets, domain experts have to define a framework for the relevant context. This aspect is not part of current metadata schemes.

An additional problem results from the limited language space of the LOM scheme relationships, which do not admit a comprehensive description of the context [1].

The solution proposed here is to separate context information from learning objects and metadata. The metadata of the learning objects remain untouched so that context information can be extended and completed as well as maintained by an individual domain expert. This assures consistence and timeliness and reduces the complexity of defining metadata. The relevant context information is represented as a classification scheme.

C. Subject-Based Classification

Subject-based classification classifies objects by relevant topics 85]. As a basis the ISO standard for topic maps allows the configuration of semantic networks which are separated from the referenced objects [11]. Therefore, the three constructs topics, associations and occurrences are available in topic maps.

All subjects but also abstract concepts and categories can be defined as topics [4]. Associations link topics to each other and also relationship types can be mapped. Relationships do not exist generically but are described in detail. A network develops from a hypergraph in which the topics are linked by associations, which may have several topics at their ends [4]. Finally, learning material which cover the topics or which are relevant for them is linked via occurrences to the topics.

Thus, topic maps have the advantage to build up a flexible model for mapping the context to an open vocabulary [5].

D. Structure-Based Classification

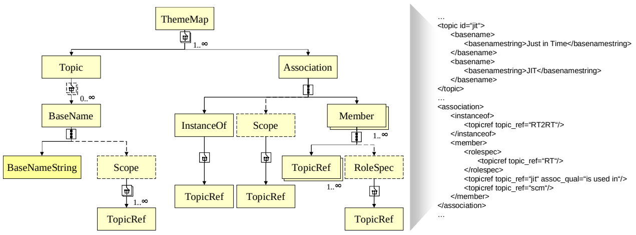
Structure-based classification groups modular objects around a given didactical structure of a course or learning material. In order to represent the curriculum structure in a didactical way hierarchically organized classification systems have to be established. Taxonomies that already exist within the LOM standard provide an appropriate basis. However, they have to be extended by the possibility of building sequences of learning objects. Interlocking chapters and subchapters represent the basic structure. For further description, they are enhanced with a numbering system, which specifies the position in the whole course. In case of a didactical and organizational motivated conversion of the course structure, learning objects do not have to be described again, as changes are made separately in the classification system.

Furthermore, chapters and subchapters have a rhetorical-didactical internal structure, which is characterized by the learning elements such as theory, exercises, and learning progress controls (see fig. 1). For the representation of this internal structure, labeling of the learning objects' types is sufficient as the sequence of the learning elements is determined independently by the learner within a self-paced learning environment.

III. XML-BASED CONTEXT SPECIFICATIONS

A. XML-ThemeMap

The ThemeMap is alligned to the ISO standards for topic maps. The main difference is that the construct occurrences is not used because authors only specify a uni-directional link from the learning object metadata to the topics due to the problems mentioned in section 2.2. Figure 2 shows the elements of the ThemeMap.



```
...
<topic id="jit">
      <basename>
            <basenamestring>Just in Time</basenamestring>
      </basename>
      <basename>
            <basenamestring>JIT</basenamestring>
      </basename>
</topic>
...
<association>
      <instanceof>
            <topicref topic_ref="RT2RT"/>
      </instanceof>
      <member>
            <rolespec>
                  <topicref topic_ref="RT"/>
            </rolespec>
            <topicref topic_ref="jit" assoc_qual="is used in"/>
            <topicref topic_ref="scm"/>
      </member>
</association>
...
```

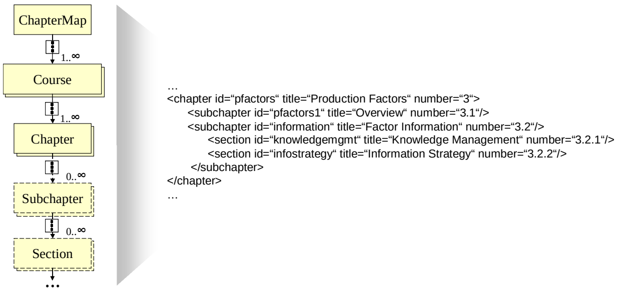
Figure 2. XML Structure of ThemeMap

Topics comprise subjects the learning material is addressing to. One or several *BaseNames* can be assigned to a *Topic*. Also synonyms can be given. Furthermore, the element *Scope* also allows the indication of a domain with a valid *BaseName*. Each *Topic* has an identification number (ID) which is referred to within the learning object metadata in order to classify the context.

In order to create relationships between topics with the element *Association* a reference type has to be chosen. All intended relationship types are defined as a *Topic*. The element *InstanceOf* refers to one of the defined relationship types. The element *Member* comprises *Topics* that will be associated by relationships. By using the attribute "*assoc_qual*" of *TopicRef* a qualified description of the relationships is possible.

B. XML-ChapterMap

Figure 3 shows the elements provided for the specification of didactical structures.



```
...
<chapter id="pfactors" title="Production Factors" number="3">
      <subchapter id="pfactors1" title="Overview" number="3.1"/>
      <subchapter id="information" title="Factor Information" number="3.2"/>
            <section id="knowledgemgmt" title="Knowledge Management" number="3.2.1"/>
            <section id="infostrategy" title="Information Strategy" number="3.2.2"/>
       </subchapter>
</chapter>
...
```

Figure 3. XML Structure of ChapterMap

All elements possess an attribute for identification within the XML data. Metadata do not have to be adapted in case of changes of the course structure because the learning object metadata is linked with the ID of the element. For rhetorical-didactical cross-references the metadata of the learning object refers to a further element-ID and allocates the attribute "*mentioned*".

IV. CONCLUSIONS

The presented contextalization of learning objects has been realized within the course "Introduction to Business Administration" of the "Bavarian Virtual University". The context is separated from the learning objects and included in context resources that permit easy mapping and updating by a domain expert. The realized self-paced and hypermedia learning environment converts the XML-based context resources and generates dynamically the user interfaces for a topic-oriented as well as an instructivistic access to the learning material.

The contextualization of learning objects using ThemeMap and ChapterMap proposes an appropriate extension for current learning technology standards. ThemeMap is a flexible instrument for context mapping which may be customized to the individual e-learning content and may be processed automatically. In a wide sense contextualization not only covers the representation of relationships between learning objects but also takes into account learning environments in real life scenarios, e. g. at work or at home, and supports social context, e. g. learning groups or communities. These are open fields for research in many ways. ThemeMap can also be used to describe such context elements.

REFERENCES

[1] P. Baumgartner, "Creating, sharing and reusing e-Learning Content. Position Paper for European Commission", DG for Education and Culture: Consultation Workshop, Brussels, October .2004.

[2] J. Caumanns, „Automatisierte Komposition von wissensvermittelnden Dokumenten für das World Wide Web", http://www.ub.tu-cottbus.de/hss/diss/fak1/caumanns_j/pdf/diss_caumanns.pdf, 2000.

[3] J. Conklin, "Hypertext - An Introduction and Survey," IEEE Computer, 1987, 20 (9), pp. 17-41.

[4] G. Flach, KnowledgeDirect–Einsatz semantikbasierter Wissens-management-Technologien im Unternehmensnetzwerk „BioCon Valley", 2002.

[5] L. M. Garshol, "Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it all." Journal of Information Science, vol. 4, 2004, pp. 378-391, doi: 30.

[6] C. Graesel, J. Bruhn, H. Mandl, and F. Fischer, „Lernen mit Computernetzen aus konstruktivistischer Perspektive," Unterrichtswissenschaft, vol. 25, pp. 4-18, 1997.

[7] S. Hoermann, A. Faatz, O. Merkel, A. Hugo, and R. Steinmetz, „Ein Kurseditor für modularisierte Lernressourcen auf der Basis von Learning Objects Metadata zur Erstellung von adaptierbaren Kursen". Proc. Tagungsband der GI-Workshopwoche: Lernen-Lehren-Wissen-Adaptivität, University of Dortmund, 2001, pp. 315-323.

[8] A. Holzinger, „Basiswissen Multimedia", Wuerzburg, vol. 3, 2001.

[9] IEEE LTSC, "Draft Standard for Learning Object Metadata," http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf, 2002.

[10] J. M. Pawlowski, and H. H. Adelsberger, " Standardisierung von Lern-Technologien", Wirtschaftsinformatik, 1 (43), pp. 57-68, 2001.

[11] S. Pepper, and G. Moore, "XML Topic Maps (XTM) 1.0," http://www. ThemeMaps.org/xtm/1.0/xtm-20010302-2.html, 2001.

[12] S.-O. Tergan, „Hypertext und Hypermedia: Konzeption, Lernmöglichkeiten, Lernprobleme" in Information und Lernen mit Multimedia, L. J. Issing, P. Klimsa, Eds., Weinheim: Beltz PVU, 1997, pp. 123-138.

[13] M. Thüring, J. Hannemann, and J. Haake, "Hypermedia and Cognition: Designing for Comprehension. Communications of the ACM," 38 (8), 1995, pp. 57-66.

[14] B. Weidenmann, „Multicodierung und Multimodalität im Lernprozes," in Information und Lernen mit Multimedia,L. J. Issing and P. Klimsa, Eds., Weinheim: Beltz PVU, 1997a, pp. 65-84.

[15] B. Weidenmann, „Abbilder in Multimedia-Anwendungen" in Information und Lernen mit Multimedia, L. J. Issing and P. Klimsa, Eds., Weinheim: Beltz PVU, 1997b, pp. S. 107-121.

[16] D. Wiley, "Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy", in The Instructional Use of Learning Objects, D. Wiley, Ed., 2000, available online: http://reusability.org/read/chapters/ wiley.doc.