International Journal on Advances in Systems and Measurements

2011 vol. 4 nr. 3&4

The *International Journal on Advances in Systems and Measurements* is published by IARIA.
ISSN: 1942-261x
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Systems and Measurements
Volume 4, Number 3 & 4, 2011

- Jean-Henry Morin, University of Geneva - CUI, Switzerland
- Igor Podebrad, Commerzbank, Germany
- Leon Reznik, Rochester Institute of Technology, USA
- Chi Zhang, Juniper Networks, USA

Sensor Networks

- Steven Corroy, University of Aachen, Germany
- Mario Freire, University of Beira Interior, Portugal / IEEE Computer Society - Portugal Chapter
- Jianlin Guo, Mitsubishi Electric Research Laboratories America, USA
- Zhen Liu, Nokia Research – Palo Alto, USA
- Winston KG Seah, Institute for Infocomm Research (Member of A*STAR), Singapore
- Radosveta Sokkulu, Ege University - Izmir, Turkey
- Athanasios Vasilakos, University of Western Macedonia, Greece

Electronics

- Kenneth Blair Kent, University of New Brunswick, Canada
- Josu Etxaniz Maranon, Euskal Herriko Unibertsitatea/Universidad del Pais Vasco, Spain
- Mark Brian Josephs, London South Bank University, UK
- Michael Hubner, Universitaet Karlsruhe (TH), Germany
- Nor K. Noordin, Universiti Putra Malaysia, Malaysia
- Arnaldo Oliveira, Universidade de Aveiro, Portugal
- Candid Reig, University of Valencia, Spain
- Sofiene Tahar, Concordia University, Canada
- Felix Toran, European Space Agency/Centre Spatial de Toulouse, France
- Yousaf Zafar, Gwangju Institute of Science and Technology (GIST), Republic of Korea
- David Zammit-Mangion, University of Malta-Msida, Malta

Testing and Validation

- Cecilia Metra, DEIS-ARCES-University of Bologna, Italy
- Rajarajan Senguttuvan, Texas Instruments, USA
- Sergio Soares, Federal University of Pernambuco, Brazil
- Alin Stefanescu, University of Pitesti, Romania
- Massimo Tivoli, Universita degli Studi dell'Aquila, Italy

Simulations

- Tejas R. Gandhi, Virtua Health-Marlton, USA
- Ken Hawick, Massey University - Albany, New Zealand
- Robert de Souza, The Logistics Institute - Asia Pacific, Singapore
- Michael J. North, Argonne National Laboratory, USA

International Journal on Advances in Systems and Measurements
Volume 4, Numbers 3 & 4, 2011

CONTENTS

Internet Traffic Distribution

Peter Danielis, University of Rostock, Institute of Applied Microelectronics and Computer Engineering, Germany
Jan Skodzik, University of Rostock, Institute of Applied Microelectronics and Computer Engineering, Germany
Jens Rohrbeck, University of Rostock, Institute of Applied Microelectronics and Computer Engineering, Germany
Vlado Altmann, University of Rostock, Institute of Applied Microelectronics and Computer Engineering, Germany
Dirk Timmermann, University of Rostock, Institute of Applied Microelectronics and Computer Engineering, Germany
Thomas Bahls, Ernst-Moritz-Arndt-University of Greifswald, Institute for Community Medicine, Germany
Daniel Duchow, Nokia Siemens Networks GmbH & Co. KG, Broadband Access Division, Germany

About the influence of virtualization and open source technologies on V&V lifecycle

Jean-Pierre Lebailly
EADS - EUROCOPTER
ETZSR
Marignane, France
jean-pierre.lebailly@EUROCOPTER.com

Nicolas Belanger
EADS - EUROCOPTER
ETZSR
Marignane, France
nicolas.belanger@EUROCOPTER.com

***Abstract* —the present paper discusses the strategic evolution of avionics test systems at EUROCOPTER as requested by virtualization turn. Until now considered as a necessary evil in the global avionics product lifecycle, test systems are taking the path to be promoted as key productivity enablers. Industry competitiveness pushes avionics actors to revisit and strengthen their Verification & Validation capabilities. To support it, EUROCOPTER has launched dimensioning projects that are introduced in this article. EASI (Eurocopter Avionic System Ide) which will be the new Test System at EUROCOPTER is presented as a case study. The benefits of open source technology compared to old fashion proprietary test systems are discussed. On top of that virtualization turn is illustrated through the SDMU (System Digital Mock Up) tool developed among others to simulate interfaces between system components in order to fix equipment specifications as early as possible to reduce the lead time of integration tests on rigs and flight tests. Perspectives are then given in terms of Test and Simulation system near convergence.**

Keywords - Model Based Testing; Open source; Avionic System Rig; Virtualization; EASI; Avionic Simulation.

I. INTRODUCTION

For the past 20 years, the major question when building test systems was to ensure the real time performance. Due to technology limitation, the solutions were proprietary operating system embedded in specific VME (VERSA Module Europa) [2] hardware. The test methodology was an important parameter but constrained by the technology. The real time expertise was the core competency for test systems departments in avionic companies. Nowadays, the evolution of hardware allows us to imagine new real time solutions based on standard and low cost components. The performance of the hardware becomes less important. At the same time, industry competitiveness becomes the major parameter. The systems are more and more complex but the cost and delay are drastically reduced. The consequence is that focusing on validation and test at hardware level is no longer enough because it occurs too late in the project life. The way to improve the process is to perform validation at each step of the V cycle. That means using: Model Driven Architecture for requirement capture and architecture definition; Simulation to validate system specification and detailed specification; Tests on rigs to qualify equipment and system integration. This approach has a major impact on test systems domain: the validation cycle is distributed along the whole V cycle; heterogeneous tools cooperate to the final result; new populations are concerned by test and validation. Whereas test systems were until now considered as a necessary evil, they are now entering into the phase of potential productivity enablers. The scope of the department Test Systems is becoming larger, services level agreement must remain at the highest level while expertise resources cannot increase significantly. Only critical components may remain EUROCOPTER internal, the others shall be open source components or delegated to third parties. Needed expertise becomes more integration oriented. This new model is illustrated in the present paper through a current status of test system at EUROCOPTER, a study of potential benefits of virtualization approach at system architecture level, the EASI (EUROCOPTER Avionic System Ide) case study. A current status about long term test systems practice at EUROCOPTER is performed, then virtualization turn in terms requirements and benefits is introduced. At least in chapter 4 we present the EASI case study with a particular focus on open source technology interest for test system change. Moreover, the case study introduces the SDMU tool for avionics interfaces simulation.

II. CURRENT STATUS

EUROCOPTER test systems activity began in middle 80's. Three tools, Mona Lisa, Anais and Artist have been developed successively. The necessity to build a new tool was justified by a major modification of the hardware environment or the addition of a major functionality. In anyway, the architecture principles stay unchanged and we can speak globally of EUROCOPTER Test System Tools in the present article.

The first intention was to take into account three main aspects:

1) Test and qualification of avionic systems: Software Test Bench and Integration Rigs;
2) Specification and debug of embedded software through simulation capabilities;
3) Development framework for training facilities.

Our ambition was to address not only the right part of the verification and validation cycle, equipment test and system integration test, but also to be able to address the left part at the specification level. We will try to understand why some

limitations have made it a relative success and why it is now the good period to reach our objectives taking advantage of recent technical gaps.

A. Test tools architecture

The architecture built in the middle of the 80's is still operational today. It is organized around:

- A real time platform
- A workstation dedicated to piloting;

For the last twenty years, full VME architecture was the best solution to manage avionics systems hard real time testing. VMEBus is particularly efficient to allow I/O event management: multi processing synchronization, access to different hardware resources (CPUs and I/O boards) in a transparent way.

EUROCOPTER has integrated VMEBus as a standard backbone for embedded helicopter systems integration test bench activities. The current architecture used at EUROCOPTER for System Integration Rig is introduced by Figure 1:

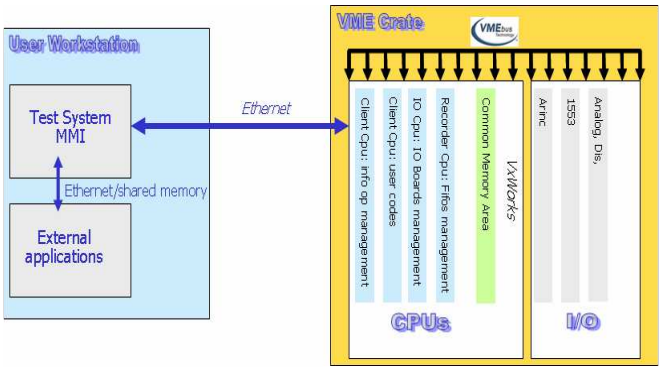


Figure 1. EUROCOPTER Integration Test Bench current architecture

Three categories of real time CPUs using VxWorks operating system are embedded in the crate. Client CPU are dedicated to computing tasks which consist in: first, coding and decoding data coming or sent to the avionic equipments; secondly, in simulation models. IO CPUs are managing the interface between the Test environment and the specific board connected on the avionic buses (ARINC 429, MIL-STD-1553, etc.). Recorder CPUs are recording data's in the dedicated protocols format with no discrepancy introduced by the tool. A common memory area allows synchronized and controlled real time exchange between CPUs.

B. Limitations

The use of EUROCOPTER tools for test, qualification and training was globally successful. Nevertheless, we have encountered some limitation against the competitiveness of the product and the use of specific functionalities, especially during the last years.

1) Structural constraints

As in other companies, we have to produce quicker, cheaper and keeping product quality at the same level. In our model, all the functions are made internally that needs maintaining an important development team and induces costs. The use of external software like: Oracle, pSOS and VxWorks generates license cost.

2) Real time debug

EUROCOPTER test system offers real time debugging capabilities for user codes. This capability is mandatory when you want to use the tool for embedded system development through simulation. In both real time environments, pSOS and VxWorks, there were some issues which prevented us to put the functionality in service. First, Multi real time environment for pSOS and Tornado environment for VxWorks induce high license cost. Secondly the real time debugging capabilities were not completely mature in case of multiple CPUs configuration. As an example, propagation of a breakpoint on all the CPUs was not possible with Multi-pSOS and generates unpredictable and unacceptable jitters with Tornado-VxWorks.

3) Customized BSP (Board Support Package)

We have been obliged to customize the CPU BSPs in order to guaranty some real time functionalities like:

- Use of the SYSFAIL to synchronize board on a physical signal;
- Message passing between CPUs;

Each time we decide to use a new type of CPU, it generates high recurrent costs to propagate these modifications.

4) Customized IO boards API (Application Programming Interface)

As some functionality's of the IO boards were not usable through the supplier standard API, we have developed our own customized API. In the same way as for the BSPs, each time we integrate a new board we have recurrent costs.

5) Test bench user profile

Since 20 years, the testers profile has changed. Test teams were composed of technical profiles aware of avionic buses protocols. Nowadays, the profile is more a generalist software engineer with good knowledge about avionic buses but some lack at the protocol level. In case of failure during a test they have difficulties to diagnose if the original cause of the problem is located: in the tool, in the procedure or due to the equipment under test. The philosophy of EUROCOPTER tools fits the needs of people familiar to avionic bus protocol but this philosophy is limited with less mature team. This may induce blocking point and increase delay.

6) Ergonomics

During the past 20 years, the industry of personal computing has increased. Everybody is now working on a personal computer for office automation and has also one or

more personal computers at home. Ergonomic evolution on these platforms is continuous and very fast. In people's minds, it should be normal to have the same evolution at work with the tools supposed to help them to improve their productivity. Even if the tools are operational and able to fit the needs, there is a rejection if the tool does not provide a familiar environment with all ergonomics facilities.

III. VIRTUALIZATION

A. Competitiveness constraint

As other companies, EUROCOPTER is strongly challenged by competitors in terms of Time to Market, quality, product costs. It requires finding new leverages to improve our competitiveness factors. Jumping to virtualization process will allow validation loop at each step of the V Cycle thus shortening time frame deliveries, improving product quality and reducing final costs. Virtualization is an emerging project at EUROCOPTER in which we are making progress in the preliminary analysis stage: no choice is already made in terms of tools, methods, architecture.

B. Brief state of the art

Even if still far from moving to "source model" concept rather than "source code", the emergence of Model Driven Architecture (MDA) [3] supported by the Object Management Group (OMG) has become a reality. In its 2006 Workshop on Model Driven Architecture [12], the Carnegie Mallon Software Engineering Institute stated that "MDA has been endorsed by various entities of the U.S. Department of Defense (DoD)". Moreover, in [13], a Defense Science Board report on the Joint Integrated Fire Support Systems (JIFSS), a large distributed system, it was stated that JIFSS "should employ the Model-Driven Architecture (MDA) development approach for designing the JIFSS architecture and in implementing its component systems. The MDA [5] approach ensures adherence to standards across the components and has been shown to substantially reduce costs in the development of large-scale systems-of-systems". In the same way, Lockheed Martin [6] in the frame of F-16 Modular Mission Computer Application Software development, proved the real interest of model driven development techniques for industrial purposes close to the ones of EUROCOPTER. On top of that, Forrester Consulting conducted a major study in 2008 on 132 companies about Model-Driven development methodologies [8], and concluded that putting the model as the central artifact of development life cycle has major benefits. Last but not least, TOPCASED project [7] recently provided some important results consecutively to a proof of concept performed in 4 different functional areas which argue in favor of Model Driven development.

C. EUROCOPTER users requirements

EUROCOPTER needs a total development approach meaning that architects, designers and engineers should be involved from start to finish of a project. The first requirements which are pushed by EUROCOPTER system architects are:

- Being able to automatically parse each requirement in order to diagnose its correctness according to potential errors formally described in [14]
- System requirements coverage analysis based on modeling approach: being capable to provide automatic validation about the requirements consistency and completeness.
- Being able to re-use use cases at corresponding steps of the V-cycle (see Figure 2)

It requires to be supported by tooling enabling Virtualization. This is the mandatory path to support avionics design office. We, therefore, plan to build a global framework (cf. Figure 2) in which interact the following components:

- Modeling, Test System Functionality, System Digital Mock-up (SDMU) [11],
- Standard Interfaces to real avionics & IT hardware,
- Modular inclusion of Tools: interface databases, Test Automation, any COTS (Commercial Off-The-Shelf),
- Available on every engineer's desktop.

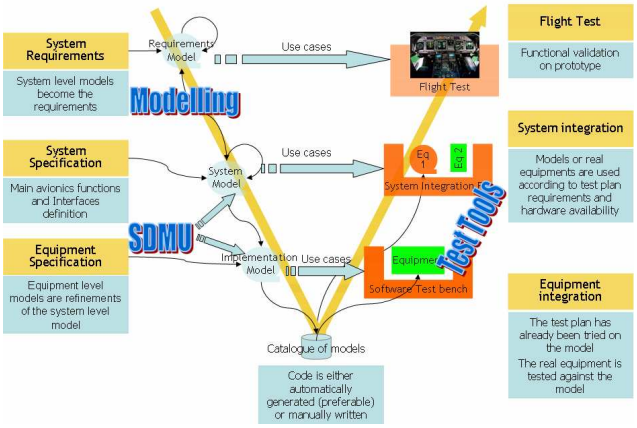


Figure 2. EUROCOPTER avionic design office framework

D. Virtualization benefits

The main benefits expected by EUROCOPTER in the jump to Virtualization perspective are:

- Reach "avionic product first time right delivery"
- Shorten Time to Market
- The capability to capitalize from one project to another through models library
- Extend the test completeness : capability to test failures in the whole flight domain in simulation mode;

IV. THE EASI CASE STUDY

In the near future, we must be able to federate heterogeneous people working in heterogeneous environment with heterogeneous tools. The success factor will be our availability to provide the good tools for each kind of job and a framework allowing EUROCOPTER teams to work together, avoiding redundancy and improving quality by optimizing the work at each level. How could we face all these new subjects without increasing unreasonably the size of our test tool team?

This can be achieved relying on 3 basic pillars (cf. Figure 3): Collaborative Approach, Open Source and SDMU:

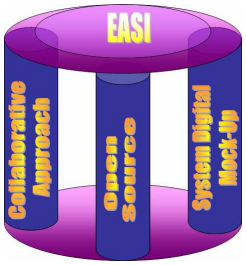


Figure 3. EASI framework pillars

A. *Collaborative approach*

After a detailed overview of the market [9], we decided to build a collaborative solution with EADS TEST&SERVICES. EASI project aims to build an EADS collaborative development platform focusing on avionic systems integration tests. The original and ambitious strategy of EADS TEST&SERVICES is to propose a modular plug-in solution at MMI (Man Machine Interface), communication and also at real time level. This relies on three mains components:

- U-TEST RTC (Universal Test Real Time Component), an open real time component distributes **third party modules** in a unique and integrated real time test system;
- A plug-in based MMI allows association of development from various parties and fits it to real time system modularity;
- An open framework manages a network based communication between all the parts of the test;

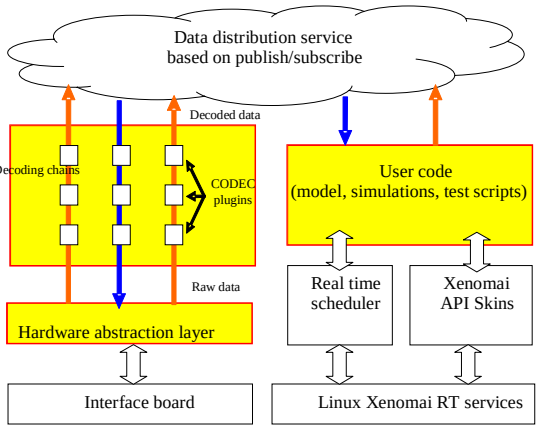


Figure 4. U-TEST RTC architecture

On figure 4, all the non blank parts of U-Test are concerned by the real time plug-in approach. This allows U-Test integrators to connect, under their own, specifics hardware to the system. Concretely it is one key feature that has driven EUROCOPTER's choice for Test System change. This is the guaranty 1) to be able to keep all our existing hardware without any modification of the wiring of the bench, 2) to keep the availability to adapt under our own the tools to future EUROCOPTER specific needs. For EADS TEST&SERVICES it is the opportunity to inherit EUROCOPTER specific developments and to integrate them to its offer. At EADS level it is a chance to share test systems development between different Business Units. The Figure 5 presents the organization proposed between EUROCOPTER, EADS TEST&SERVICES, EADS other Business Units and Open Source world in the frame of EASI Project.

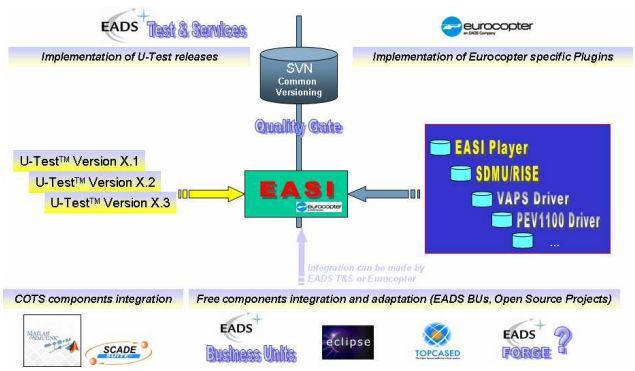


Figure 5. EASI project organization

The innovative collaborative model imagined for the EASI project proposes to share the developments between EUROCOPTER and EADS TEST&SERVICES under a common versioning (SVN). The provider/customer classical model is then slightly changed. In order to face the Virtualization step and its test systems renewal, EUROCOPTER needed to be partnered by a test system core

specialist which role was to become a pure "services backbone" as keeping the ability to contribute into the developments in terms of close helicopter test systems specificities. EUROCOPTER will then keep test system expertise knowledge to design specific developments. These developments will be performed through plugins and used as temporary versions. The specific plugins shall pass a development quality gate to be accepted by EADS TEST&SERVICES and to be integrated to major versions delivered twice a year. After having being integrated to major versions, the EUROCOPTER plugins will then be maintained (fixes) by EADS TEST&SERVICES.

B. *Open Source*

An important challenge for EADS TEST&SERVICES and EUROCOPTER is the capability to develop a new test system from scratch with strong constraints in term of budget and delay. We have made the choice of using Open source components; the potential advantages are well known:

- Ease and accelerate the development phases;
- License free
- Supported by large communities;

One challenging aspect of using Open Source is to have in minds that our Test systems must have a long term life. These are costly systems that must support avionic system development during all the product life which exceeds tens of years. It is important to choose well known and stable Open Source components:

- U-Test real time tasking is based on orocos (Open Robot Control Software) components;
- U-Test MMI is based on Eclipse. Consequently, it inherits of all the C C++, fortran and ADA Eclipse development and debugging platform capabilities. This is a great advantage for developing simulation models;
- U-Test data distribution service is based on the popular EPICS framework. That opens to U-Test all the data control and visualization tools which shares this framework.

Another challenging aspect is to rely on communities whose constraints and goals may be completely disconnected of ours. We have no assurance that these components will evaluate in a way which fits our needs. Perhaps that means that we must become actors in Open source projects.

C. *Advantage of openness by examples*

Since the end of our initial study, the project has completed two important steps

- A prototype phase;
- An operational mockup phase;

The goal of the prototype phase was to validate the main principles of this new collaborative approach, demonstrating the capability to work in a collaborative model between two EADS Business Units and that the plug-in approach proposed by EADS TEST&SERVICES allows EUROCOPTER to implement real-time connection to its specific hardware without degrading performances.

On another hand, the goal of the operational mockup phase may be seen as the first industrialization step. Its goal was not to demonstrate the technical ability of the project, but to show that based on the Open Source strategy it is possible to build a complex test integration platform at the state of the art in term of ergonomics and which is well accepted by our customers.

1) *U-TEST RTC Real-time plug-in*

The goal of this paragraph is to describe how it has been possible to reach EUROCOPTER wish to implement by our self the connection of EASI to our VME interfaces boards.

To avoid costly modifications, the main requirement of EUROCOPTER is to keep as much as possible the hardware interface of the rigs unchanged and particularly the wiring. It is why we impose that EASI must be able to drive the system through the IO boards in VME crates which are used with ARTIST, which is the actual test system. The problem is that U-TEST RTC component is based on a PC platform and consequently it is using the PCI bus for board connection.

The first and mandatory step was to find a way to connect the VME crate and the PCI Express. Of course this link must provided all the services we are using for IO boards management:

- DMA;
- Interrupt management;
- Mail boxes

We had the opportunity to be a privileged partner of IOxOS technology for the development of the PEV1100 [15] PCIe bridge. The very impressive results of the test with a prototype of the PEV1100 convinced us to select it as the basis of the connection of EASI to the VME world.

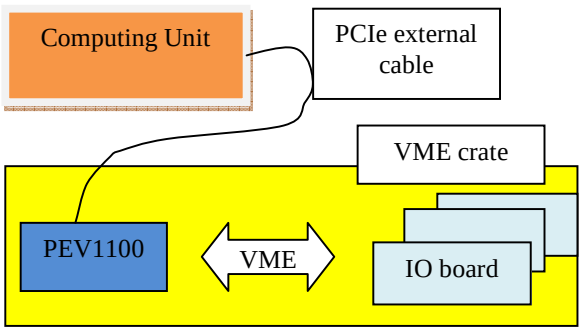


Figure 6. PEV1100 Overview

At this step, the hardware platform is completely defined and the remaining task is to implement the access of EASI test system to the IO boards embedded in VME crates through the PEV1100 Bridge using the plug-in approach allowed by U-TEST RTC.

The plug-in approach of the U-TEST RTC component is based on codec and protocol. Data is decoded / encoded using a chain made up of a list of user definable modules. These modules can be either used to extract the payload from a message (e.g., identifying the real data associated to an ARINC 429[21] label based on the label given – these are called protocols) or to convert data (called codecs) such as endianness conversion, scaling, BCD decoding, etc. Protocol and codec chains are executed sequentially in order to transform raw data to system level data. The result of this process is published in the Global Data Space making it available to other system parts (which can interconnect with via a LAN). In addition, it is possible to publish intermediate calculation between codecs to Global Data Space so that the user can analyze the result of each step inside the decoding / encoding chain. This modular approach eases capitalization between projects by sharing real time data processing code. For a given test and for each system variable, the decoding chains are defined in an XML Interface database. At the beginning of the test or on demand, the chains are instantiated according to the database definition. Each codec and protocol plug-in contributes to enlarge a C++ Interface Class library. An Xml description of the plug-in allows to make a dynamic link between U-TEST RTC and the library where the entry point of the plug-in implementing the methods of the Interface Class are provided.

An important point is that the way to access each IO resource is described in the Xml Interface database and may consist of multiple protocols and codecs chained sequentially. So, what is the application to the particular case of accessing IO boards in a VME crate?

First EUROCOPTER has implemented the elementary protocols and codec plug-in:

- A protocol to open a PCI expresses access to the VME area through the PEV1100 bridge;
- For each type of IO board embedded in the VME crate, a protocol to manage the board;

It has not been necessary to implement codecs because codecs to encode/decode engineering values inside ARINC 429 labels (BCD, BIN) or MIL-STD-1553[22] message were already provided by U-TEST RTC;

Secondly, we have described in the Xml Interface database that the board embedded in the VME crates must be accessed by chaining:

- The PEV100 protocol plug-in;
- The dedicated IO board protocol plug-in
- The codecs plug-in;

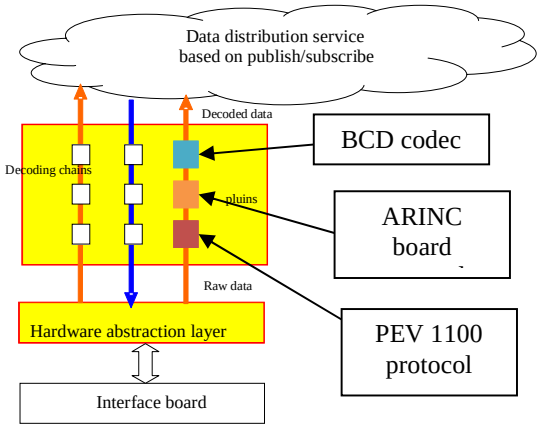


Figure 7. VME chain of codecs and protocols

At the end of the operational mockup phase, the real-time plug-in strategy, as shown on Figure 7, is fully demonstrated. Of course some problems are remaining, for example, we are looking for a way to ensure a complete synchronization between the PCI express world and the VME world especially in term of event dating. There are known solutions like IRIG [24] link, but it is a costly solution and not available with all our park of IO boards. In any way, these kinds of problems are not linked to the plug-in and collaborative approach.

2) Open source strategy the Eclipse and CSS (Control System Studio) example

Another main strategy of the EASI project is to remain on Open source. Here are two typical example of how Open source could help to build complex environment.

The Eclipse RCP framework

The modular approach proposed by U-TEST RTC on the RTOS (Real Time Operating System) side had to be matched with the same level of customizability on the MMI side. Thus, EADS TEST&SERVICES used Eclipse RCP as the base framework for its client side application.

Eclipse RCP is a platform for building and deploying rich client applications, supporting the famous Eclipse Java Integrated Development environment. It provides fundamental workbench functionalities such as movable and stackable window components (editors and views), menus, tool bars, push buttons, tables, trees, and much more with an OS native look and feel for applications and features.

Moreover, Eclipse RCP offers commercial off the shelf modules needed in our EASI project test system:

- Development environment for user code, eclipse CDT [16].
- Decoding chains Database modeling (EMF) [17]
- Test and projects configuration management (Subversive) [18] [19]

Each of these modules can be “overloaded” in order to customize it to match our needs. For instance, EADS TEST&SERVICES added auto-completion capabilities to

the user Code Development Module (CDT), based on avionic signal declaration in the Interface Configuration Database. That means that when the end user wants to access an avionic signal in a simulation code, he may use auto-completion based on the definition of the avionics signals done by the architect designers.

The Control System Studio

EASI is an avionic test system, it must be able to interface and control complex avionic configuration. It is then mandatory to provide an ergonomic interface to interact with the system. Based on Open source approach, the choice for EASI has been to do an investigation on existing products. Of course the criteria have been to find an existing tool fulfilling the requirement in term of ergonomic but also in term of integration with EASI. We have just seen in the previous paragraph that EASI is based on the Eclipse RCP framework, it has also been mentioned that the data distribution service used in EASI is the popular EPICS. Comparing all the solutions based on Eclipse and EPICS, we have chosen the CSS (Control System Studio) [20] tool because it is worldwide used in scientific community this ensures its perenniality at mid and long term, its components cover a large scope of EASI needs and will drastically improve graphical control capability compared to our actual tools.

The main effort to integrate the CSS to the EASI project has been to merge the EASI avionic configuration database and the CSS signal description. But, the task has been really eased by the fact that the two projects are base on the same EPICS frameworks. Finally, with a reasonable effort, we can take advantage of a very powerful tool usable as it is. For the future, we plan to developed some new graphical control more avionic oriented. Of course these new components will be shared with the CSS community as CSS plug-ins.

At the end of the operational mockup phase of EASI, we can say that the Open source strategy is a real success. The two previous examples show that if we build a coherent architecture, for example in our case, choosing Eclipse framework for GUI, EPICS framework for Data distribution service and then CSS based on the two previous frameworks, it reduces drastically the cost and delay for building complex IDE. Another advantage is that it lEADS in a natural standardization of the interface between the different components of the system. Therefore, it will be Easier to develop and integrate new functionalities in our test system. In a simplified way, providing an EPICS connection to our text system extension seems sufficient to ensure connectivity. The Open source strategy based on popular component is also, at mid and long term, a guaranty to stay at the state of the art in term of functionality and ergonomic due to the large community which improve continuously the quality of the products. In EUROCOPTER, and especially in the Test system department, it is really a 180-degree turn compared to the 20 passed years. We must go from a complete proprietary model to an Open source one and of course if we hope to maximize the advantages we can expect of this new way of work, we must become actors and contributors of the Open Source.

D. SDMU

SDMU [11] is an internal project dedicated to simulation of avionic systems. One of the main difficulties in avionic system development is to produce a complete and consistent interface definition. The main goal of SDMU will be to simulate interfaces between system components (as shown on Figure 8) in order to fix equipment specifications as early as possible to reduce the lead time of integration tests on rigs and flight tests.

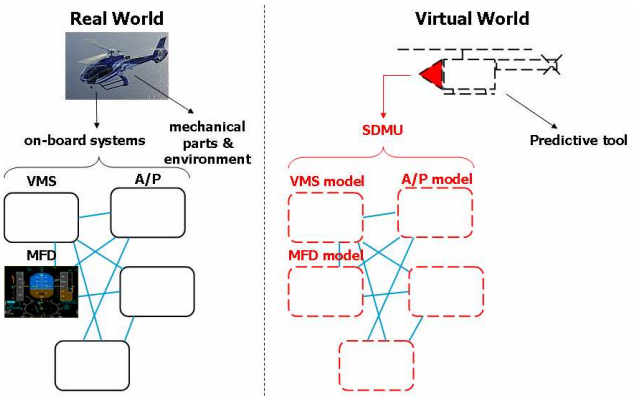


Figure 8. SDMU concept

1) Description of SDMU objectives

The simulation general objective is to increase Validation & Verification activities earlier in the development cycle on a mock-up before performing classical rigs and flight tests qualification.

Specifically, SDMU will be used in a static mode to refine the consistency of systems interface definitions and in a dynamic mode to control the functional behaviour. It allows the simulation of complex (redundant) architectures using the real data flow (A429, A653, Analogues, Discretes) for a realistic representation and an easier coupling with real equipment (hardware in the loop). Last, but not the least; SDMU is an automatic generation process based on the Avionic System definition.

2) GILDA

GILDA is a new tool developed by EUROCOPTER for IMA/ARINC 653 [23] avionics and takes also into account ARINC 429, analogs and discretes. GILDA is based on a set of XML files which are managed by configuration via SVN. The goal of GILDA is to collect information in order to:

- Describe system and equipment communication (as well as internal A653)
- Define equipment and partitions outputs (extension of ADBS - Avionics Data Base System which the central database for all avionics signals - to data structures exchanges)
- Define equipment and partitions inputs

GILDA description is then used to generate IRS (Interface Requirement Specification) documents (for CIGALHE equipment and for partitions), to feed ADBS in order to configure bench and flight test tools and to feed SDMU.

3) SDMU simulation building process

The 5 steps of SDMU process presented on Figure 9 are explicated hereunder.

Step 1:

The first step required to build SDMU simulation would be Interface analysis. Due to processes included in the forecast SDMU tool set, all the interfaces would be analysed to check the consistency of data exchanges, then a set of files would be produced for all the signals exchanged through the system and all the functions needed to manage these signals.

Step 2:

The second step would consist of using a specific process to translate all the interfaces described in the database into files that are usable by the SDMU simulation (libraries, headers).

Step 3:

Today, almost all of EUROCOPTER's sub-system specifications are entirely defined using SCADE® tool. The forecast tool would integrate a C-code generator to translate functional description sheets into C-code files. These generated files would be then embedded in SDMU simulation to be managed as other simulation models are.

Step 4:

The HMI partition used to display flight and vehicle data to the crew is defined using SCADE® for the logical part and VAPS® for the graphical part. These two tools include a C-code generator to obtain a graphical model communicating with other models through ARINC 429 exchanges.

Step 5:

In order to emulate the assessed sub-systems, a set of models are to be developed to "feed" the simulation with representative data from the physical world. All these models would produce or receive data in engineering format (physical data) which will have to be translated into interface value like frequencies or voltages (raw data). To perform this translation, a set of sensor models or/and actuator models will be developed. To encode or decode values from engineering data format to raw data format (or vice-versa) these models would use functions and services provided by the forecast SDMU translator. They would be thus automatically updated when a new simulation is to be integrated after delivery of a new version of the centralised interface data base or a software upgrade.

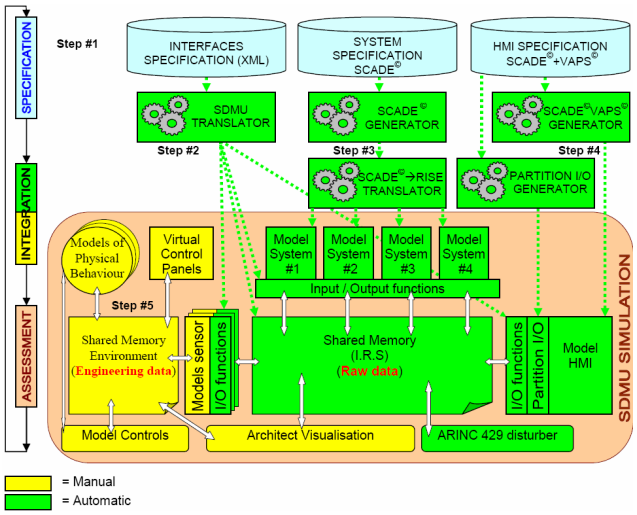


Figure 9. SDMU simulation building process

4) SDMU features and application domains

The main characteristic of SDMU is a powerful simulation addressing low cost hardware platform, running on standard PC or Laptop and compatible with Unix/Linux environments.

It covers many application fields like Rapid prototyping, simulation including hardware in the loop capability, training media framework. SDMU is based on RISE (Real time Interactive Simulation Environment), which is the EUROCOPTER internal simulation tool. Later, SDMU and RISE will be integrated into the EASI framework.

An SDMU implementation in RISE tool is presented hereunder on Figure 10.

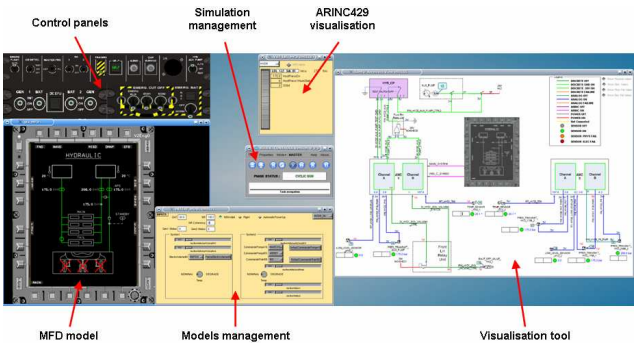


Figure 10. SDMU implementation in RISE

E. CENTRAL REPOSITORY FOR MODELS

To reach the virtualization objectives, it will be necessary to have first a repository where the interfaces of every equipment will be managed in configuration including all input and output signals and secondly a flexible simulation platform allowing to integrate models together building a realistic simulation.

The models forecasted to be used for simulating the system would be:

- Realistic simulation of the helicopter flight
- Realistic simulation of avionic equipments
- Models of the physical behavior of components included in the system simulation (hydraulic system, electrical generation system, flight dynamics, …)
- Sensor or actuator models that translate engineering data from the physical world to system format parameters,
- On-board software models (Vehicle Management System, Automatic Flight Control System, etc).

The management of models is becoming a central topic. In this idea, it is planned to develop a new tool called EASI Repository. This tool aims to centralize all objects used to develop, test and simulate Avionic systems as described on Figure 11.

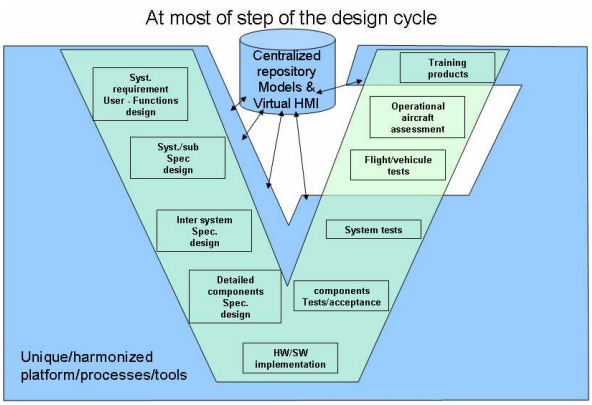


Figure 11. Central repository of models – EASI repository

In a first step, EASI Repository shall contain Avionic Interfaces (equipment, frame & data definitions) and hardware logic definition (FPGA, BSP, relay, diodes).

In a second step, EASI Repository should contain models (Equipment models or re-hosted code, Virtual Panels), Environment codes (Aerodynamic, atmospheric, sensor), simulation panels and dashboards.

EASI repository will improve sharing Virtual aircraft unitary bricks and services between developers through project life cycle. It will ensure that avionic definitions, system models (generic and specific) and environment models are defined and developed once by the specialist and offer to the system community. All EASI repository services shall be based on a strong and rigorous configuration management linked to the project.

Some standard formats should be defined to ease integration and adaptation of repository bricks to the simulations. These standards will authorize automated process for building simulations by connecting models according to the avionic architecture and interface stored in EASI repository. The automation based on specific communication layers services allows retargeting avionic software without modification constraints. The communication between the simulated avionic interfaces shall be transparent and must offer internal monitoring capability with no added effort of the end user.

V. PERSPECTIVES

EASI Tester prototype phases are completed. The EASI Tester operational mockup phase is on the way and will be completed in September 2010. On a Linux quad core platform, we have demonstrated:

- The capability for EUROCOPTER to integrate real-time plug-in to manage its current hardware including hardware located in external VME crates;
- The performance level to perform representative avionic real-time simulation connected to avionic equipments through multiple IO including ARINC 429, Mil1553B, analog, discrete;
- The advantage which should be expected of the Open source strategy.
- The possibility for EUROCOPTER and EADS TEST&SERVICES to share the development of a complex avionic IDE.

We plan to deliver into production the first EASI bench in early 2011. SDMU is already operational on desktop benches and allows anticipating system interface tests and validation.

The next tasks will consist in:

- Launching a model driven approach proof of concept at system requirement level;
- Building progressively a convergence between EASI and SDMU in order to set up hybrid platforms mixing simulation and hardware
- Building a global repository allowing to share within EUROCOPTER: avionic definition, simulation models, flight test data.

REFERENCES

[1] N. Belanger and JP. Lebailly, “Promoting tests system as productivity enablers: The EASI case study”, IARIA, ICONS 2010, International Conference on Systems, pp. 66-70 , April 11-16 2010, Les Menuires, France.

[2] VERSA Module Eurocard (IEEE 1014) http://en.wikipedia.org/wiki/VMEbus

[3] Object Management Group, “Executive Overview: Model-Driven Architecture”. http://www.omg.org/mda

[4] N. Belanger, N. Favarcq and Y. Fusero, “An open real time test system approach”, IARIA, VALID 2009, IEEE International Conference on Advances in System Testing and Validation

Lifecycle, 20-25 September 2009, pp. 38-41, Porto, Portugal
[5] D. Flater, "Impact of Model-Driven standards", National Institute of Standards and Technology, Gaithersburg, USA, HICSS, vol. 9, pp.285, 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 9, 2002
[6] Lockheed Martin aeronautics, Lockheed Martin (MDA Success Story), www.omg.org/mda/mda_files/LockheedMartin.pdf, January 2003
[7] N. Pontisso, D. Chemouil, "TOPCASED Combining Formal Methods with Model-Driven Engineering", French Space Agency, Automated Software Engineering, ASE '06. 21st IEEE/ACM International Conference, pp. 359-360, Toulouse, France, 2006
[8] "Modernizing Software Development Through model-Driven Development", Forrester Consulting, A commission study conducted by Forrester Consulting on behalf of Unisys, 13rd of August 2008, Headquarters, Forrester Research, Cambridge, USA.
[9] T. Schlichtherle, "Open Source Business Models in Transition, Part 2: life cycles, market sectors and new competition strategies", yeebase GbR, 2007
[10] A. Raundahl Gregersen, N. Jørgensen, "Extending eclipse RCP with dynamic update of active plug-ins", Journal of object technology, Vol. 6, No. 6, pp. 67-89 July-August 2007
[11] E. Guillon, N. Brisset, N. Damiani and M. Goasdoué, "System Digital Mock-Up – A new approach for the design and development of helicopter avionic systems", 35th European Rotorcraft Forum conference, 25-28 September 2009, Hamburg, Germany
[12] G. A. Lewis, B. C. Meyers and K. Wallnau , "Workshop on Model-Driven Architecture and Program Generation", Software Engineering Institute, Carnegie Mallon University, Technical Note CMU/SEI-2006-TN-031, 2nd June 2006, Pittsburgh, USA
[13] Office of the Under Secretary of Defense For Acquisition, Technology, and Logistics, "Report of the Defense Science Board Task Force on Integrated Fire Support in the Battlespace", Washington, DC, October 2004, http://www.acq.osd.mil/dsb/reports/ADA428791.pdf
[14] I. Hooks, "Writing good Requirements", Proceedings of the Third International Symposium of the NCOSE - Volume 2, 1993
[15] N. Belanger, J. Bovier, JF. Gilot, JP. Lebailly and M. Rubio, "Multi core computers and PCI express, the future of data acquisition and control system", ETTC 2009, European Telemetry Conference, 24-26 June 2009, Toulouse, France.
[16] Eclipse CDT homepage, http://www.eclipse.org/cdt/
[17] Eclipse EMF homepage, http://www.eclipse.org/modeling/emf/
[18] Subversion homepage, http://subversion.apache.org/
[19] Eclipse subversive plug-in, http://www.eclipse.org/subversive/
[20] CSS homepage, http://css.desy.de/content/index_eng.html
[21] ARINC 429 Wiki, http://en.wikipedia.org/wiki/ARINC_429
[22] MIL-STD-1553 Wiki, http://en.wikipedia.org/wiki/MIL-STD-1553
[23] ARINC 653 Wiki, http://en.wikipedia.org/wiki/ARINC_653
[24] IRIG, Inter-Range Instrumentation Group time codes Wiki, http://en.wikipedia.org/wiki/IRIG_timecode

Federated Identity Management in a Tactical Multi-Domain Network

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

***Abstract*—Identity Management maintains information regarding actors of an Information System, like users, equipment and services. One important service is to disseminate and validate *credentials* for the purpose of *authentication* and *access control*. Within the context of military tactical communication network the identity management services should, due to the disadvantaged nature of these networks, minimize their network demand and connectivity requirements. Security protocols for tactical network should be efficient, prudent and be based on well justified use cases. The contribution of this paper is the rationale and the prototype of an identity management system designed with these properties in mind, including services for authentication and access control. The discussion will suggest a set of architectural patterns for the development and deployment of an identity management system, as well as justifications for the simplified protocol operations.**

***Keywords*-Identity management, Disadvantaged networks, Authentication**

I. Introduction

This article is an extension of a previously published conference paper [1].

Current systems for Identity Management (IdM) are constructed with an emphasis on "Single Sign On" (SSO) service and authentication of clients. *Federations* of IdM systems are supposed to recognize clients from other domains, but are often found to require replication of user registries or the creation of a separate "federated" user registry. They are observed to require identification and enrollment separate from any existing public key infrastructure (PKI) and do not offer authentication of services. Existing IdMs are also seen to rely on invocation of authentication servers for each authentication process, resulting in larger bandwidth and connectivity requirements. [2]

These properties are inadequate and inefficient in the perspective of a dynamic military network of large scale. A military network will extend its operation across wired and wireless links, and observe a wide range of data rates and connectivity conditions. One network will connect to other networks belonging to other parties of a coalition in order to offer well managed services without loss of autonomy.

A. Federablity and Ubiquity

The ability for an IdM to enter into federations with other IdM domains in a scalable, controllable and manageable manner is called its *Federability*. The *Ubiquity* of an IdM indicates its ability to operate on a wide range of equipment and network environments. The properties of an IdM which decide its ubiquity and federability will be identified in the course of this article.

Associated with identity management is the process of *Authentication*. Authentication relies on the provision of identity credentials and offers "establishment of identity". Authentication protocols can be designed for disadvantaged environments with small demand for network resources based on a realistic threat analysis: some threats are too far-fetched to justify expensive protocol details.

It is the purpose of this article to offer an analysis on how an identity management system may be deployed and operated in a tactical, disadvantaged network. The article will also provide a description of a prototype IdM for disadvantaged networks, and the rationale of its design.

The remainder of the article is organized as follows: Section II provides a short introduction to Identity Management Systems, and Section III discusses necessary properties for a successful IdM. Section IV briefly discusses limitations of existing IdM systems, and Section V introduces the prototypical Gismo IdM which is used for experimentation and a main object for this article. Section VI discusses the properties of authentication protocols used in disadvantaged networks, while Section VII presents the experimental protocols in detail. Section VIII presents the challenges when porting the Gismo IdM to the Android platform. Finally, Sections IX and X presents experimental plans and some conclusive remarks.

II. Motivational Background

Identity Management (IdM) are collections of services and procedures for maintaining subject information (key pair, roles) and to issue credentials for the purpose of authentication, message protection and access control. From the client perspective, the credentials issued by the IdM services enable it to access many services inside a community under the protection of mutual authentication and encryption. From the server perspective, IdM enables it to offer its credentials to clients in order to provide service authentication and to control access to its resources.

A. Federated Identity Management

Several federated IdM schemes have been developed, some of which offer single sign on (SSO) for web clients [3], [4], [5]. The SSO protocols exploits the redirection mechanism of HTTP in combination with cookies and POST-data so that an Identity Provider (IdP) can authenticate the client once and then repeatedly issue credentials for use within the federation. This arrangement requires IdP invocation for each "login" operation, and does not offer mutual authentication, i.e., service authentication. [2].

In the situation where the client is an application program (rather than a web browser), there are more opportunities for the client to take actively part in the protocol operations, e.g., by checking service credentials, contacting the IdP for the retrieval of own credentials, caching those credentials etc. The research efforts presented in this paper assume that the clients are able to run custom built programs.

The usual meaning of the word "federated" is that several servers share their trust in a common IdP for subject management and authentication. It does not necessarily imply any trust relationship between independent IdPs so that they can authenticate each others' clients. For the following discussion, we will call the group of clients and services which put their trust in the same IdP as a *community of interest* (COI). A trust relation between independent IdPs is called a *cross-COI relation*.

B. Mobile and Federated IdM requirements

An essential property of an IdM is its ability to integrate with other components for management of personnel and equipment. The list below contains other necessary properties, some of which are to be explained later in the article.

- An IdM should be able to use resources from the existing PKI (keys, certificates, revocation info) and offer its services to different platforms, with different presentation syntax and for different use cases.
- An IdM instance should be able to form trust relations with other IdM instances in order to accommodate guests and roaming clients.
- An IdM must provide support for role/attribute based access control
- An IdM must support protocol operations for mutual authentication.

For IdM used in mobile systems, there are requirements related to the resource constraints found in these systems:

- The protocol operations of an IdM must use small PDU sizes and allow the use of caches to reduce the system's connectivity requirements.

C. The relation between IdM and Access Control

Services can enforce access control on the basis of the *identity* of an authenticated client, or based on *roles* or *attributes* associated with the client. For the purpose of the accommodation of roaming users, it is absolutely necessary to make access control decisions based on roles/attributes, not identity. "Identity based" access control requires that all roaming clients are registered into the visited IdM, which is an unscalable solution.

The principles of *Role/Attribute Based Access Control* (RBAC/ABAC) are well investigated [6]. The names and meaning of the roles/attributes that are used to make access decisions must be coordinated as a part of an IdM trust relationship. For this reason, the number of roles/attributes used for access control needs to be kept low.

It is the responsibility of an IdM to manage the roles/attributes of a subject, some of which may enter into access control decisions, others may be used by the service for configuration purposes etc. The presence of subject attributes is the main functional difference between IdM credentials and X.509 public key certificates.

III. Candidate Design Patterns

One of the contributions of this article is a set of proposed design patterns for the construction of a scalable IdM with loose coupling between management domains. The patterns are:

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 prevent/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 retained 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 the public sector. Other PKIs will provide keys for banks, others 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 control [6] 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 any 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 [7].

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 [8]. 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 resource requirements do 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 time, and will not need to be revalidated during 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 IdP 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 [9].

IV. Existing IdM Architectures

The proposed design is related to the SAML 2.0 architecture for federated identity management [10] and the WS-Security [11] and WS-Trust standards [12], 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 [5], Shibboleth [3], and OpenID [4], 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) [6] 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 article refers to federation of servers from different communities with independent 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.

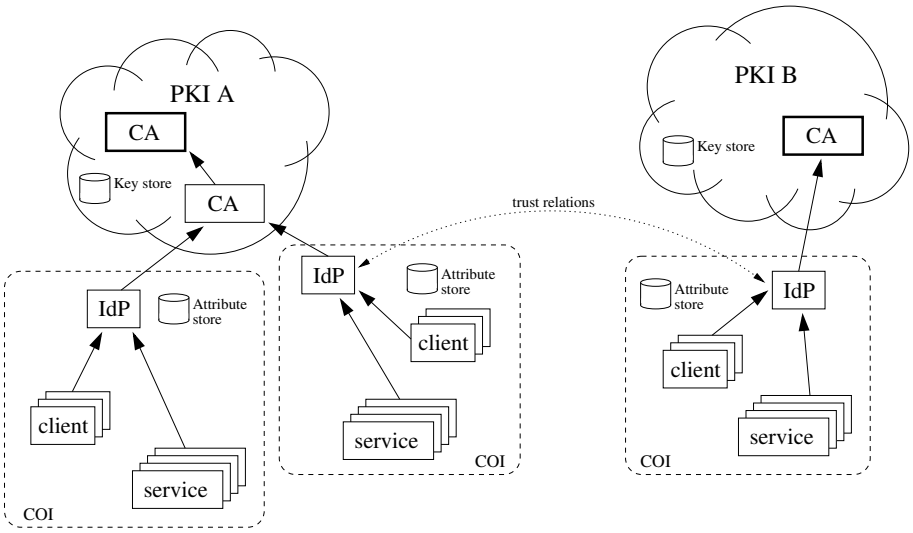


Figure 1. The functional components of the Gismo IdM system.

V. The Gismo Architecture

Following the guidelines given in Section III, an IdM prototype has been built for the purpose of experimentation [13]. The prototype has been implemented in Java for operation in a service oriented environment. The protocol data units have been given two different syntax representations:

1) For the operation in a Web Services environment the PDUs are coded in XML-based syntax like SOAP, WSSec, SAML etc.
2) In environments with poor support for WS-standards (like the Android platform), the PDUs have been coded as serialized Java objects, from now on simply called "POJO" (Plain Old Java Objects).

The presentation layer diversity is discussed in Section VIII.

The functional components of the Gismo IdM and their relations are shown in Figure 1. 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. ("Gismo" is the acronym for the Norwegian expression "Fundamental IT security for mobile operations".)

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 subjects (either client or server), and they can be member 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 if they are 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 the 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, generated by this IdP.

Please observe on Figure 1 that even there exist PKIs and CAs which issue X.509 certificates, they are only visible to the IdP. The COI members only relate to the IdP services and the identity statement. The PKI services may therefore be replaced with different technology without affecting the COI members.

E. 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 IS* link the signature on an IS from a remote COI to the IdP of the local COI, and conveys the delegation of trust from the local IdP to the remote IdP.

F. 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 Section III-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 since it generates much network traffic and breaks the encapsulation of the underlying PKI.

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 may have several forms, depending on whether the trust anchor 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 are 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. One cannot assume that a key is valid only because it is not listed as revoked. What is needed is a positive revocation status (meaning not revoked), which is 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 Section III-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.

In essence, the separation of the IdP from the trust anchor is a matter of reducing the number of trust anchors, as well as a matter of *trust anchor protection*. An IdP is reachable for everyone, and with a trust anchor key inside it becomes an attractive attack target.

G. 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, that 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 with 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).

In those cases where the intra-COI traffic need to be protected from other activities on the same network links, a Virtual Private Network arrangement should be employed.

Since the use case of our prototype is related to military applications, the value of privacy protection has not been highly regarded.

VI. The Authentication Protocols

Several authentication protocols have been devised 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.

The "proof of possession" principle is in common use in authentication protocols. The requester proves that it is in possession of a secret (that only this subject knows) in order to prove its identity. The secret can be a private key, and the proof of possession can be implemented in at least two ways:

- The requester can sign the request message with its private key
- The responder can encrypt the response with the public key of the requester.

Although these options both offers authentication, only the first also offers protection of message integrity. Without this protection, an attacker may alter the content of the request without detection.

Another attack scenario is the *replay attack* where the attacker gathers messages in transit and re-injects them into the network at a later stage. For these attacks to be detected, messages can be timestamped (so excessively old messages are discarded) and duplicates should be recognized within the allowed time frame.

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.

A. Stateless services do not require replay protection

Under the conditions that the service is stateless, i.e., a request is not altering the state of the system (e.g., 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. For protection of message integrity, the client should add a signature to the request message.

B. Authentication "as we go"

Another matter is the number of protocol round trips. During a separate authentication phase, client and service can mutually authenticate themselves before the actual service call is made. A more effective approach is to piggyback the client authentication on the service request, and the service authentication on the response, as shown in Figure 4. 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 eavesdropping than a "hit and run" tactic. A fraudulent service which is not able to authenticate itself would trigger an intrusion alarm and a 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.

The replay protection of request messages is difficult because client may approach the service for the first time, when no shared state can exist common to both. For the response message the protection is easier to obtains. A nonce in the request message may be copied to the reply message and protected under the signature of the service. The request message establishes a shared state which simplifies the subsequent replay protection.

C. Clock skew elimination

For the purpose of reducing the period during which messages need to be remembered for duplicate detection may be greatly reduced through the presence of a common clock source and an associated synchronization protocol. While in a tactical network a separate synchronization protocol like NTP consumes costly bandwidth, the choice of this experiment has been to piggyback clock information in the messages from the IdP.

The experimental clock protocol is illustrated in Figure 2 and its operations are performed as follows:

1) The IdP includes with every issued identity statement a clock value, which is the number of milliseconds since a chosen t_0.
2) Upon reception of the IS, the subject starts counting milliseconds starting with the clock value included in the IS.
3) When sending a request message, the client includes the current millisecond counter value in the request, and protects this value with its signature.
4) When receiving a message, the service compares the counter value in the request with its own counter value (initiated with the counter value of its IS). The message is rejected if the difference is outside an allowed range.

Figure 2 shows how an IS issue at time t_x is marked with the timer value t_x, but received at $t_x + \Delta t_x$. The subject's counter at value at time t_z is therefore $t_x + t_z - (t_x + \Delta t_x)$, so this value is included in the service invocation which takes place at t_z. The IS issued for the service is marked with t_y, but received at $t_y + \Delta t_y$, at which point the service starts counting from the value t_y. The service receives the request at $t_z + \Delta t_z$, at which time its counter value is $t_y + t_z + \Delta t_z - (t_y + \Delta t_y)$. The difference between this value and the value included in the request is $\Delta t_z + \Delta t_x - \Delta t_y$. The network delays during IS issue cancel each other out and reduce the variance

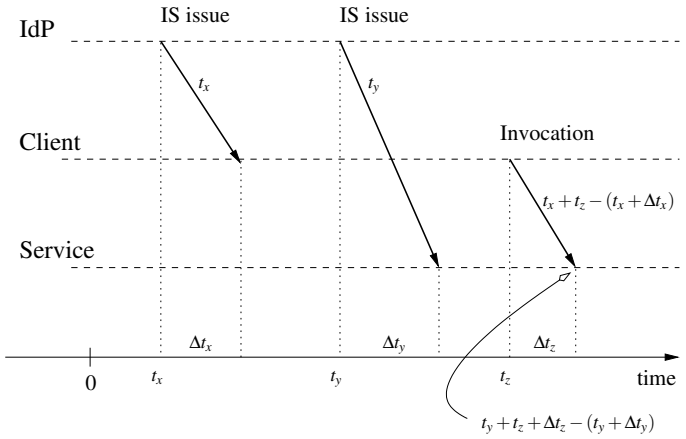


Figure 2. The clock synchronization mechanism included in the IdP and service invocation protocols.

of the expression. The clock drift during the IS issue interval is regarded as negligible.

The request message should be discarded if the value difference exceeds a threshold interval. The treshold interval should be chosen to give a low probability for false positives (due to network delay) yet maintaining a fair amount of protection.

Even without detection of duplicate messages, replays are now given a very short time window to succeed. It is a claim by the author, supported by observations in the cyber defense community, that replay attacks that must take place within one second (example value) have very limited application. A successful replay attack is likely to be the result of a period of traffic eavesdropping and analysis, followed by injection of carefully selected messages, all in a manner that minimizes the probability of detection. A replay attack for the purpose of *Denial of Service* is highly unlikely since the advanced position and capabilities of the attacker will be immediately detected and eliminated.

It is therefore proposed that with a clock synchronization scheme like the one outlined above, detection of duplicates is not strictly necessary. If duplicate detection is required, it is much simpler to implement since the time window is much smaller. If the server crashes, it can simply withhold its service for the duration of the time window during restart. This details relieves the server from the need to recognize duplicates across service incarnations, which was identified as a costly operation earlier in this section.

VII. 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 article will commence with a description of the data structures and protocols in greater detail.

A. The Identity Statement

As previously described in Section V-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:

$(Id_x)_a = Name_x + PublicKey_x + Attributes_x + TimeCounter + Signature_a$

$Attributes_x$ 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 COI a. The IdP of COI a will from now on be denoted IdP_a.

In the proposed system, the identity statement is formatted according to one of the following methods:

- 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.
- As serialized Java objects.

A discussion on the existence of parallel syntax representations will be given in Section VIII.

B. Identity Statement Issuance

The discussion in Section V-G identified the need to protect subject attributes outside the Community of Interest (COI), which means that only members of a COI should 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. Prior to the issue of an IS, the subject need to have its private key loaded, through which it can prove its identity to the IdP. The proof can be implemented as:

- an SSL-based authentication phase prior to the IdP invocation
- a signature on the IdP request (susceptible to replay attacks)
- encryption of the IdP response with the subject's public key.

In Figure 3, which shows the IS issue protocol, the client authentication is not shown, but regarded as a protocol implementation detail.

A part of the IdP 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 III-A) the IdP should use the available PKI-based validation mechanisms for this purpose, and deny the issue 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 V-D. Presented with

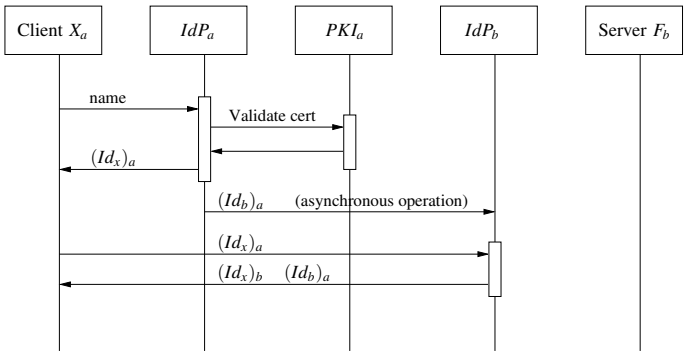


Figure 3. The identity statement issuing protocol. In this case a guest IS is being issued in two steps.

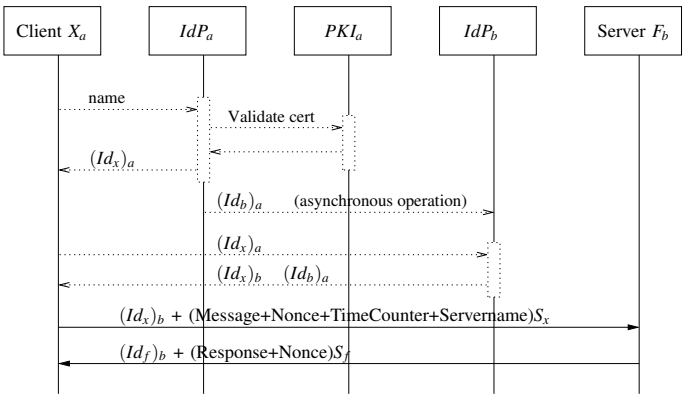


Figure 4. The authentication protocol for the stateful service.

$(Id_x)_a$, 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 $(Id_x)_b$, 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 issuance.

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

Figure 3 illustrates the guest IS issuing protocol as a two stage process involving two IdPs. Key validation takes place only in the first stage. The optional proof of validity (Section V-F) is assumed to have been issued at an earlier occasion.

D. The Authentication Protocol

Section VI 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 4 shows a protocol suited for a server with the necessary resources to implement replay protection. The data elements needed for mutual authentication (signature, timecounter, nonce, servername) are *piggybacked* on the request and response messages in order to save a protocol round trip. The remaining security

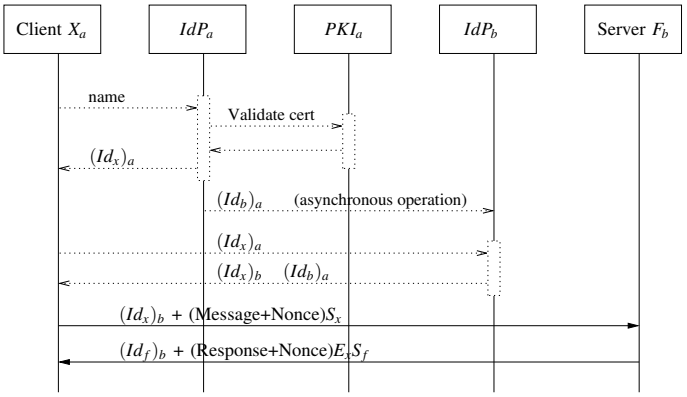


Figure 5. The authentication protocol for the stateless service.

risk, which results from this choice is marginal, is pointed out in Section VI. The optional duplicate/replay detection happens locally in the server and does not affect the protocol data units.

Figure 5 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 request is signed to protect its integrity. 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. A replacement for the X.500 Distinguished Name

In an X.509 public key certificate, the subject is identified though the use of an X.500 Distinguished Name (DN). The author's DN might be: `CN=Anders Fongen, O=FFI,C=NO`. Several forms are likely, and each form can have a number of string representations, e.g., `/CN=Anders Fongen/O=FFI/C=NO`.

During an authentication process, this is the subject identifier that relates to the operation taking place.

A signed document, sent from the E-mail address `anders.fongen@ffi.no`, relates its signature to the identifier `CN=Anders Fongen,O=FFI,C=NO`. It is not possible to decide if the two identifiers relate to the same subject. It is possible to write X.500 DN in a form that maps to an E-mail identifier in the RFC-822 form, but in general, there are no mapping rules.

The X.500 DN made sense back when there was X.400 E-mail which used this form for message addressing. Today, when the RFC-822 form is prevalent, authentication should take place on the E-mail address of the subject, or other identifiers that relates to the subject in subsequent operations (like DNS-name or IP address). One reason that X.500 DN is still in use is that the LDAP directory protocol mandates its use as a lookup key.

The Gismo IdM has been built to use the *Subject Alternative Name* extension of the X.509 certificate (created by the PKI) in the generation of identity statements, which means that the X.500 DN is visible only to the IdP, not to the subjects. This extension can hold the subjects RFC-822 E-mail identifier, a DNS domain name, and more. It allows for a more straightforward processing of access rights, usage policies etc., since the authenticated identifier is more intuitively related to the subject.

Due to the use of X.500 DN in the LDAP lookup protocol, it is not feasible to lookup a certificate using the Subject-Alt-Name extension. The X.500 DN must be used in the IdP request for an identity statement. This poses no problem, since the subject knows its own X.500 DN from its certificate (which must be preinstalled). It is also necessary to use X.500 DN in cross-COI identity statements for the same reason.

VIII. Identity management for mobile units

The inclusion of mobile units in a framework for identity management and mutual authentication is highly desired since the focus of this research is tactical military systems. The resulting design will to a large extent be influenced by the resource situation in mobile networks; Narrow bandwidth, frequent disconnections and network partitions, limitations of software platform and user interface.

The choice of this prototype has been to include mobile units which are based on the Android operating system [14]. Android systems are readily programmable in the Java programming language. Java programs are compiled into a distinct bytecode and executed in a virtual machine called Dalvik. Source code portability from Java Standard Edition (Java SE) to the Android platform is good, but limited by the availability of some builtin packages. Packages related to user interface, remote invocation or J2EE operations will not port. Otherwise, the portability allows the relevant Gismo IdM code to compile to a Dalvik engine with little effort. An important property of Dalvik is that the object serialization format is compatible with the Java VM so that the two platforms can exchange their native objects in serialized form.

Gismo IdM was initially built with the use of XML syntax representation of external data units. The identity statements were represented as SAML assertions [10] and the service requests/responses were encoded according to the SOAP message standard and the authentication protocol data was put in the SOAP headers using the WSSec standard [11]. The software library used to support the manipulation of SAML and WSSec objects (Sun XWSS 2.0) is not available under Android and will probably never be. Besides, the SAML and WSSec standards are so complex that "barefoot" processing using string operations etc. is not feasible.

Therefore, the port of Gismo IdM to Android required a different syntax representation of external protocol data units. The choice was made to use serialized Java objects for this purpose. Serialized Java objects (from now on denoted POJO - Plain Old Java Objects) can be exchanged between any Android/Dalvik, Scala and Java SE systems (although not Java ME, which lacks a serialization engine). The choice was made for reasons of network efficiency and ease of programming.

Of course, moving from a platform neutral XML syntax to a platform specific POJO representation affects interoperability. Gismo IdM participants do not talk to other IdM systems, since it employs non-standard protocols. Therefore, interoperability becomes more of a *portability* problem, i.e., which systems can the Gismo IdM implementation code be ported to?

The SOAP based implementation can only be ported to systems with library support for WSSec and SAML. The Java library used in the prototype (Sun XWSS 2.0) provide good portability for Java SE enabled platforms. For other systems, lack of SAML support may inhibit the port since it is not feasible to write that code from scratch. An in-house attempt to port the authentication protocol to .NET platform failed due to bugs in the WSSec library.

The POJO based implementation can only be ported to systems with the Java serialization engine. Such systems include Java SE, Dalvik and Scala, although the latter has not been tested.

It is therefore not true that an IdM is portable simply because it uses "open standards", due to the sheer complexity of the specification. Neither is it true that a POJO based protocol is less interoperable than a SOAP/SAML based protocol.

No "one-size-fits-all" solution appears to be available, but rather than making separate stovepipe IdM systems for the different environments, it appears sensible to design an IdM which allows different syntax representations of the data elements to co-exist.

A. SOAP vs. POJO interoperability

The GISMO IdM contains nodes which use different presentations for identity statements and service invocations. In order for two nodes to communicate, they must use the same communication stack, including the presentation layer. A client using serialized POJOs can therefore not communicate with an IdP or a service requiring SOAP message syntax and vice versa. For a client to reach the services it needs, regardless its choice of presentation syntax three approaches can be taken:

1) Make services (and the IdP) dual-stack.
2) Make a general proxy for automatic conversion between the presentation forms (POJO and SOAP), e.g., based on Sun JAXB.
3) Make a specific proxy for each service

Option 1 is not a possible solution, since we introduced the dual representation form precisely due to the lack of SAML/WSSec support in mobile systems. Some nodes may be equipped with two stacks, but not all.

Option 2 has not been studied in detail, but requires a combination of WSDL-compilation and JAXB-assisted conversion. It is not likely to be possible to convert on-the-fly any POJO to a SOAP message which conforms to the WSDL-file of a particular web service.

Option 3 has been studied and tested, and represents an attractive approach. A proxy service takes the parameter values and passes them to a precompiled web services stub (generated by the WSDL compiler). The return value from the stub is passed back to the caller of the POJO service. Example code lines required for this function are shown below:

```
public class MainClass {
  public Serializable service(WeatherRequest wr,
                          Properties props) {
    try {
      Weather w = new Weather();
      String result = w.getWeatherSoap()
         .getWeather(wr.town);
      return result;
    } catch (Exception e) { return e; }
  }
}
```

Option 3 is also attractive since it gives the developer control over service aggregation and orchestration. One service call to a POJO service need not be passed on as one single web service invocation. Many individual calls may be made, and they may be sequenced or tested in any manner. Aggregated operations are useful because they potentially reduce the network traffic to and from the mobile unit, which is likely to be connected through a disadvantaged link. The proxy can even cache results for subsequent service calls.

For options 2 and 3 there is a problem related to signature values. Equivalent POJO and SOAP messages will have different signature values, and the integrity of the message is broken during a conversion. The proxy can sign the converted object using its own private key, which would require that the service accepts that the proxy vouches for the original client in the authentication phase.

IX. Experimental evaluation

The Gismo IdM has been subject to a series of experimental evaluations, mostly for testing correctness of algorithms and implementation, and to study configuration and deployment options. The experiments have confirmed the correctness of the protocol design and the feasibility of the implementation. During 2012 the Gismo IdM will be a part of larger field experiment where the secure exchange of information within a military coalition will be in focus. Technologies in use will include IPSec, IPv6, XMPP messaging protocol, Gismo IdM, security gateways, Android, Linux etc. The Gismo IdM will be evaluated with the perspectives of interoperability with other security technologies, its performance when XMPP is used as transport protocol, and its stability and resilience when run in a disadvantaged network.

X. Conclusion

The necessity to offer identity management across a wide range of equipment and communication technologies is the background for this article. A number of properties has been identified as important for the successful construction, deployment and operation of an identity management system. Investment protection, domain autonomy and prudent resource consumption are key elements.

The Gismo IdM prototype has been described in detail. It is an implementation of the proposed design principles and a basis for experimental evaluation. Its dual stack implementation of PDU syntax representation raises interoperability concerns which have been discussed.

Further work on the Gismo IdM includes an experimental evaluation in a military field maneuver of mobile coalition partners, where performance and interoperability properties will be assessed. Also, the work on interoperable syntax representation for identity statements and service requests/responses will be continued. The goal is to look for representations that are able to retain the signature integrity to a greater extent that what is presently possible.

References

[1] A. Fongen, “Architecture patterns for a ubiquitous identity management system,” in *ICONS 2011*. Saint Maartens: IARIA, Jan. 2011.

[2] J. Hughes, S. Cantor, J. Hodges, F. Hirsch, P. Mishra, R. Philpott, and E. Maler, *Profiles for the OASIS Security Assertion Markup Language (SAML) V2.0.*, OASIS Standard, March 2005.

[3] "Shibboleth." [Online]. Available: http://shibboleth.internet2.edu/ [retrieved Dec 21, 2011]

[4] "OpenID." [Online]. Available: http://openid.net/ [retrieved Dec 21, 2011]

[5] "The Libery Alliance." [Online]. Available: http://www.projectliberty.org/ [retrieved Dec 21, 2011]

[6] 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.

[7] 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.

[8] A. Fongen, "Scalability analysis of selected certificate validation scenarios," in *IEEE MILCOM*, San Diego, CA, USA, Nov. 2010, pp. 1–7.

[9] 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.

[10] 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.

[11] K. Lawrence and C. Kaler, *Web Services Security: SOAP Message Security 1.1*, OASIS Standard Specification, 2004.

[12] ——, *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 Dec 21, 2011]

[13] A. Fongen, "Identity management without revocation," in *SECURWARE 2010*. Mestre, Italy: IARIA, July 2010.

[14] ——, "Federated identity management for Android," in *SECURWARE 2011*. Nice, France: IARIA, July 2011.

Reconstruction Quality of Congested Freeway Traffic Patterns Based on Kerner's Three-Phase Traffic Theory

Jochen Palmer
Daimler AG
GR/PTA - HPC: 050-G021
D-71059 Sindelfingen, Germany
Email: jochen.palmer@daimler.com

Hubert Rehborn
Daimler AG
GR/PTF - HPC: 050-G021
D-71059 Sindelfingen, Germany
Email: hubert.rehborn@daimler.com

Iván Gruttadauria
UTN-FRVM
Av. Universidad 450
X5900HLR Villa Maria, Argentine
Email: igruttadauria@frvm.utn.edu.ar

***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. In this research Kerner's three-phase traffic theory is applied, which distinguishes two different phases in congested traffic: synchronized flow and wide moving jam. In the presented approach spatio-temporal congested traffic patterns are reconstructed from intelligent probe vehicle information generated by an on-board 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. Comparing *Ground-Truth* with the reconstructed traffic pattern shows that a reconstruction quality comparable to that of established traffic flow models is achievable with probe vehicle equipment rates of about 0.5 %. At higher equipment rates of about 1-1.5 % the achievable quality already exceeds the quality established traffic flow models are able to offer based on a dense detector network with average detector distances of 1-2 km.**

Keywords-Traffic monitoring; Traffic state detection; Traffic data fusion; Probe vehicle data; Three-phase traffic theory; Models ASDA and FOTO; Traffic data quality; Traffic quality indices.

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. In this research the question is answered how many probe vehicles are needed to deliver a quality of traffic data comparable to a dense high quality detector network.

In order to assess the quality of reconstructed traffic patterns, three different quality indices are introduced and assessed. In [16] some basic results have been already presented for the quality index for travel time and the quality index for regions of synchronized flow and wide moving jams. The quality index for fronts of synchronized flow and wide moving jams has been briefly introduced but not evaluated. The other two quality indices have been evaluated based on a congested traffic situation representing a general pattern (GP) according to Kerner's Three Phase Traffic Theory. General patterns occur when there is just a single effective bottleneck. Hence the emerging spatio-temporal patterns are fairly simple.

In this research based on [16] a more detailed assessment of the reconstruction quality is conducted based on an extended traffic situation. The examined situation represents an expanded pattern (EP) according to Kerner's Three Phase Traffic Theory. In this situation traffic patterns emerging at several adjacent effective bottlenecks are overlapping. Hence the resulting combined spatio-temporal traffic patterns show a very complex structure and dynamics. In addition, for the first time, a thorough assessment of the quality index for fronts of synchronized flow and wide moving jams is presented. This allows a detailed discussion of the achievable quality of probe vehicle based models in comparison to stationary loop detector based systems. This research takes all quality aspects of spatio-temporal congested freeway patterns, covered by the three three presented quality indices, into consideration.

Outline

The paper starts with a brief introduction to Kerner's Three Phase Traffic Theory. Basic elements as well spatio-temporal traffic patterns are explained. The traffic models ASDA (Automatische Staudynamikanalyse; automatic congestion analysis) and FOTO (Forecasting of Traffic Objects) based on this theory are described. They reconstruct congested traffic patterns from data measured by stationary loop detectors.

After that the impact of spatio-temporal traffic pattern on vehicles moving through these patterns is discussed.

New methods, which allow the reconstruction of spatio-temporal traffic patterns from data measured by probe vehicles, are briefly introduced. This approach allows a high quality reconstruction of traffic patterns in the whole traffic network without having to rely on stationary loop detectors.

Then three quality indices are introduced and discussed in detail. This includes quality indices for (i) travel time, (ii) fronts of synchronized flow and wide moving jams as well as (iii) regions of synchronized flow and wide moving jams. These quality indices allow the assessment of the reconstruction quality on spatio-temporal traffic patterns. An approach to the simulation of a realistic test dataset is presented.

Based on this dataset and the presented quality indices, the reconstruction quality of the new methods is compared to the reconstruction quality of the traffic models ASDA and FOTO.

II. Kerner's 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) ([5], [8]).

A. Elements of Three Phase Traffic Theory

Empirical macroscopic spatio-temporal objective criteria for traffic phases as elements of Kerner's three-phase traffic theory ([5], [8]) 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.

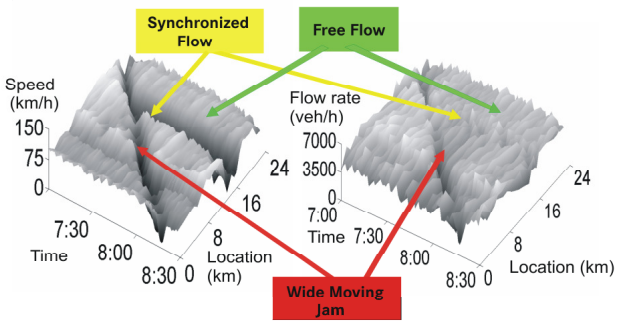


Figure 1. Explanation of traffic phase definitions from empirical data: Spatio-temporal overview of speed (left) and flow rate of traffic (right) on a selected freeway section

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 ([19], [20]).

C. Models ASDA and FOTO

For recognition, tracking and prediction of the spatio-temporal congested traffic patterns, based on stationary loop detectors, the models ASDA and FOTO ([5], [8]) 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 [9]. 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). In recent years other models, which are similar to the models ASDA and FOTO, have also been published [22].

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).

Traffic state changes between the different traffic phases have impacts of different strength on the vehicle and vehicular assistance applications [2]. 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.

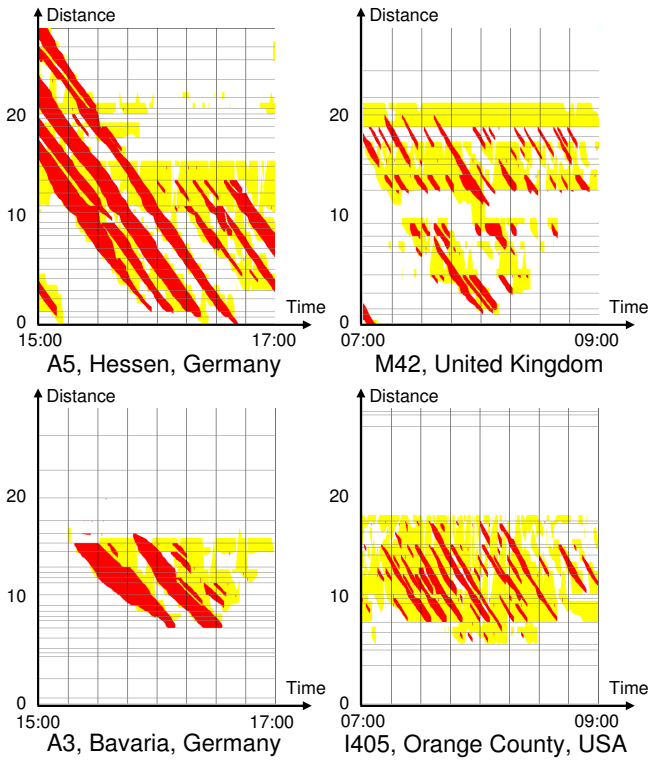


Figure 2. Resulting empirical spatio-temporal traffic patterns when applying the models ASDA and FOTO to traffic data measured in different countries [15]. All these patterns show the same spatio-temporal structure and dynamics: Regions of synchronized flow fixed at the location of the bottleneck and wide moving jams propagating upstream with constant speed.

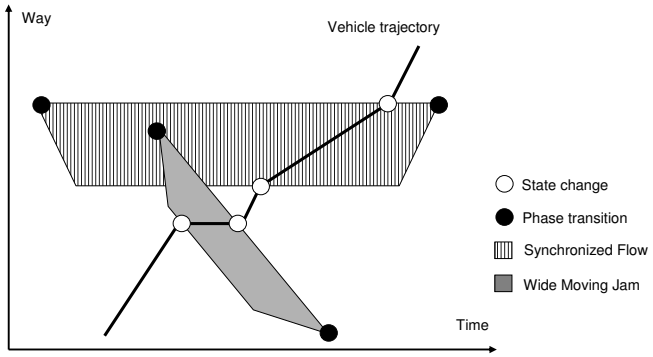


Figure 3. Qualitative explanation of traffic phase transitions and traffic state changes a vehicle experiences on its way through a spatio-temporal congested traffic pattern

Table I
VEHICLE SPEED: CHANGE DEPENDING ON SPECIFIC TRAFFIC STATE TRANSITIONS

State	to F	to S	to J
F	-	Deceleration	Strong deceleration
S	Acceleration	-	Deceleration
J	Strong acceleration	Acceleration	-

Table II
VEHICLE DENSITY: 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	-

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 ([7], [11]). Each traffic state change represents a control and parameter adaption point for these applications. The better the reconstruction of the spatio-temporal structure of congested traffic patterns, the better the knowledge about current and future spatio-temporal positions of traffic state changes and hence the higher the potential of improving the open- and closed-loop control of vehicular assistance applications. Having a high quality knowledge about current and future traffic state changes opens the door for a complete new class of vehicular assistance applications:

- Safety related systems: Vehicular assistance applications, which improve vehicle safety by being aware of sudden speed breakdowns on the path of a vehicle, for example caused by traffic state changes from phase F to phase J.
- Energy efficient systems: Vehicular assistance applications, which improve the energy efficiency of vehicles by reducing fuel consumption by being aware of current and future traffic states.

In addition a high quality reconstruction of spatio-temporal traffic patterns allows the application of existing vehicular assistance applications in better quality. For example a better knowledge of the travel time loss caused by spatio-temporal congested traffic patterns improves the quality of existing dynamic route guidance systems.

IV. RECONSTRUCTION OF SPATIO-TEMPORAL CONGESTED TRAFFIC PATTERNS

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 [4]. 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, [5], [8]).

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

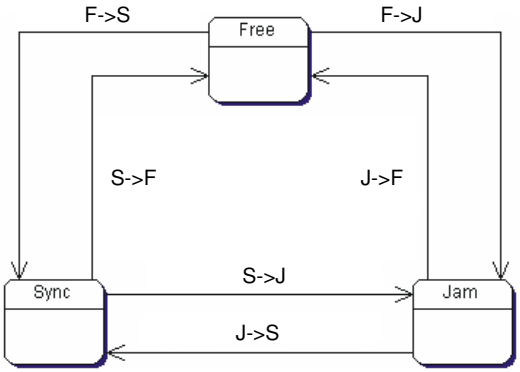


Figure 4. State diagram for three traffic phases [6]

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 ([6], [8]). 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 on-board traffic criteria [10].

B. Cooperative Reconstruction of 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 [17] (Figure 5).

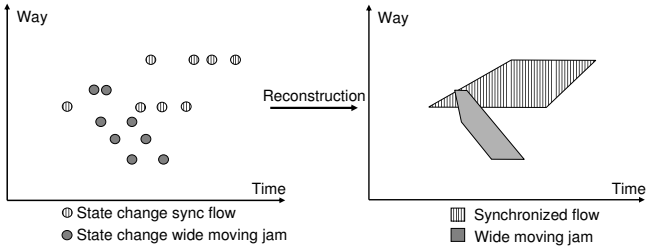


Figure 5. Using clustering to reconstruct traffic patterns from traffic state changes [15]

V. Reconstruction Quality Indices

The reconstruction quality of spatio-temporal congested traffic patterns can only be assessed in the case that the real spatio-temporal traffic pattern is known. Full knowledge about spatio-temporal traffic patterns is only possible, when the positions of all vehicles participating in traffic flow are known for any time instant. This is still difficult to achieve, even with the recent advances in traffic measurement technology. Hence any measured traffic pattern, which represents the *best-known* information about a real traffic pattern, exhibits *some* deviation to reality. The best known traffic information is called *Ground-Truth*. Therefore, when the quality of reconstructed spatio-temporal congested traffic patterns is assessed by comparing them to *any* measured *Ground-Truth* information about this patterns, it is only possible to assess their error compared to *Ground-Truth*, but not their deviation to reality.

In the following sections a formal definition of spatio-temporal congested traffic patterns is introduced. Based on this definition, three different quality indices are presented: a quality index for travel time, a quality index for fronts of synchronized flow and wide moving jam and a quality index for regions of synchronized flow and wide moving jam. Each of the quality indices assesses a different feature of spatio-temporal congested traffic patterns.

A. Formal Definition of Traffic Patterns

In a continuous traffic reality a traffic state TS is assigned 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 spatial borders x_s and x_e as well as the temporal borders and t_s and t_e, 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

$$M := \{TS(x,t) | x_s \leq x \leq x_e \wedge t_s \leq t \leq t_e\} \tag{3}$$

$$Q = D_{M \to R} \tag{4}$$

As mentioned above, currently it is difficult to measure traffic reality. In addition storing and processing continuous spatio-temporal information is not feasible. For example, when stationary loop detectors are used to measure traffic reality, the continuous values of time and distance are actually measured as discrete values. These values 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. In common systems using stationary loop detectors the temporal resolution is in most cases $t_d = 1$ min, while the spatial resolution is in the best case $x_d = 1-2$ km. In many cases the spatial resolution is even lower. This means that the traffic reality is known with even less quality.

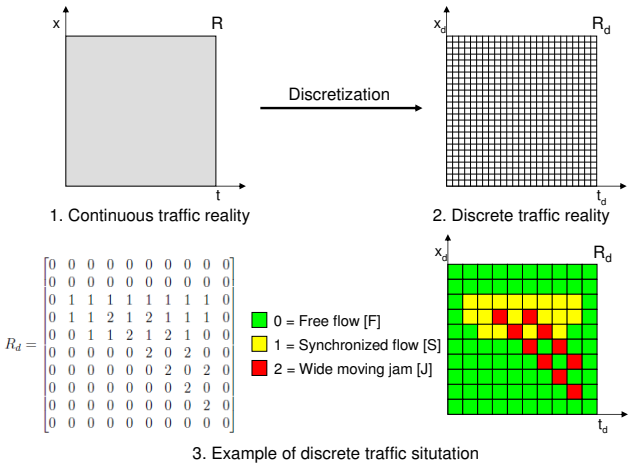


Figure 6. Discretization of spatio-temporal congested traffic patterns

Using x_d and t_d the continuous traffic reality R becomes the discrete traffic reality R_d as shown in Figure 6. By taking the border spatial borders x_s and x_e and the temporal borders t_s and t_e into account, R_d represent information about a traffic pattern TP_{R_d} within some spatio-temporal boundary:

$$TP_{R_d} := \{TS_d(x,t) | x_s \leq x_0, x_1 ... x_n \leq x_e \wedge t_s \leq t_0, t_1 ... t_n \leq t_e\} \quad (5)$$

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. This model represents the reconstructed traffic pattern TP_M:

$$TP_M := \{TS_d(x,t) | x_s \leq x_0, x_1 ... x_n \leq x_e \wedge t_s \leq t_0, t_1 ... t_n \leq t_e\} \quad (6)$$

For Q_d this leads to

$$Q_d = D_{TP_M \rightarrow TP_{R_d}} \quad (7)$$

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 [18]

$$T_{total}(t) = \frac{L_F(t)}{v_F(t)} + \frac{L_J(t)}{v_J(t)} + \frac{L_S(t)}{v_S(t)} \quad (8)$$

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 travel time $T_{total}(t)$ can be determined for road segment based on a reconstructed traffic pattern TP_M and based on a traffic pattern TP_{R_d}, which has been defined as *Ground-Truth*. In this case the travel time through the reconstructed traffic pattern is represented by T_M and the travel time through *Ground-Truth* is defined as T_{GT}. By calculating the relative average deviation ΔT between T_M and T_{GT}, is is possible to assess the quality index for travel time for a given spatio-temporal region [14]. The assessment takes spatio-temporal regions between the spatial border x_s and x_e on N time instants between the temporal boundaries t_s and t_e into account:

$$\Delta T = \frac{1}{N} \sum_{t=t_s}^{t_e} \frac{|T_M(t) - T_{GT}(t)|}{T_{GT}(t)} \quad (9)$$

The quality index given by (9) is represented by a number $\Delta T \geq 0$. A result of $\Delta T = 0$ represents the best possible quality, which is achieved, when T_M and T_{GT} are the same for all time instants t.

C. Quality Index for Fronts of S and J

Positions of phase fronts of the traffic phases *S* and *J* are denoted by $x_{up}(t)$ and $x_{down}(t)$ for each time instant t. Hence, expanded spatio-temporal congested traffic patterns can show several phase fronts at a given time t. Phase fronts can be reconstructed based on a reconstructed traffic pattern TP_M or based on a traffic pattern TP_{R_d}, which has been defined as *Ground-Truth*. For each time t and each *phase* the positions of the upstream front are defined as $x_{up_M}^{Phase}(t)$ and $x_{up_{GT}}^{Phase}(t)$, while the positions of the downstream front are designated as $x_{down_M}^{Phase}(t)$ and $x_{down_{GT}}^{Phase}(t)$, respectively. The reconstruction quality of each single phase front is determined by the relative average deviation of both positions at N time instants between the temporal boundaries t_s und t_e.

The quality of an upstream phase front is given by Δ_{up}^{Phase}

$$\Delta_{up}^{Phase} = \frac{1}{N} \sum_{t=t_s}^{t_e} |x_{up_M}^{Phase}(t) - x_{up_{GT}}^{Phase}(t)| \quad (10)$$

while the quality of a downstream phase front is given by Δ_{down}^{Phase}:

$$\Delta_{down}^{Phase} = \frac{1}{N} \sum_{t=t_s}^{t_e} |x_{down_M}^{Phase}(t) - x_{down_{GT}}^{Phase}(t)| \quad (11)$$

Both quality indices give as a result the relative average deviation in meters between a reconstructed phase front and a phase front, which has been defined as *Ground-Truth*. Hence $\Delta_{up}^{Phase} \geq 0\ m$ and $\Delta_{down}^{Phase} \geq 0\ m$. The best quality is achieved, when both phase fronts are the same. This is represented by $\Delta_{up}^{Phase} = 0\ m$ and $\Delta_{down}^{Phase} = 0\ m$.

Within a traffic pattern, which has been defined as *Ground-Truth*, a phase front occurs on $N(GT)$ time instant between the temporal borders t_s and t_e. In this period a single phase front shows no gaps. If it would be the case it would not be a single phase front, but instead *several* phase

fronts. Hence, in addition to the reconstruction quality of a phase front, its coverage is assessed. Coverage is defined as the ratio of $N(M)$ time instants where the phase front was reconstructed by a model with respect to $N(GT)$ time instants where the phase front was present in *Ground-Truth*. This leads to two addition quality indices, which measure the coverage of both the upstream and downstream phase fronts.

The coverage of the upstream phase front is given by C_{up}^{Phase}

$$C_{up}^{Phase} = \frac{N_{up}^{Phase}(M)}{N_{up}^{Phase}(GT)} \tag{12}$$

while C_{down}^{Phase} gives the coverage of the downstream phase front:

$$C_{down}^{Phase} = \frac{N_{down}^{Phase}(M)}{N_{down}^{Phase}(GT)} \tag{13}$$

Both C_{up}^{Phase} and C_{down}^{Phase} give a result between 0 and 1, describing the coverage of a *Ground-Truth* phase front by a reconstructed one. Best coverage is achieved by $C_{up}^{Phase} = 1$ and $C_{down}^{Phase} = 1$.

All quality indices for phase fronts emphasize some specific aspects of a spatio-temporal congested traffic pattern. They assess the quality of the reconstruction of a *specific* phase fronts, hence they do not represent a quality index for *all* phase fronts of a traffic pattern.

D. Quality Index for Regions of S and J

The quality index for regions of *S* and *J* assesses the general reconstruction quality between a traffic pattern TP_M reconstructed by a model M and traffic pattern TP_{GT}, which has been defined as the *Ground-Truth* reference information. For each congested traffic phase the spatio-temporal areas they cover are compared against each other with regard to hits and false alarms. Figure 7 illustrates this concept, representing an adaption of the ROC[1]-analysis for traffic data. The origin of ROC-analysis is medical diagnostics [23]. Since a few years its concepts have also been applied to machine learning [24].

The ROC-analysis is previously been applied for assessing the quality of traffic data [1]. However, in this case only the classes *congested* and *not congested* have been considered [12]. In [16], the ROC-analysis was extended and applied to traffic data containing three distinct traffic phases.

The definitions of TP_M and TP_{R_d} given in the previous section can directly be applied to yield a computationally efficient calculation of this quality index. The *Ground-Truth* reference information use for quality assessment is defined as TP_{GT}. For both TP_M and TP_{GT} three binary matrices are calculated, one for each traffic phase. These matrices

[1]ROC - Receiver operating characteristic

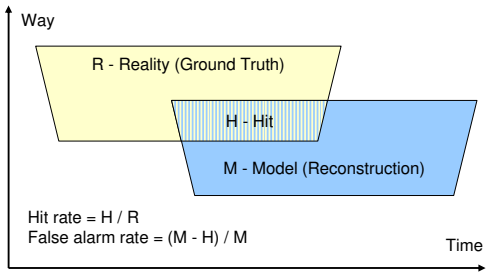


Figure 7. Definition of the Quality Index for Regions of Synchronized Flow and Wide Moving Jam

contain the number 1 at all spatio-temporal positions where a specific traffic phase occurs and the number 0 at all other positions.

For TP_M this leads to TP_M^F, TP_M^S and TP_M^J and for TP_{GT} to TP_{GT}^F, TP_{GT}^S and TP_{GT}^J, respectively. Using a row vector r_v and a column vector c_v with correct dimension for TP_M and TP_{GT} with all $a_{ij} = 1$ allows to count the number c of elements a_{ij} within matrix A where $a_{ij} = 1$.

$$c = r_v * A * c_v \tag{14}$$

The reconstruction quality of spatio-temporal traffic patterns can now be assessed based on ROC-analysis [3] using the ROC measure TPR (true-positive-rate),

$$TPR = \frac{t_p}{t_p + f_n} \tag{15}$$

the measure FPR (false-positive-rate)

$$FPR = \frac{f_p}{t_n + f_p} \tag{16}$$

and finally the measure FAR (false alarm rate)

$$FAR = \frac{f_p}{t_p + f_p} \tag{17}$$

whereas t_p represents the true-positives, f_p the false-positives, t_n the true-negatives and f_n the false-negatives. Figure 8 shows this for a feature space having the features A and B, where the classification of a class G is performed by a classificator C.

This leads to different definitions of t_p, f_p, t_n, f_n and TPR, FPR and FAR for both congested traffic phases *S* and *J*.

1) Traffic phase S*:* For synchronized flow *S* the classification of a traffic state TS at a position P_d results in the following different cases according to ROC definitions:

1) True positive t_p: $TS_M = S \wedge TS_{GT} = S$
2) False positive f_p: $TS_M = S \wedge (TS_{GT} = J \vee TS_{GT} = F)$
3) True negative f_n: $(TS_M = J \vee TS_M = F) \wedge (TS_{GT} = J \vee TS_{GT} = F)$

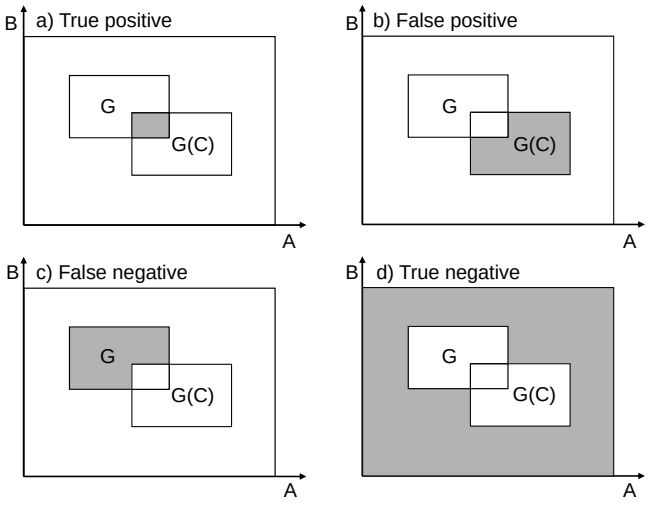


Figure 8. Basic ROC definitions

4) False negative t_n: $(TS_M = J \vee TS_M = F) \wedge TS_{GT} = S$

Taking these definitions into account, the corresponding values for TPR_S, FPR_S and FAR_S are calculated as follows:

$$TPR_S = \frac{z_v * (TP_M^S \wedge TP_{GT}^S) * s_v}{z_v * TP_{GT}^S * s_v} \tag{18}$$

$$FPR_S = \frac{z_v * (TP_M^S \wedge TP_{GT}^J + TP_M^S \wedge TP_{GT}^F) * s_v}{z_v * (TP_{GT}^J + TP_{GT}^F) * s_v} \tag{19}$$

$$FAR_S = \frac{z_v * (TP_M^S \wedge TP_{GT}^J + TP_M^S \wedge TP_{GT}^F) * s_v}{z_v * TP_M^S * s_v} \tag{20}$$

All quality indices TPR_S, FPR_S and FAR_S give values between 0 and 1.

The index TPR_S describes the ratio of TP_{GT}, which is covered by TP_M. A value of $TPR_S = 1$ corresponds to the best possible quality. In this case TP_{GT} is completely covered by TP_M. The worst possible quality is represented by $TPR_S = 1$, when the reconstructed traffic pattern does not cover any region of the real traffic pattern.

In contrast to TPR_S, the index FPR_S describes the ratio of TP_{GT}, which is by mistake classified as traffic phase *S*. Hence the best possible quality corresponds to $FPR_S = 0$, while the worst possible quality corresponds to $FPR_S = 1$.

Finally, the false alarm rate FAR_S describes the ratio of TP_M, which by mistake classifies the traffic phase *S*. The best possible quality corresponds to $FAR_S = 0$, while the worst possible quality corresponds to $FAR_S = 1$.

The special case $TPR_S = 1$, $FPR_S = 0$ and $FAR_S = 0$ represents the best possible quality, when all three indices are taken into account. In this case TP_M and TP_{GT} are completely congruent, the traffic pattern reconstructed by the model is in accordance to real traffic pattern.

2) Traffic phase J*:* For the traffic phase wide moving jam *J* the classification of a traffic state TS at a position P_d results in the following possible cases:

1) True positive r_p: $Z_M = J \wedge Z_{GT} = J$
2) False positive f_p: $Z_M = J \wedge (Z_{GT} = S \vee Z_{GT} = F)$
3) True negative f_n: $(Z_M = S \vee Z_M = F) \wedge (Z_{GT} = S \vee Z_{GT} = F)$
4) False negative r_n: $(Z_M = S \vee Z_M = F) \wedge Z_{GT} = J$

The corresponding values for TPR_J, FPR_J and FAR_J are calculated as follows:

$$TPR_J = \frac{z_v * (TP_M^J \wedge TP_{GT}^J) * s_v}{z_v * TP_{GT}^J * s_v} \tag{21}$$

$$FPR_J = \frac{z_v * (TP_M^J \wedge TP_{GT}^S + TP_M^J \wedge TP_{GT}^F) * s_v}{z_v * (TP_{GT}^S + TP_{GT}^F) * s_v} \tag{22}$$

$$FAR_J = \frac{z_v * (TP_M^J \wedge TP_{GT}^S + TP_M^J \wedge TP_{GT}^F) * s_v}{z_v * TP_M^J * s_v} \tag{23}$$

Again TPR_J, FPR_J and FAR_J give all values between 0 and 1. Regarding maximum and minimum quality they behave like their counterparts for the traffic phase *S*.

E. Summary

Table III gives an overview of all defined quality indices and the values, which represent the maximum and minimum quality.

Table III
OVERVIEW OF ALL DEFINED QUALITY INDICES

Index	Phase	Maximum Quality	Minimum Quality
ΔT	F, S, J	0	∞
Δ_{up}^S	S	0	∞
Δ_{down}^S	S	0	∞
C_{up}^S	S	1	0
C_{down}^S	S	1	0
Δ_{up}^J	J	0	∞
Δ_{down}^J	J	0	∞
C_{up}^J	J	1	0
C_{down}^J	J	1	0
TPR_S	S	1	0
FPR_S	S	0	1
FAR_S	S	0	1
TPR_J	J	1	0
FPR_J	J	0	1
FAR_J	J	0	1

Indices, which relate to a deviation Δ, show the maximum quality, when the deviation is small $\Delta = 0$. Hence the minimum quality is represented by a large deviation $\Delta >> 0$.

The maximum quality of the coverage C is full coverage $C = 1$, while the minimum quality is represented by no coverage $C = 0$.

The indices TPR, FPR and FAR have their maximum quality at a maximum of correct classifications and a minimum of false classifications. Hence the maximum quality is represented by $TPR = 1$, $FPR = 0$ and $FAR = 0$, while the minimum quality is represented by $TPR = 0$, $FPR = 1$ and $FAR = 1$.

VI. Test Environment

The Kerner-Klenov microscopic three-phase traffic model ([5], [8]) 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 ([5], [8]).

Ground-Truth denotes in this context that the model output represents the reference information or the *reality*, which should be reconstructed by a traffic model using the vehicle trajectories as base data. 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 9).

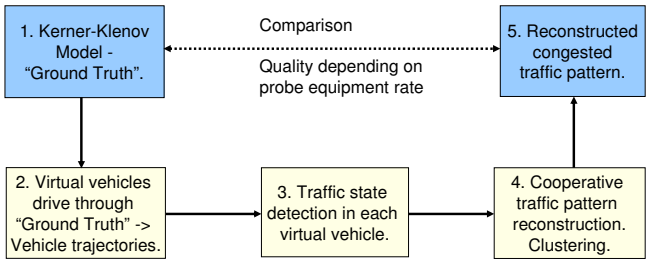


Figure 9. Steps necessary for probe equipment rate investigations

VII. Results

A. Examined Situation

For evaluation the Kerner-Klenov three-phase microscopic model was used to simulate an expanded spatio-temporal congested traffic pattern. The simulation was configured to simulate a 30 km highway stretch with three lanes and three junctions containing on- and off-ramps. The incoming and outgoing flow at these three junctions leads to the realization of three bottlenecks, which could lead to an $F \rightarrow S$ traffic breakdown and the formation of several regions of synchronized flow. Within the synchronized flow regions several wide moving jams emerged (Figure 10).

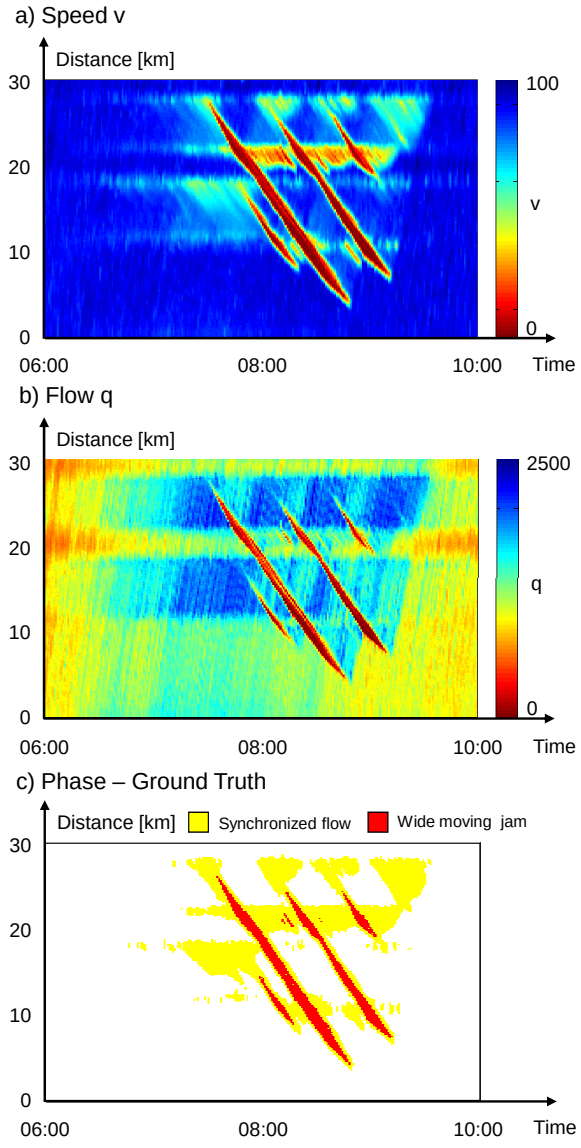


Figure 10. Kerner-Klenov simulation output of an expanded pattern. Average vehicle speed is shown in Figure a), average vehicle flow is shown in Figure b), while Figure c) shows the traffic phase of *Ground-Truth* at a given spatio-temporal position. Wide moving jams can be identified in all three figures as stable structures moving upstream with a constant speed. In contrast the downstream front of the traffic phase synchronized flow is stationary fixed at the location of the bottlenecks.

In the following sections, the quality index results of model using probe vehicle data are compared with the results of the ASDA and FOTO models based on stationary loop

detector data. In order to achieve this, the simulated situation was reconstructed by both the model using probe vehicle data and the ASDA and FOTO models using different configurations of input data. Then the achievable quality was assessed by using the quality indices introduced earlier in this paper.

The models ASDA and FOTO were used with different detector configurations. The average distance ranged between about 1 km until up to up 9 km. Results of the models ASDA and FOTO with two different detector configurations are shown in Figure 11.

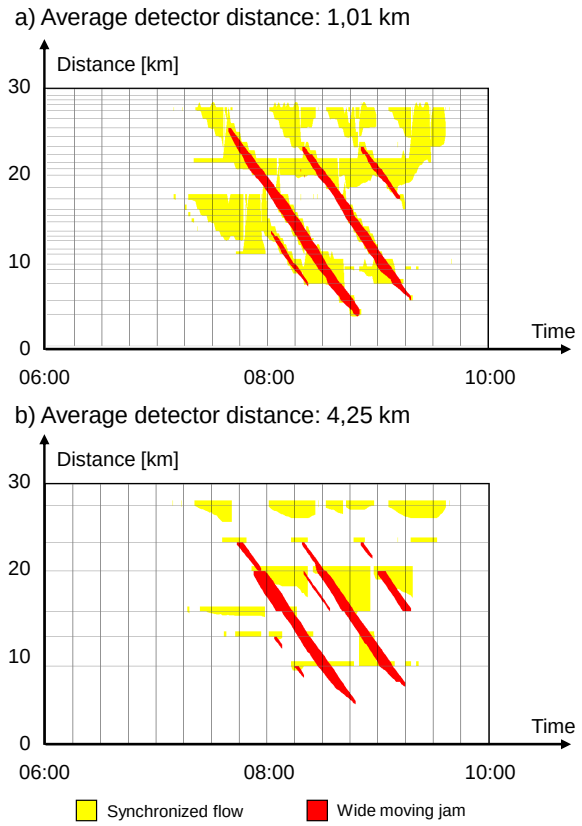


Figure 11. Example results for the models ASDA and FOTO

The traffic models based on probe vehicle data were used with different equipment rates of probe vehicles ranging between 4 % and 0.25 %. Results of the vehicles based models are shown in Figure 12. Each assessed equipment rate was examined based on 100 random distributions of this equipment rate. Each random distribution of a given equipment rate was selected from all vehicles, which are part of the traffic flow. Each selected random distribution represents one possible realization of this equipment rate. By comparing the results of all random distributions of one equipment rate, the influence of different realizations on the achievable results at a given equipment rate can be derived.

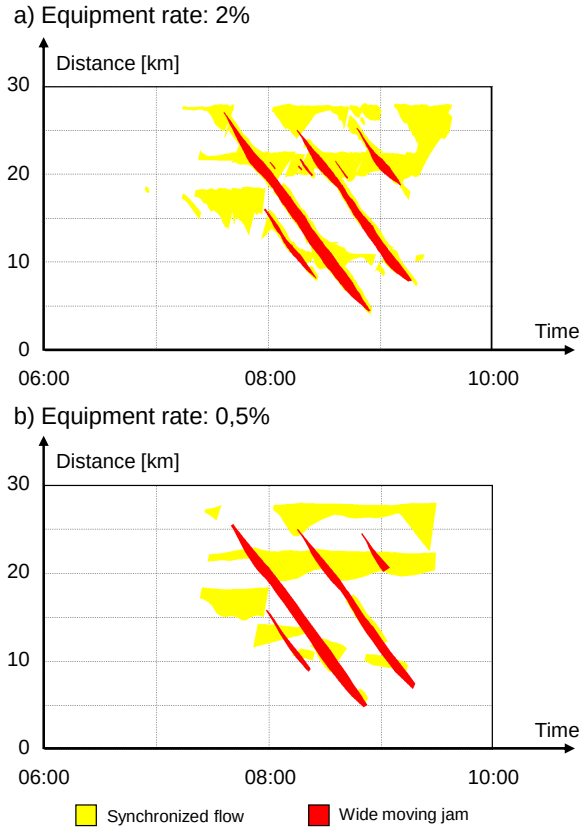


Figure 12. Example results for vehicle based methods

B. *Quality Index for Travel Time*

The results of both models were assessed using the quality index for travel time. Figure 13 shows the results for the ASDA and FOTO models, while Figure 14 shows the results of the model based on probe vehicle data.

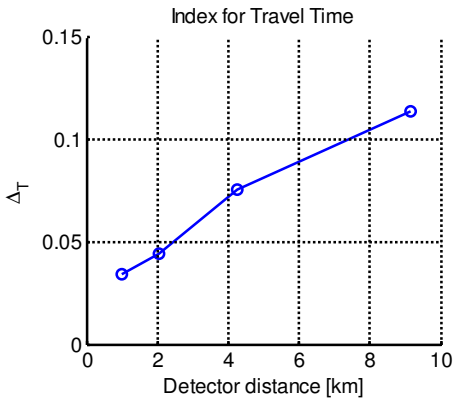


Figure 13. Quality Index for Travel Time: ASDA and FOTO

Using the quality index for travel time the dependence of the model result from the available input data is visible. The probe vehicle based methods generally offer a better quality

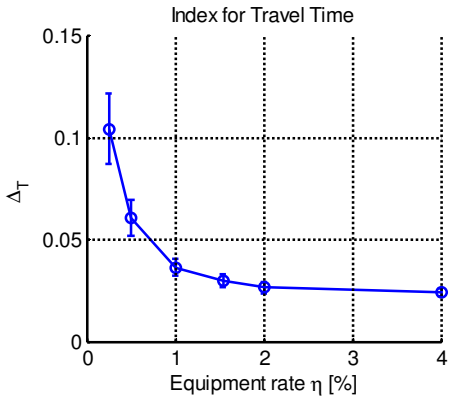


Figure 14. Quality Index for Travel Time: Vehicles

compared to the ADSD/FOTO models. At a equipment rate of about 1 % the achievable results are comparable to a high quality detector network having an average detector distance of 1-2 km.

C. Quality Index for Fronts of Synchronized Flow and Wide Moving Jam

The results of both models were also compared using the quality index for fronts of synchronized flow and wide moving jam. At first the deviation of fronts depending on the input data were examined. Figure 15 shows the results for the models ASDA and FOTO, while Figure 16 shows the results for the models based on probe vehicle data. For both examined jams the models based on probe vehicle data show superior results. At the best detector configuration having an average detector distance of only 1 km the ASDA and FOTO models achieve a front deviation of about 200 m for both jams. This increases up to 300 - 800 m when the average detector distance increases up to over 9 km. The models based on probe vehicle data achieve a front deviation of at most 200 m down to a equipment rate of only 0.5 %. At higher equipment rates the front deviation is even smaller. This corresponds to an even better quality.

After that the front coverage was examined for both models. Figure 17 shows the results for the ASDA and FOTO models. In Figure 18, the results for the models based on probe vehicle data are shown. The quality index for front coverage shows comparable results: the models based on probe vehicle data are superior to the models ASDA and FOTO. Even at an equipment rate of only 1 % the results of the vehicle based models are comparable to ASDA and FOTO used with an average detector distance of only 1 km.

D. Quality Index for Regions of Synchronized Flow and Wide Moving Jam

Finally, the results of both models are assessed using the quality index for regions of synchronized flow and wide moving jam. Figure 19 shows the results for the ASDA and FOTO models, while Figure 20 shows the results for

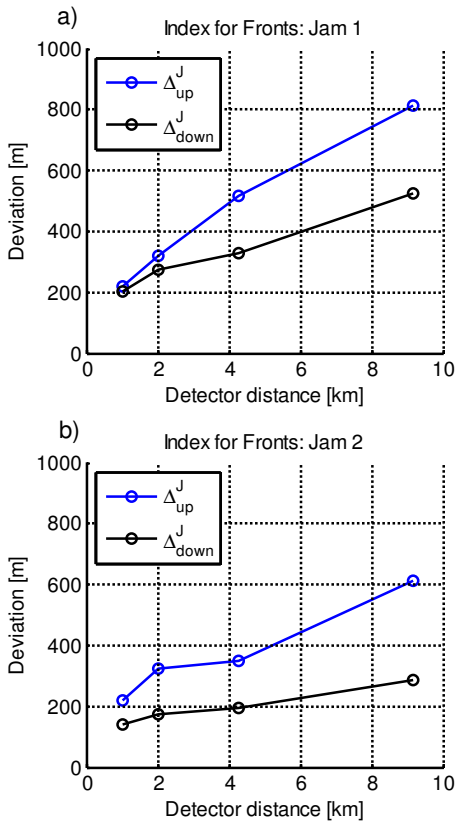


Figure 15. Quality Index for Fronts (Deviation): ASDA and FOTO

the models based on probe vehicle data. Again the models based on probe vehicle data are superior to the ASDA and FOTO models. At an equipment rate of about 1-1.5 % the results are comparable to a dense detector network with an average detector distance of 1 km. In addition the false alarm rates are very low for all examined equipment rates. In comparison the models ASDA and FOTO show an increasing false alarm rate as the detector distance increases.

In addition the results of both models can be compared directly, when both of them are plotted as a ROC curve in the ROC space. In this case TPR is plotted over FPR. A random decision is represented by the line $TPR = FPR$. All results above this line are better than the random decision, all results below this line are worse than the random decision. The maximum quality in the ROC space is represented by the point $TPR = 1$ and $FPR = 0$. Figure 21 shows the results of both models in the ROC space. These results underline the previously discussed results. The probe vehicle based methods offer higher hit rates combined with lower false alarm rate. This leads to an overall better quality. The reconstruction quality is much better for the traffic phase J when compared to the traffic phase S. The reason for this is the stable upstream propagation of the traffic phase J, which allows a high quality prediction of the future movement of the traffic phase and thus a high quality reconstruction of

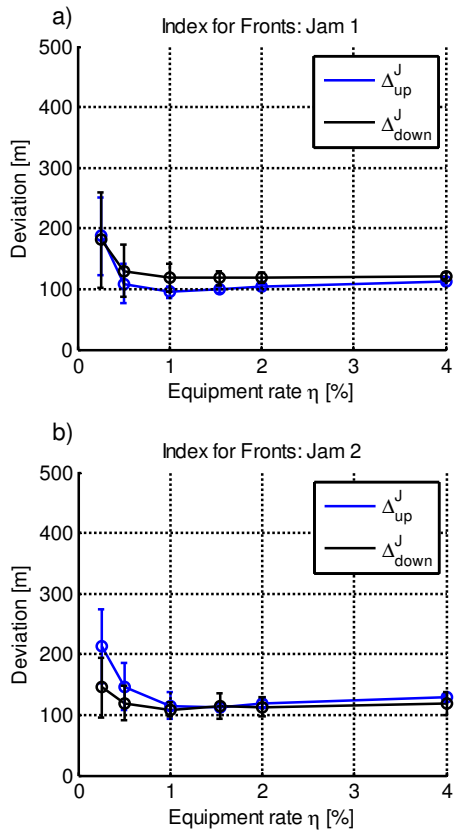


Figure 16. Quality Index for Fronts (Deviation): Vehicles

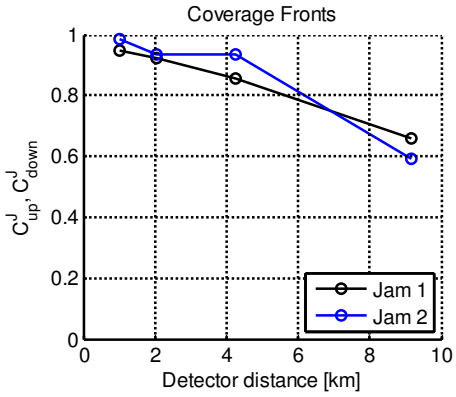


Figure 17. Quality Index for Fronts (Coverage): ASDA and FOTO

the overall region by a model.

E. Application for Empirical Data of Moving Probe Vehicles

The results shown above have been achieved using simulated traffic data. A cooperation with TomTom gave the authors the opportunity to check the methods and results with real-world data from a chosen traffic situation [21]. This real traffic situation is a congested situation on the freeway A5 in Germany on the 10th December, 2009. The situation is similar to the simulated example from Figure 10.

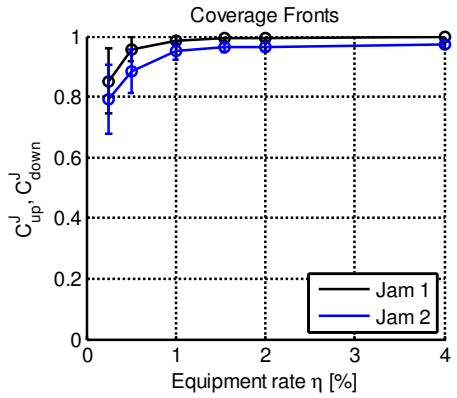


Figure 18. Quality Index for Fronts (Coverage): Vehicles

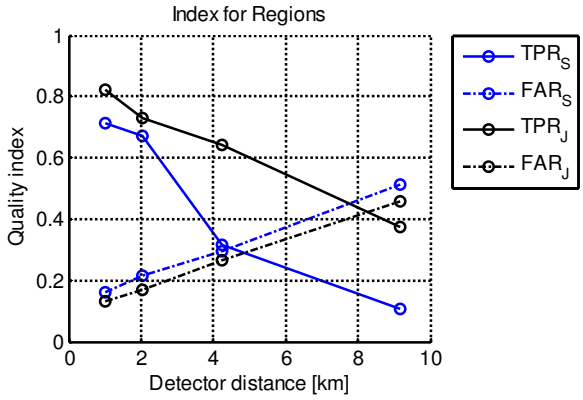


Figure 19. Quality Index for Regions: ASDA and FOTO

The number of probe vehicles collecting data for TomTom reach an equipment rate of approximately $\eta \approx 1\%$ for this traffic situation. Figure 22 a) illustrates the results after processing the TomTom probe vehicle data with the traffic state detection algorithm, while Figure 22 b) shows the results of the phase front detection. In Figure 22 c) consequently the overall spatial-temporal reconstructed congested pattern

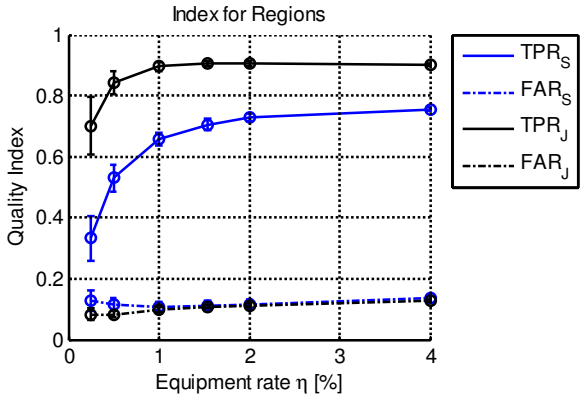


Figure 20. Quality Index for Regions: Vehicles

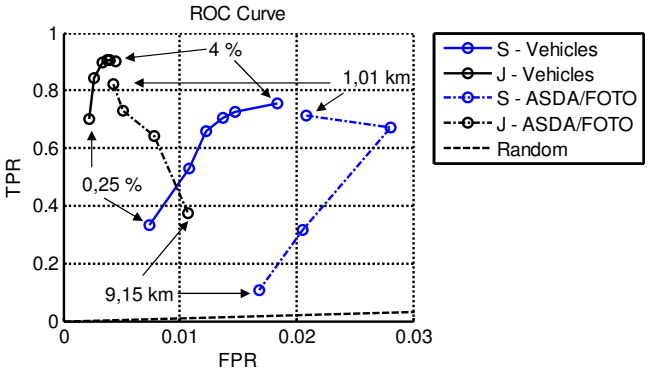


Figure 21. Quality Index for Regions: ROC Curve

based on real probe vehicle data is presented.

The following Figure 23 presents an empirical comparison for the same traffic situation based on different measurement techniques and data processing methods. At the top the results of the ASDA and FOTO models for the measured detector data and at the bottom the congested pattern based on TomTom probe vehicle data and reconstructed with vehicle based methods. The results prove that based on an equipment rate of approximately $\eta \approx 1\%$ the possible data reconstruction quality is in the order of the ASDA and FOTO models for average detector distances of ≈ 1 km.

Qualitatively a good correlation of both congested traffic phase regions *S* and *J* can be identified. Between the positions 0 km and 5 km the pattern reconstructed from probe vehicle data (Figure 23 b)) shows smaller regions of the traffic phase *S*, which do not exist in the ASDA and FOTO reconstructed pattern (Figure 23 a)). At the location between 0-5 km the detectors are more sparse than at other sections of the freeway and the traffic phase *S*, therefore, can be detected and reconstructed less precise in the ASDA and FOTO models. Those congested regions can been identified only with probe vehicle data based reconstruction methods. Overall, in this real-world case study the processing of *different* raw traffic data sources with *different* reconstruction methods has led to *comparable* results.

VIII. Conclusion and Future Work

Probe vehicle equipment rates of about 1-1.5 % processed with the proposed method are comparable to detectors with distances of 1-2 km processed with the existing ASDA and FOTO models. 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. It should be noted, that this 1-1.5 % of vehicle data has been processed optimally, i.e., without data failures and without additional latencies of communication channels. A real-world case of traffic

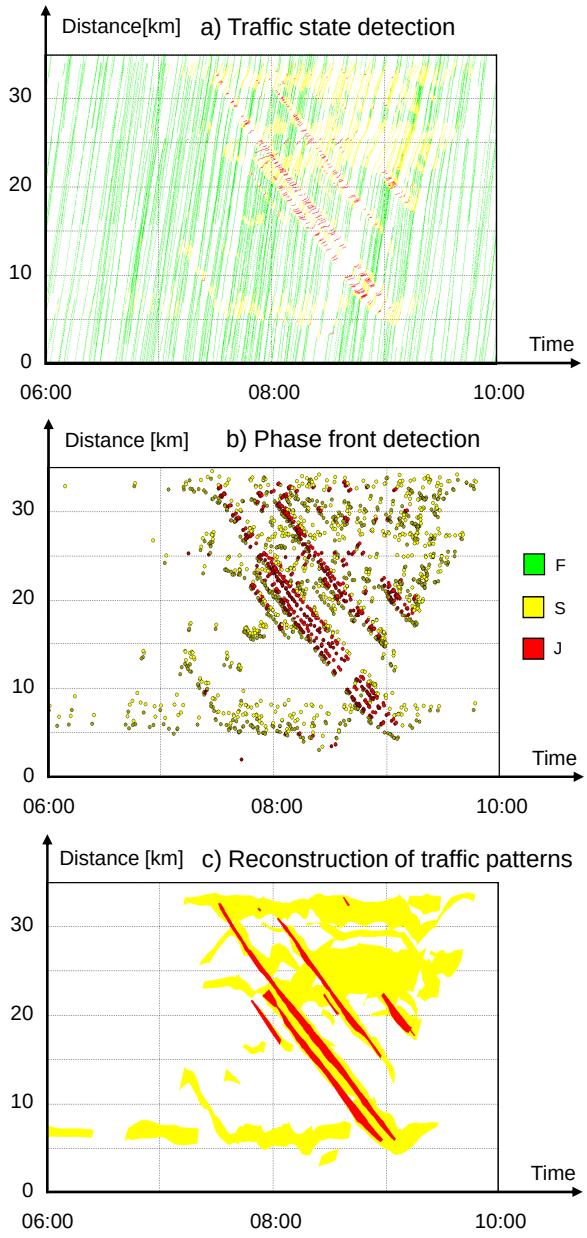


Figure 22. Reconstruction based on TomTom's probe vehicle data on 10th December, 2009 on the A5-South in Hessen, Germany [13]

congestion has proven that the processing of *different* raw traffic data (detectors and probe vehicles) have led with *different* methods (ASDA and FOTO models and probe data processing) to *comparable* results.

Subjects of further ongoing research are the evaluation and prototyping of new vehicular assistance applications based on high quality information about traffic patterns. Different types of vehicular assistance applications benefit from different parts of information about congested traffic patterns. For example having good results in the quality

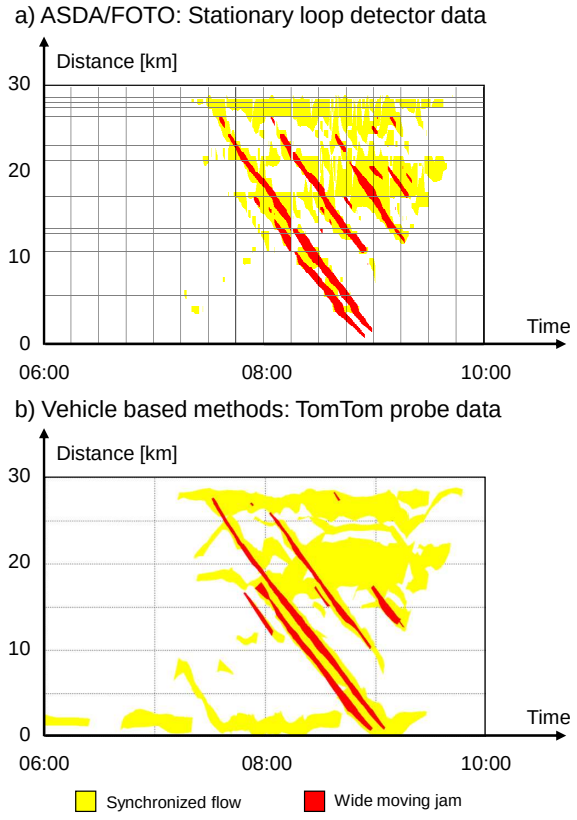


Figure 23. Traffic congestion on 10th December, 2009 at A5-South - Comparison of different reconstruction results for different empirical measurements [13]

index of travel time is relevant for routing applications, while applications controlling an hybrid engine would benefit from good results in the quality index for fronts of synchronized flow and wide moving jam. Future research has to answer the important question about the performance of specific vehicular assistance applications on the reconstruction quality of congested traffic patterns. This might include the definition of additional quality indices, which are tailored to the requirements of specific vehicular assistance applications. For example the reconstruction quality of traffic state changes from F to J could be more important to some applications than traffic state changes from F to S. An additional quality could take these differences into account.

Another relevant field for future research is the evaluation of system communication and system architecture alternatives. Different system architectures, like for example central server based systems and fully distributed systems using only communicating vehicles, have different advantages and disadvantages, when it comes to the overall system performance in comparison to the investment needed to establish and maintain these system architectures.

ACKNOWLEDGMENT

The authors would like to thank Boris S. Kerner and Sergey L. Klenov for support regarding the microscopic three-phase traffic model, the simulated vehicle data, and many fruitful discussions. In addition, we would like to thank Ralf-Peter Schaefer and Nikolaus Witte from TomTom for their support with evaluating probe vehicle data.

REFERENCES

[1] K. Bogenberger, *Qualitaet von Verkehrsinformationen* Strassenverkehrstechnik, Kirschbaum Verlag GmbH Bonn, pp. 10:518–526, 2003.

[2] D. Ehmanns, H. Wallentowitz, C. Gelau, and F. Nicklisch, *Zukuenftige Entwicklungen von Fahrerassistenzsystemen und Methoden zu deren Bewertung* In Proceedings: 9. Aachener Kolloquium Fahrzeug- und Motorentechnik, 2000.

[3] T. Fawcett, *ROC graphs: Notes and practical considerations for researchers* Machine Learning, pp. 1–38, 2004.

[4] J.C. Herrera, D.B. Work, R. Herring, X.J. Ban, Q. Jacobson, and A.M. Bayen, *Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment* TTransportation Research Part C: Emerging Technologies, pp. 2:135–166, 2008.

[5] B.S. Kerner, *The Physics of Traffic* Berlin, New York: Springer, 2004.

[6] B.S. Kerner, S. L. Klenov, J. Palmer, M. Prinn, and H. Rehborn, *Verfahren zur Verkehrszustandsbestimmung in einem Fahrzeug* German Patent Publication DE 10 2008 003 039, 2008.

[7] B.S. Kerner, *Betriebsverfahren fuer ein in einem Fahrzeug befindliches verkehrsadaptives Assistenzsystem* German Patent Publication DE 10 2005 017 560, 2005.

[8] B.S. Kerner, *Introduction to Modern Traffic Flow Theory and Control* Berlin, New York: Springer, 2009.

[9] B.S. Kerner, H. Rehborn, M. Aleksic, A. Haug, and R. Lange, *Online automatic tracing and forecasting of traffic patterns* Traffic Engineering & Control, Hemming, pp. 10:345–350, 2001.

[10] B.S. Kerner, H. Rehborn, J. Palmer, and S.L. Klenov, *Using probe vehicle data to generate jam warning messages* Traffic Engineering & Control, Hemming, pp. 3:141–148, 2011.

[11] A. Kesting, *Microscopic Modeling of Human and Automated Driving: Towards Traffic-Adaptive Cruise Control* PhD Thesis, Technical University of Dresden, 2008.

[12] S. Lorkowski, *Fusion von Verkehrsdaten mit Mikromodellen am Beispiel von Autobahnen* PhD Thesis, Technical University of Berlin, 2009.

[13] J. Palmer, *Fahrzeugautonome und verteilte Erkennung rauemlich-zeitlicher Verkehrsmuster zur Nutzung in Fahrerassistenzsystemen* PhD Thesis, University of Tuebingen, 2011.

[14] 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* In Proceedings: 16th World Congress on ITS, Stockholm, 2009.

[15] J. Palmer and H. Rehborn, *Vehicular Assistance Applications in the Scope of Kerner's Three-Phase Traffic Theory* In Proceedings: Networks for Mobility - 5th International Symposium, Stuttgart, 2010.

[16] J. Palmer and H. Rehborn, *Reconstruction Quality of Congested Freeway Traffic Patterns from Probe Vehicles Based on Kerner's Three-Phase Traffic Theory* In Proceedings: 6th International Conference on Systems - ICONS, St. Maarten, The Netherlands Antilles, 2011.

[17] J. Palmer, H. Rehborn, and B.S. Kerner, *ASDA and FOTO Models based on Probe Vehicle Data* Traffic Engineering & Control, Hemming, pp. 4:183–191, 2011.

[18] H. Rehborn and J. Palmer, *ASDA/FOTO based on Kerner's Three-Phase Traffic Theory in North Rhine-Westphalia and its Integration into Vehicles* In Proceedings: 2008 IEEE Intelligent Vehicles Symposium, Eindhoven, 2008.

[19] H. Rehborn, S. L. Klenov, and J. Palmer, *An empirical study of common traffic congestion features based on traffic data measured in the USA, the UK, and Germany* Physica A 390, pp. 4466-4485, 2011.

[20] H. Rehborn and S. L. Klenov, *Congested Traffic Prediction of Congested patterns* In: Meyers, Robert (Ed.) Encyclopedia of Complexity and Systems Science, Springer New York, pp. 9500-9536, 2009.

[21] R.-P. Schaefer, S. Lorkowski, N. Witte, J. Palmer, H. Rehborn, and B.S. Kerner, *A study of TomToms probe vehicle data with three phase traffic theory* Traffic Engineering & Control, Hemming, pp. 5:225–230, 2011.

[22] M. Schoenhof and D. Helbing, *Empirical Features of Congested Traffic States and Their Implications for Traffic Modeling* Transportation Science, pp. 4:568–583, 2010.

[23] J.A. Sweets, *Measuring the Accuracy of Diagnostic Systems,* Science, 240(4857), pp. 1285–1293, 1988.

[24] K. Veropoulos, C. Campbell, and N. Cristianini, *Controlling the Sensitivity of Support Vector Machines* In Proceedings: IJCAI - International Joint Conference on Artificial Intelligence, Workshop on Support Vector Machines, Stockholm, Schweden, 1999.

Testing Platform for Hardware-in-the-Loop and In-Vehicle Testing Based on a Common Off-The-Shelf Non-Real-Time 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. The timing behavior of the Driver Assistance Functions is usually tested on real-time capable Hardware-in-the-Loop platforms as well as by in-vehicle tests, where the timing behavior is evaluated with the help of data loggers. Both, the data loggers and the Hardware-in-the-Loop platforms are mostly custom-made, proprietary and in consequence expensive. Moreover, many software developers usually have to share few instances. Existing inexpensive solutions show deficits in their real-time capabilities, which means for Hardware-in-the-Loop platforms that the real-time behavior cannot be guaranteed and for data loggers that they do not provide a common time base for relating data from different vehicles involved in a maneuver. 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 and an extended I/O interface, which can be used for in-vehicle tests as well. Thereby, the simulation software runs on the developer's desktop computer while the extended I/O interface provides a global time base and ensures the real-time communication with the System Under Test even for complex timing requirements. Two examples show how the introduced setup can be used to test Driver Assistance Functions on a Hardware-in-the-Loop platform and as a data logger for in-vehicle tests. Questions such as "How much time is needed by the Adaptive Cruise Control System to determine the relative speed of the preceding vehicle?" can be answered.**

Keywords-Hardware-in-the-Loop Testing; In-Vehicle Testing; Embedded Real-Time Systems; Temporal Behavior; Driver Assistance Systems.

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 testing of the timing behavior on the target hardware as presented in [1]. 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 these platforms very expensive. Often, these testing platforms are based on standard PC hardware in combination with a 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' desks, which inhibits the rapid prototyping development cycle.

The approach introduced in this paper uses a special, 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 developer's 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 interfaces. Even ECUs communicating via analog 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 suitable 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) [2] 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 more efficient way–enabling a higher test frequency.

Testing Driver Assistance Functions in vehicles adds another challenge to the test equipment. Having now the ECU in the vehicle means that a surrounding vehicle on which the Driver Assistance Function reacts has to be present. Furthermore, the surrounding vehicle has to be considered for evaluating the test result. The Driver Assistance Function Adaptive Cruise Control (ACC) [3], e.g., is a closed-loop control for the vehicle's speed considering the speed of the preceding vehicle. Common ACC systems have a radar sensor, which determines the relative speed and distance of the preceding vehicle. In some ACC systems, the driver can set the desired distance to the preceding vehicle and the ACC system tries to keep this distance by accelerating or decelerating the vehicle. If such a system is to be tested on the road it is necessary to be able to correlate the information about the preceding vehicle with the information of the vehicle with the ACC system. Issues to be tested might be:

- How much time elapsed between pressing the brake pedal in the preceding vehicle and a deceleration demand of the ACC system?
- How much time is needed by the radar device to determine the relative speed of the two vehicles?
- What is the difference between ACC algorithms on the behavior of the vehicle?
- How does an ACC algorithm perform compared to a human driver?

All questions have in common that having data from the vehicle with the ACC system is not enough. It is necessary to know the state of the preceding vehicle and to be able to correlate this information. Independent data loggers in the vehicles do not allow to record data on a common time base, which means that data cannot be correlated in time. It is therefore necessary to synchronize the data loggers in the different vehicles. The introduced RTA is able to synchronize its internal clock to the time provided by the Global Positioning System (GPS). Since the RTA is not only able to record data with a highly precise GPS-based time stamp but is also able to use the GPS-time to send data, it is possible for applications to replay data recorded in a vehicle with precisely the same timing behavior on a test platform in the laboratory.

Section II of this paper gives an insight into the testing of interconnected ECUs followed by a section that compares current HiL testing platforms relating to timing issues. In Section IV, the operating principles of the RTA as intelligent I/O device are introduced as well as an approach for a HiL testing platform based on the RTA. Finally, Section V shows examples for a HiL test setup and an in-vehicle test of an ACC system, which show the current usage of the RTA 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 [4].

The distribution of software on 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 ACC–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 [5], 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 seen cycle time of this message on the bus will differ from the desired 20 ms. The ECUs will tolerate such an inaccuracy as long as the deviation is below a specified limit.

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. Usually, the ECU for Driver Assistance Functions monitors if the timing of the received messages is within the predefined limits. In some cases, the data content is additionally monitored on its in-time arrival as explained in Section V-B. Considered from a testing perspective, it is thus

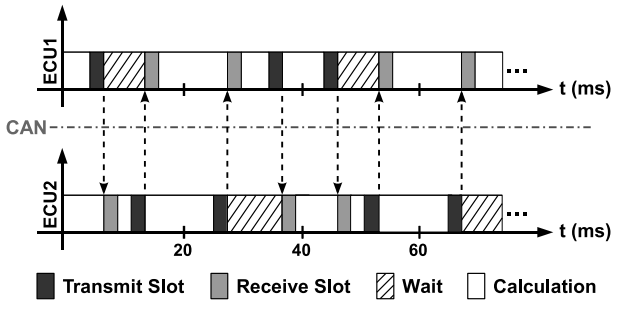


Figure 1. Cyclic communication of ECUs

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 ECU software on the target hardware. A common approach is to do this integration testing on HiL testing platforms.

Further on, it is useful having the HiL testing platform close to the software-developers' desks, 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 the ECU software, 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.

After having tested the combination of ECU software and hardware on a HiL platform, the ECU can be integrated into a vehicle to see the Driver Assistance Function work in its target environment. For Driver Assistance Functions such as an ACC, the target environment is not limited to the vehicle but the vehicle and its surrounding vehicles that the ACC reacts on have to be considered [6], too. The same also applies for the test of a Forward Collision Warning System (FCWS). The National Highway Traffic Safety Administration (NHTSA) requests for testing a FCWS in [7] that the instance in time when the driver has been warned can be correlated to the position of both vehicles involved in the test. This means that the measured data of both vehicles have to be correlated.

III. Current Testing Platforms

In the following, three different HiL approaches currently used in the automotive industry are discussed with a closer look on their timing behavior.

A. HiL Platform Based on an RTOS

Current HiL platforms, as they are introduced in [8], usually focus on ECU testing from both, a functional and a 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 [9], 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. To guarantee the real-time behavior of the CAN communication, an RTOS is used to implement the RBS for the CAN bus and the additional 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.

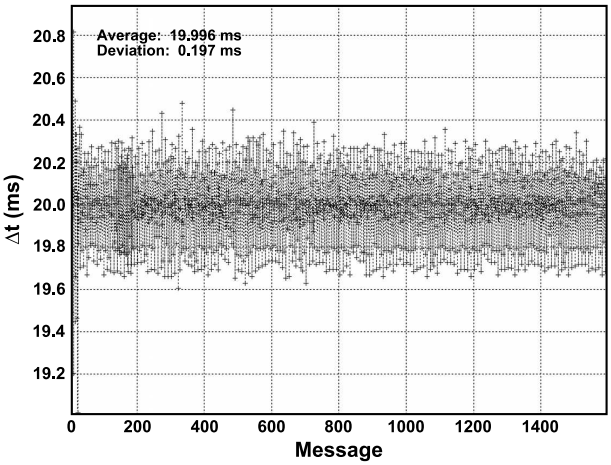


Figure 2. RTOS HiL – 20 ms cycle time message

Figure 2 shows the measured time between two CAN messages with the same message ID during a HiL test at a platform based on an RTOS [10][11][12]. According to the CAN schedule, the message is supposed to have a 20 ms cycle time. The plot displays that this implementation achieves an average cycle time of almost 20 ms with a standard deviation of about 197 µs. Single outliers are reaching up to a period of 20.826 ms between two consecutive messages. In this example, the measured timing still fulfills 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. In this case, the COTS computer can be 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

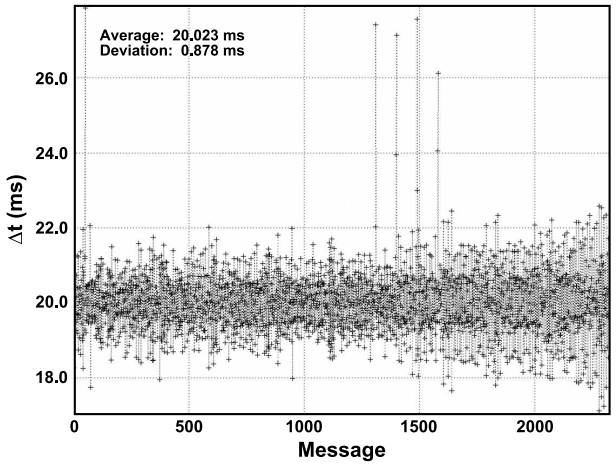


Figure 3. Non-real-time OS HiL – 20 ms cycle time message

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 3 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 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 development tool chain is necessary to implement the RBS on the CAN interface. This leads to a fixed 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. Data Logger for In-Vehicle Testing

Testing a Driver Assistance Function in the vehicle is usually done with the help of a data logger. The data logger collects and stores the data, which is transferred between ECUs within the vehicle. The test result is based on the collected data. Common data loggers like [13] are able to record several different buses within one vehicle on a common time base. A time synchronization for data loggers of several vehicles is mostly custom made and thus expensive. For evaluating Driver Assistance Functions such as an ACC, this feature is important. Since the interaction between vehicles raise questions such as ”How much time is needed by the Adaptive Cruise Control System to determine the relative speed of the preceding vehicle?”.

There are several ways like RFC 958 [14] or IEEE 1588 [15] to synchronize independent clocks. As mentioned by Luther [16] IEEE 1588, also referred to as Precision Time Protocol, reaches a temporal synchronicity among vehicles with less than five milliseconds deviation. This turned out to be sufficient for the verification of vehicle functions with a velocity of up to 20 m/s. For synchronization at least at the beginning of each measurement, a connection between the devices is required. A single adjustment of the clocks may lead to inaccuracies of the time stamps at longer test runs, whereas a continuous adjustment with IEEE1588 can be difficult to be implemented especially for moving vehicles without the existence of a direct wired connection. Thereby, the clocks are usually synchronized amongst themselves and not to a global time like provided by the Global Positioning System (GPS) [17].

IV. Testing Platform optimized for Driver Assistance Systems

In this section, the RTA and the approach for a HiL testing platform based on the RTA device that addresses the requirements for an agile usage as well as for the timing issues are introduced.

A. The Real Time Adapter

The RTA [18] combines the functionality of a mobile data logger and an intelligent I/O device for CAN. Its core functions [19] are implemented in VHDL to increase the execution speed and run them as parallel as possible on the built-in FPGA. Additionally the RTA’s time base, which is implemented in VHDL as well, can be synchronized to the time provided by a GPS receiver with a deviation of less than 10 µs [18]. 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. In both cases, incoming as well as outgoing CAN messages can be processed with respect to the GPS-time.

As illustrated in Figure 4, 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 it detects timing violations caused by messages with time stamps that cannot be transmitted in time. For the timing violation detection, the RTA compares the intended sending point of each message with the actual transmission time taking account of an adjustable tolerance as described in [20]. The RTA decouples the non-real-time behavior of the PC from the precise real-time behavior towards the SUT, which allows the execution of complex test cases with a repeatable precise timing behavior 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 at its limits.

The use of RTA devices in test environments allows the recording and the replay of CAN messages with high temporal precision. But without time synchronization between RTA devices the time stamps may not be generally comparable, because they refer to built-in oscillators with individual drifts. For the verification of Driver Assistance Functions during road tests, e.g., of an ACC as illustrated in Figure 5, it is necessary to record CAN messages in different vehicles with synchronized time stamps. This allows a comparison of the time stamps within the limits of the clock synchronization accuracy. Having these synchronized CAN traces enable us, e.g., to answer questions such as "How much time elapsed from pressing the brake pedal in the vehicle in front until the ACC decelerates the following vehicle?". Another use case with a detailed example is given in Section V.

The synchronization of the RTA devices uses a GPS based approach as shown in Figure 6 that ensures a comparable global time base for the time stamps as well as a continuous adjustment of the internal clocks during the test runs. In addition to the position and time information, GPS receivers

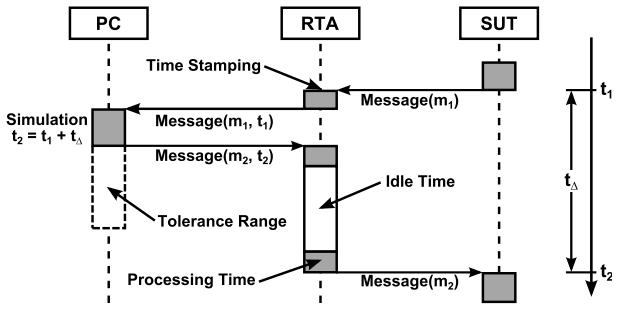


Figure 4. Sequence diagram of CAN RX/TX with an RTA

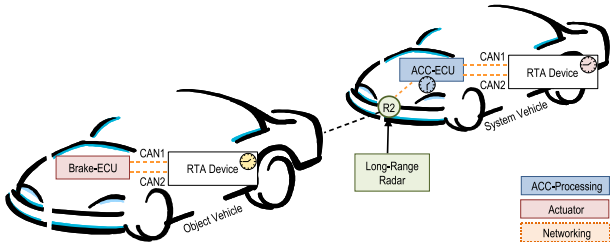


Figure 5. Schematic representation of an ACC test case

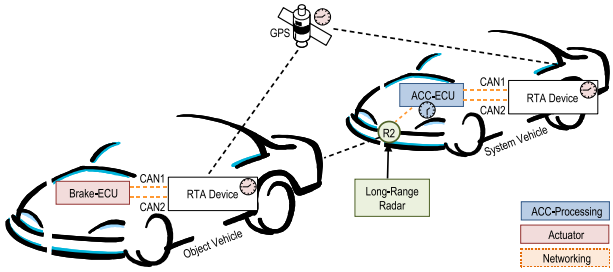


Figure 6. GPS synchronized RTA devices

provide a high-precision output so-called Pulse Per Second (PPS), which signalizes the start of a second. The time information initially preloads the clock of the RTA, whereas the PPS and the built-in oscillator increment the clock. Assuming that the drift of the RTA's oscillator does not abruptly change under ordinary circumstances but rather is slowly influenced by, e.g., temperature and / or aging. These assumptions did hold throughout all our experiments in the last four years. The RTA counts the number of oscillator ticks per GPS-second and thus obtains the time step per tick to be used by the clock for the next second. This procedure ensures a highly precise time base. Combined with the RTA's Time Stamping Unit it is possible to attach a time stamp based on the GPS-time to each CAN message.

B. HiL Testing Platform Based on an RTA and a COTS Computer

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 3 is not acceptable anymore. Additionally, if the implemented function, e.g., for interacting vehicles [21], 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 [18] and [19]. 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

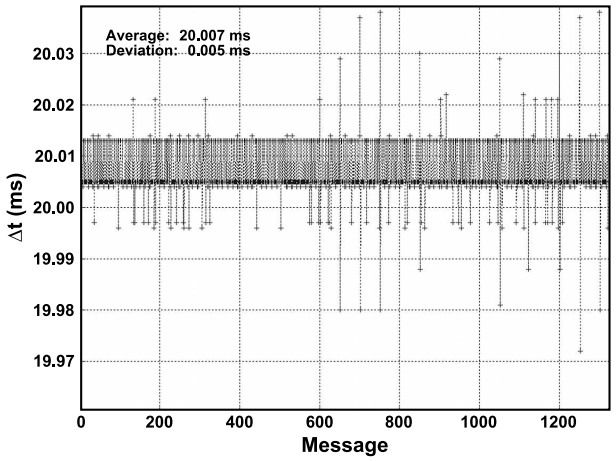


Figure 7. RTA HiL – 20 ms cycle message

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 time consuming or concurrent executions during a test run, e.g., anti-virus scanners, mail software, automatic update clients or mouse movements. 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 7 shows the result of the introduced solution for a current ECU with Driver Assistance Functions. The intended cycle time of 20 ms is kept by the RTA HiL 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.

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. Application of the Introduced Platform in Real World Examples

In the following, an in-vehicle verification of a Driver Assistance System as well as a test of a monitoring algorithm are presented. These examples describe the current usage of the RTA in the automotive industry.

A. In-Vehicle Verification of a Driver Assistance System

In Figure 6, the test setup for an in-vehicle verification of an ACC system is shown. Both the preceding and the following vehicle are equipped with an RTA device. The vehicles drive one behind the other as illustrated in the figure. The goal of this in-vehicle test is to verify the information about the relative speed between the preceding vehicle and the following vehicle delivered by the long-range radar.

Definition (System Vehicle)**.** *The system vehicle is the vehicle equipped with the Driver Assistance System that is to be tested.*

Definition (Object Vehicle)**.** *The object vehicle is the preceding vehicle in the test case that is tracked by the long-range radar of the following vehicle.*

In this test, the system vehicle is the following vehicle that tracks the preceding vehicle with its long-range radar. One RTA records thereby the velocity of the object vehicle, whereas the other RTA records the velocity of the system vehicle as well as the relative velocity between both vehicles determined by the long-range radar. Overall, the following four road tests were performed in which the object vehicle decelerates to a standstill and the system vehicle is stopped to avoid a rear-end collision:

- Manual stop of the system vehicle carried out by a defensive human driver.
- Manual stop of the system vehicle carried out by an aggressive human driver.
- Automatic stop of the system vehicle carried out by the ACC with the desired distance set to the maximum possible value.
- Automatic stop of the system vehicle carried out by the ACC with the desired distance set to the minimum possible value.

After the test runs the velocity information of the object vehicle (v_{obj}) and the system vehicle (v_{sys}) recorded by the two GPS synchronized RTA devices is merged and visualized on a common time axis. Figure 8 shows the velocity curve of the object vehicle and of the system vehicle as well as the difference of the velocities ($v_{rel,calc}$). Furthermore, the relative velocity provided by the long-range radar ($v_{rel,radar}$) is shown as a dotted line of the provided data points. The steep curve of v_{sys} relative to v_{obj} in Figure 8 shows that the defensive driver of the system vehicle responds very quickly and with strong braking to the

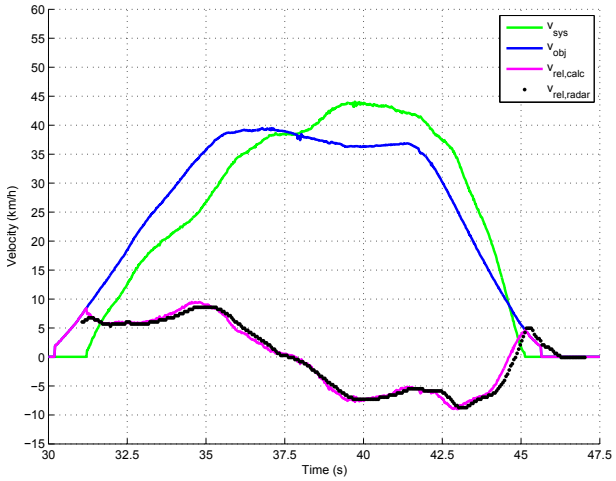


Figure 8. Driver with a defensive braking response

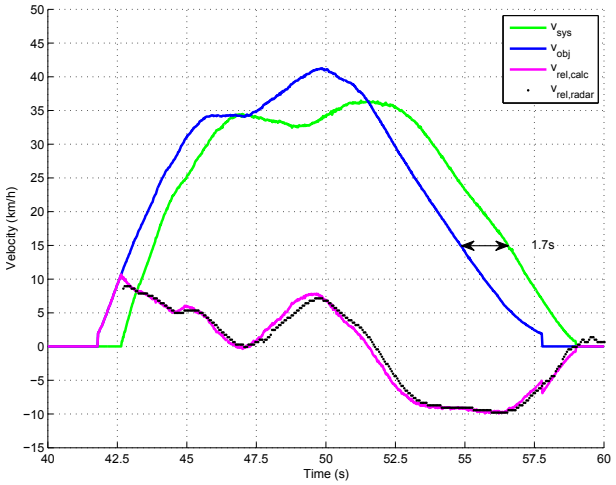


Figure 10. ACC with the desired distance set to the "MAXIMUM"

deceleration of the object vehicle. Thus, the system vehicle stops at 45.13 s while the object vehicle comes to a standstill at 45.65 s.

Figure 9 shows a stopping process of the system vehicle carried out by an aggressive driver. Thereby, the system vehicle stops in the test run just behind the object vehicle and nearly at the same point in time. This is shown in the graph by the fact that the velocity curves of both vehicles reach 0 km/h almost simultaneously. Moreover, it is seen that there are points in time where the long-range radar does not provide information about the relative velocity between the object vehicle and the system vehicle. In the time range between 54.91 s and 56.81 s, the object vehicle could not be correctly recognized.

Figure 10 shows a stopping process of the system vehicle carried out by the ACC with the desired distance set to the maximum possible value. Thereby, the Driver Assistance Function decelerates the system vehicle depending on the object vehicle. The ACC is trying to keep the specified distance. It is apparent that the system vehicle responds in the road test with a delay of about 1.7 s. The nearly parallel course of v_{sys} and v_{obj} indicates that the ACC keeps the driver's distance setting most of the time. In contrast, the defensive driver in Figure 8 has decelerated the system vehicle more than the driver of the object vehicle did, which can be seen by the intersection of the velocity curves at the end of the measurement.

Figure 11 shows a stopping process of the system vehicle carried out by the ACC with the desired distance set to the minimum possible value. It is apparent that the curves of v_{sys} and v_{obj} run parallel to each other with a larger temporal distance of approximately one second compared to the test in Figure 10. Thereby, the larger temporal distance results in a standstill of the system vehicle at a later time.

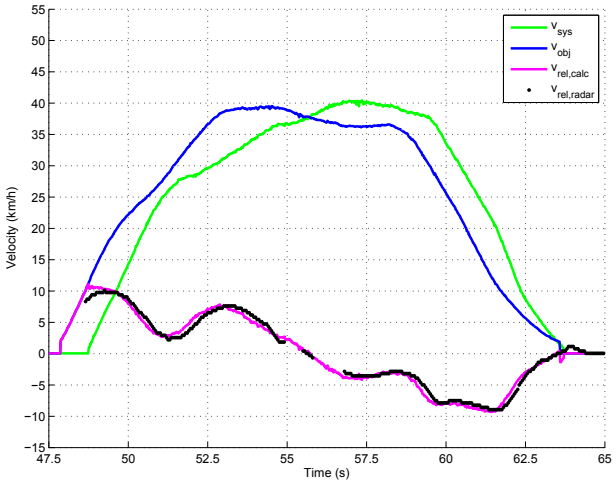


Figure 9. Driver with an aggressive braking response

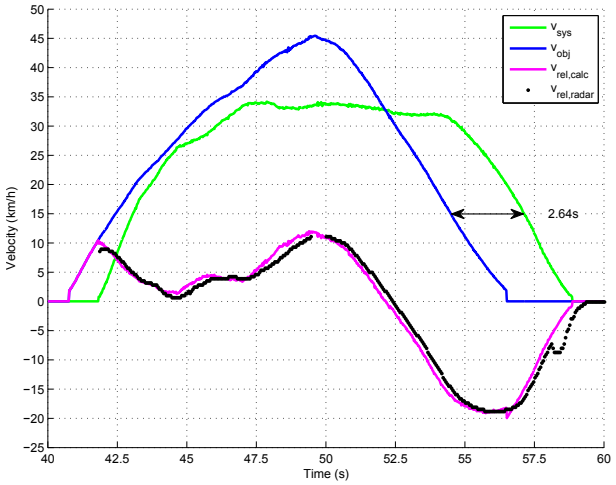


Figure 11. ACC with the desired distance set to the "MINIMUM"

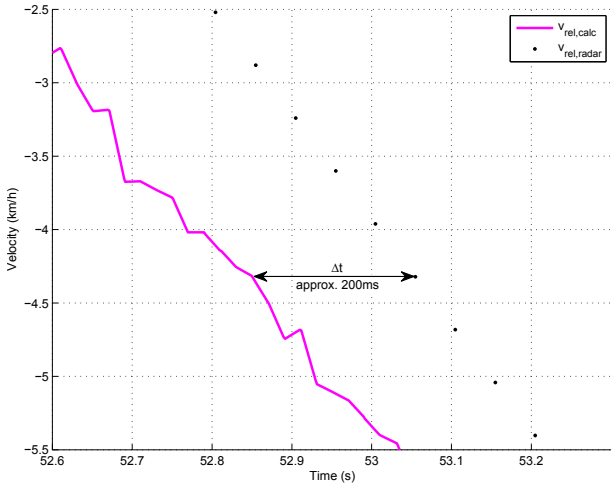


Figure 12. Processing time of the radar ECU

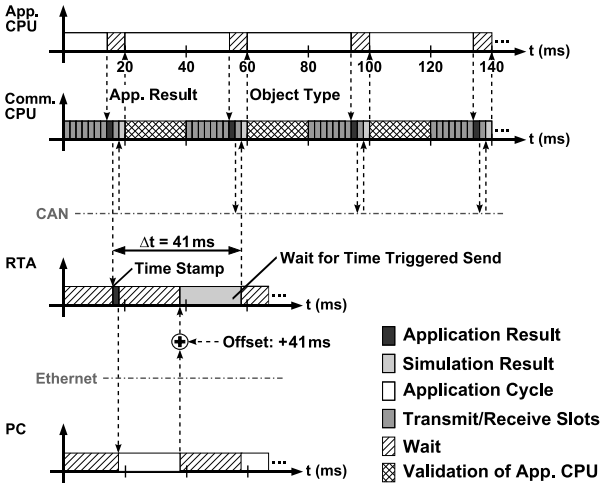


Figure 14. Synchronization of the testing platform with the SUT

Just before standstill $v_{rel,radar}$ is inaccurate. Beginning from 58.10 s, $v_{rel,radar}$ diverges from $v_{rel,calc}$.

In accordance to [22], the radar ECU should provide the information with a maximum delay of 250 ms. To verify this precept, an enlargement of Figure 11 is shown in Figure 12, in which the delay (Δt) between $v_{rel,radar}$ and $v_{rel,calc}$ is marked. Thereby, Δt is calculated according to

$$\Delta t = t_{v_{rel,radar}}(v) - t_{v_{rel,calc}}(v)$$

with $v = -4.32\,km/h$. In this case, Δt is about 200 ms.

B. Test of a Monitoring Algorithm

An example for a HiL test setup used in the automotive industry is displayed in Figure 13, 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 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

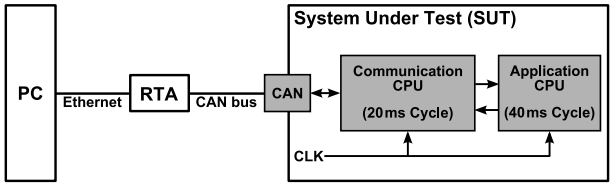


Figure 13. Example for a test setup

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 14. 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 illustrates 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.

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.

Figure 15 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 developer's 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 slice based RTOS to the SUT would have meant to change the complete simulation kernel and therefore several days of implementation work.

```
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

```
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

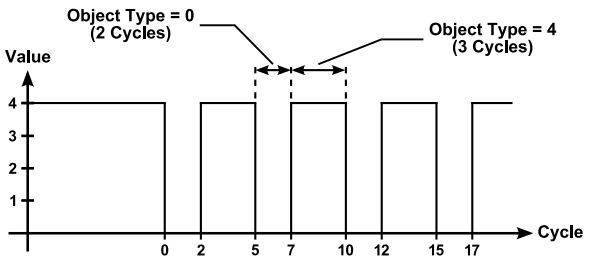


Figure 15. CAN trace for cyclic stimulation

VI. Conclusion and Future Work

The measurements demonstrated 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. For each variant, the tests can be executed at maximum in real-time. 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. An additional benefit of the introduced RTA is that the same hardware supports in-vehicle testing of Driver Assistance Functions. It has been shown that the precise time stamping synchronous to the global GPS-time can be used to correlate the information of several vehicles in one diagram on a common time axis. The reaction of the system vehicle to an action executed by an object vehicle can be evaluated quantitatively. It has been shown that the information about the object vehicle provided by the long-range radar can be verified independently by the information of the two RTAs.

Combining the synchronization to GPS with the time triggered sending mechanism of the RTA enables the highly precise replay of data that is recorded during an in-vehicle test. It is left for future work to evaluate how the RTA can support ECU testing with recorded data. For recorded data as well as for in-the-loop simulations, the GPS synchronization combined with the time triggered sending ensures that

independently of time, location and RTA device a test case can be repeatedly executed with precisely the same timing.

REFERENCES

[1] Daniel Ulmer, Steffen Wittel, Karsten Hünlich and Wolfgang Rosenstiel, "A Hardware-in-the-Loop Testing Platform Based on a Common Off-The-Shelf Non-Real-Time Simulation PC," in *ICONS 2011, The Sixth International Conference on Systems*, 2011.

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

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

[4] Christoph Marscholik and Peter Subke, *Road vehicles - Diagnostic communication: Technology and Applications*. Hüthig, 2008.

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

[6] Bart Broekman and Edwin Notenboom, *Testing Embedded Software*. Addison-Wesley, 2002.

[7] National Highway Traffic Safety Administration, "Forward Collision Warning System Confirmation Test," 2008.

[8] Christoph Marscholik and Peter Subke, *Datenkommunikation im Automobil*. Hüthig, 2007.

[9] ETAS GmbH, "LABCAR System Components," Access Date: December 28, 2011. [Online]. Available: http://www.etas.com/en/products/labcar_system_components-details.php

[10] ETAS GmbH, "LABCAR-RTPC - Real-Time Simulation Target for HiL Testing," Access Date: December 28, 2011. [Online]. Available: http://www.etas.com/en/products/labcar_rtpc.php

[11] Gerd Wittler and Jürgen Crepin, "Real-time and Performance Aspects of Hardware-in-the-Loop (HiL) Testing Systems," *ATZonline*, 2007.

[12] Jan Kiszka, "Xenomai: The RTOS Chameleon for Linux," Real-Time Systems Group, Leibniz Universität Hannover, Tech. Rep., 2007.

[13] G.i.N. Gesellschaft für industrielle Netzwerke GmbH, "MultiLog," Access Date: December 30, 2011. [Online]. Available: http://gin.de/index.php?device=1002&lang=de

[14] David Mills, "Network time protocol," RFC 958, Internet Engineering Task Force, 1985.

[15] Kang Lee and John Eidson, "IEEE-1588 Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems," in *In 34th Annual Precise Time and Time Interval (PTTI) Meeting*, 2002.

[16] Jürgen Luther and Hans-Werner Schaal, "Luftbrücke für Busdaten," *Elektronik automotive*, 2010.

[17] Guochang Xu, *GPS - Theory, Algorithms and Applications*. Springer, 2007.

[18] IT-Designers GmbH, "Real Time Adapter Datasheet (RTA-C4ENa)," 2010.

[19] Daniel Ulmer, Andreas Theissler and Karsten 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.

[20] Daniel Ulmer and Steffen Wittel, "Approach for a Real-Time Hardware-in-the-Loop System Based on a Variable Step-Size Simulation," in *Proceedings of the 22nd IFIP International Conference on Testing Software and Systems: Short Papers*, 2010.

[21] Daniel Ulmer and Andreas 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.

[22] Hermann Winner, Stephan Hakuli and Gabriele Wolf, *Handbuch Fahrerassistenzsysteme*. Vieweg+Teubner Verlag, 2009.

Autonomous Geo-referenced Aerial Reconnaissance for Instantaneous Applications

A UAV based approach to support security and rescue forces

Axel Bürkle, Florian Segor
Matthias Kollmann, Rainer Schönbein
Department IAS
Fraunhofer IOSB
Karlsruhe, Germany
{axel.buerkle, florian.segor, matthias.kollmann,
rainer.schoenbein}@iosb.fraunhofer.de

Dimitri Bulatov, Christoph Bodensteiner,
Peter Wernerus, Peter Solbrig
Department SZA
Fraunhofer IOSB
Ettlingen, Germany
{dimitri.bulatov, christoph.bodensteiner,
peter.wernerus, peter.solbrig}@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. One focus 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. If the videos taken from these swarms are geo-referenced, a significantly better coordination and distribution of missions and tasks can be achieved. Therefore, the ability of geo-referencing in unknown terrain is highly demanded. In the absence of an onboard internal navigation system, an image-based approach is proposed. It consists of registration of video frames or even whole sequences stemming from different UAVs (Unmanned Aerial Vehicles) into a common coordinate system and matching with an orthophoto. These processes are called quasi-intrasensorial and intersensorial registration. For image-based geo-referencing, a differentiation must be made between scenes with negligible spatial depth (2D situation) and those where the depth cannot be neglected any longer (2.5D situation). Several applications of geo-referenced UAV-borne videos as well as ideas how the task of image-based geo-referencing can be accelerated for online processing are presented.**

Keywords - aerial situation image; unmanned aerial vehicles; swarm; geo-referencing

I. INTRODUCTION

This paper presents our recent work on a universal ground control station called AMFIS ("Aufklärung mit Miniatur-Fluggeräten im Sensorverbund", or reconnaissance with miniature aerial vehicles in a sensor network) [1]. AMFIS is a component-based modular construction kit that supports various aerial reconnaissance and surveillance tasks. It has served as the basis for developing specific products in the military and homeland security market. Applications have been demonstrated in several exercises for the PASR (Preparatory Action for Security Research) program of the European Union, the 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 or UGVs (Unmanned Ground Vehicles), stationary platforms (network cameras), ad hoc networked sensors, and a superordinated control center.

Several software modules assist the user in obtaining aerial situation pictures. One module is the photo flight tool. This is a special property of the flight route planning in AMFIS that allows creating highly up-to-date aerial pictures of a predefined area in a short time. The software module itself is designed to work both as independent standalone software and as a part of the complex control system AMFIS. Another important property of an airborne drone is its ability to orientate autonomously in an unknown terrain. This is not only necessary to guarantee situation awareness and flexibility, but also to achieve a better coordination with other drones in the swarm. If precautions are not made, commonplace applications such as mosaicking and detection of moving objects usually suffer from error accumulation and false alarms caused by 3D structures. These problems can be solved if, ideally, each pixel of each frame is assigned a 3D coordinate. The 2D registration can be achieved if this frame is geo-referenced onto an orthophoto while the 3D component could be obtained from photogrammetric or architectural databases. Unfortunately, the obvious method of registration, namely, the use of the navigation equipment onboard of the UAV, becomes less reliable for rather inexpensive and miniaturized MUAVs (Micro UAVs). The reason is the low accuracy data stemming from such a lightweight, inexpensive navigation unit. In addition, GPS data (putting aside photogrammetric data-bases) are hardly reliable in a considerable number of applications, for example, near building walls, due to multi-path propagation. Therefore, an alternative, image-based group of approaches was developed to perform geo-referencing. We will describe the methods where the spatial depth can be neglected (2D applications) and a projection of pixels is given by a

perspective transformation of plane (2D homography). However, in the case of a relatively low sensor altitude and a moderate focal length, needed in order to achieve a satisfactory resolution of the acquired images and videos, the presence of buildings and vegetation cannot be interpreted as a disturbing factor for the upcoming computations any longer without a significant loss of accuracy. As a consequence, it is important for urban terrain to extend the existing concepts by 3D or, at least, a 2.5D method. By 2.5D situation, we understand the function of the terrain altitude depending from its longitude and latitude. This assumption is reasonable for videos taken from nadir perspective. The counterpart of homography is, in this case, a depth map that assigns a depth value to almost every pixel. An intermediate result is given by a Euclidean reconstruction of sensor trajectory and a sparse point cloud.

With respect to matching tools, there must be a trade-off between a reliable registration and a reasonable computational load. We differentiate between a (quasi)-intrasensorial registration that links neighboring frames of video sequences or even different video sequences recorded by different UAVs and an intersensorial registration that allows matching scenes of a rather different radiometry. Here, a particular contribution of this work consists in creation of synthetic images for 3D-registered mosaics and modification of the well-known adaptive self-similarity approach [2] for such synthetic images. All situations mentioned in this paragraph are summarized in Table 1.

TABLE I. GEO-REFERENCING, OVERVIEW OF RELEVANT SITUATIONS

	Intrasensorial registration	***Quasi-intrasensorial registration***	***Intersensorial registration***
Participate	Neighboring frames of the same video	Different subsequences of the same video or different, but similar videos	(Mosaicked) video and orthophoto
Geometry	From frame to frame: almost the same	Different in scaling, angle of view	Can be adjusted, rasterization in 3D see Section VI.B)
Radiometry	From frame to frame: negligibly small differences	Very small differences	Large differences
Matching tool	KLT-tracking, [3]	SIFT, [4]	SIFT for moderate differences in radiometry or adaptive Self-similarities, [2] (otherwise)
2D case	Registration by 2D homography	Registration by 2D homography, [5]	Registration by 2D homography, see Section VI.A
2.5D resp. 3D case	Fundamental matrix, Euclidean reconstruction, depth maps extraction, [6], [7]	Registration by 3D homography, Section VI.B	Creating a synthetic image and registration with orthophoto by 2D homography, see Section VI.B

In the current implementation, the methods for geo-referencing can only be performed offline. However, efforts are being made to integrate them into the AMFIS station.

The paper is structured as follows: After a short survey of related work an overview of the application scenarios is presented in Section III, 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 in Section VI. The paper concludes with a summary (Section VII) and an outlook on future work (Section VIII).

II. RELATED WORK

Related work is discussed regarding the two focuses of this paper, namely flight platforms and geo-referencing.

A. Flight Platform

With respect to the photo flight, there are some projects with a similar scope.

At the "Universität der Bundeswehr" in Munich, a UAV for precision farming [8] is currently being developed. 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 off-the-shelf fixed wing model is equipped with an autopilot and either a near infrared or a high quality camera. This technique allows monitoring the biomass development and the intensity of the photosynthesis of the plants.

The AirShield project (Airborne Remote Sensing for Hazard Inspection by Network Enabled Lightweight Drones) [9][10], 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 operational picture in case of huge disasters. The aim is to detect potentially leaking CBRNE (Chemical, Biological, Radiological, Nuclear, Explosives) contaminants in their spatial extent and to carry out danger analysis without endangering human life. The swarm is supported by a highly flexible communication system, which enables 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 show 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 Section VIII).

B. Geo-referencing

In the field of geo-referencing, we refer to our previous work ([5], see also references therein), where only 2D-relevant situations were taken into account. If there is a large parallax between the frame and the orthophoto, triangular networks can be mapped onto the orthophoto, which allows creating multi-homography-based mosaics [6]. This has advantages in situations where no 3D reconstruction can be performed from the video (e.g., in the case of zooming and rotating cameras). The relevant work on registration of 3D point clouds to 2D images and the closely related problem of

pose estimation can be found in the recent work of [11] and in references cited there. In order to learn about producing synthetic images for matching, we refer to [12]. Finally, for creating dense depth maps from a set of images that allow rendering 3D clouds, a survey [13] can be recommended. Examples on algorithms for multi-view dense matching recently developed can be found in [7] and [14].

III. Application Scenarios

The sense of security in our society has significantly changed over the past several 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 support the rescue forces in decision making. The search for people buried alive after a building collapses, or the analysis 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 slim and therefore the present situation is often very unclear to the rescue forces at the beginning of a mission. However, these are exactly those situations, for which it is extremely important to understand the context as fast as possible to initiate suitable measures.

An up-to-date aerial image can be a valuable additional piece of information to support the briefing and decision making. 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 until 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”, explained in Section V, can provide the missed information by creating an overview of the site of interest in a very short time. The application can be used immediately at the beginning of the mission with relative ease and 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 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 [15], the photo flight tool was designed to make use of a UAV swarm. By using more than one 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 work 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.

The application is responsible for all additional work, such as the composition of the respective flight routes, the control of the single UAVs including the observation of the aerial security to avoid collisions as well as setting the return flight.

IV. Flight Platform and System Architecture

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.

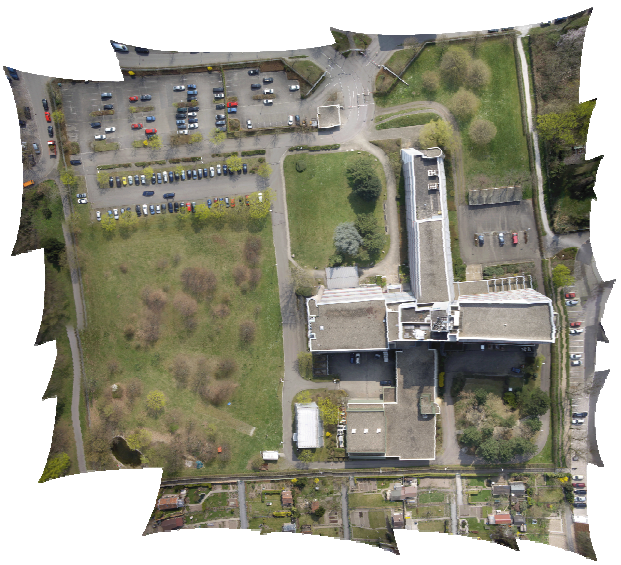


Figure 1. Situation picture from photo flight (ca. 9500 x 9000 pixel)

Figure 2. Sensor platform AirRobot 100-B.

A. *Flight Platform*

Enormous 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 for 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 an autonomous flight. A platform that fulfils these requirements is the quadrocopter AR100-B by AirRobot (see Figure 2). It can be controlled both 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 MD4-200 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.

B. *Software Architecture*

The AMFIS ground control station's software architecture is basically 3-tiered, following a pattern similar to the MVC (Model-View-Controller) paradigm best known from web application development. The central application is the so-called AMFIS Connector (see Figure 1), a message broker responsible for relaying metadata streams within the network. Metadata is transmitted using an XML-based

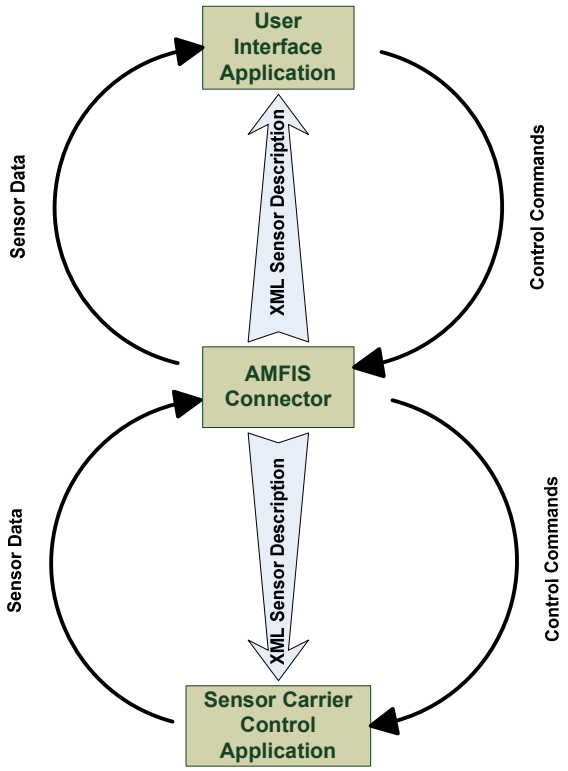


Figure 1. Software architecture of the AMFIS ground control station.

message protocol and may represent both sensor measurements as well as sensor carrier control commands. Since the biggest amount of data transmitted during a typical scenario is video (thus binary) data, the connector is tightly coupled with a second server application, the Videoserver. It is responsible for storing and distributing video streams, serving the dual purpose of providing time shifting capabilities to the network as well as reducing the load on the usually wireless links between sensor carriers and the ground control station. Since time shifting or archiving is not always required, this functionality was not integrated in the Connector in order to keep it as light-weight as possible.

Upon connection, each client application first receives an XML document describing the various sensor carriers currently active within the ground control station's network along with a unique ID used to control metadata flow. A communication library (AmfisCom) builds an object tree from the XML data, providing the application developer with a type-safe, object-oriented view of the network of sensors and sensor carriers.

A client application in this context is any application that either includes one of the numerous AmfisCom implementations (.NET, Qt, Java) or implements the AMFIS message protocol directly. Prominent examples are the GUI (Graphical User Interface) applications (analyst's interface, pilot's interface, and situation overview) or the photo flight or various transcoder applications responsible for translating metadata between the AMFIS message protocol and a

proprietary protocol used, for example, on a low-bandwidth radio data link to a distant sensor node.

After successful establishment of a connection, the Connector supplies the application with a constant stream of live sensor data. Optionally, the Videoserver provides time shifted video and metadata streams, in case an operator requires reviewing a critical situation.

V. PHOTO FLIGHT

To map 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 larger areas and should therefore receive longer flight plans than systems with a lower performance.

The flight routes must also 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. If proceeding precisely in this manner, no special flight behavior is necessary, 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 was developed at Fraunhofer IOSB, which compensates the roll, pitch and yaw angles. Nevertheless, if the photo points are flown by without a stop, an enlargement of the flight radii must be considered at turning points. As fixed-wing aircrafts may be used in future versions, more 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). Secondly, 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.

An elegant method to divide an area into subareas is the so-called “Delaunay-Triangulation” [16]. 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 had to be checked for their neighborhood relations in order to recompose them accordingly. 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. 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 recursively 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 the area being divided into more than two parts.

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 [17] also known from computer graphics. Likewise in this case, it is important to know the neighborhood relationship of the cells. Nonetheless, 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. This means a significant reduction of the required rotary and turn maneuvers of the UAV, which leads to a better cost-value ratio if 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.

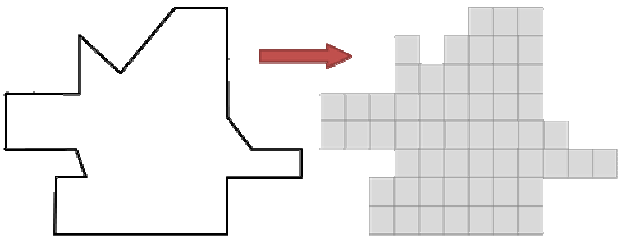


Figure 3. Scanning procedure.

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 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 then, an optimal result cannot be achieved in reasonable time (see the problem of the travelling salesman [18]).

To achieve acceptable results under the constraint to keep expenditures as low as possible, different variations were checked concurrently.

As mentioned earlier, a very steady flight route with as few as possible direction changes offers huge economic advantages, a method was developed, which processes the polygon according to its expansion in columns or line-by-line similar to a typewriter. The resulting flight route shows a clearer construction in particular with bigger areas.

Afterwards the calculated flight route is complemented with safe approach and departure air corridors to avoid collisions between the swarm members.

C. Transferring the images

In order to mosaic and geo-reference the data accumulated by the drones, the high-resolution images have to be transferred to the ground control station. The most elegant way to transfer the images is to use the downlink of the drone. This requires that the UAV has 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 disadvantage of this solution is that the frequencies for the digital video downlink of the drones are in the 2.4GHz band, which is also mostly used by Wi-Fi for broadcasting. For this reason, it can be assumed that at least the Wi-Fi transmission will experience heavy interferences. The best solution for this problem is to move either the digital video downlink or the Wi-Fi to the 5GHz 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 5GHz band.

For now, the images have to be transferred manually to the ground station.

VI. GEO-REFERENCED MOSAICKING AND APPLICATIONS

To benefit fully from the advantages of the photo flight, the images taken must be merged to an overall situation picture. In addition to the offline mosaicking based on high resolution still images, there is also the possibility to create a near real-time mosaic using the live video stream as described below.

A. 2D geo-referencing of video frames and aplications

We start a description of our geo-referencing algorithm of a video taken by a straight-line-preserving camera for the case of negligibly small depth of the scene compared to the sensor's altitude. This makes a 2D homography (plane perspective transformation, see [19]) a suitable model for registration. We denote the (global) homography between the frame I_t captured at time t of the sequence and the orthophoto by H_t, while the local transformation between the frames I_t and I_j is denoted by $H_{j,t}$. Interest points in the frame I_t will be denoted by x_t and in the orthophoto by X.

As previously discussed, the availability of GPS/INS is not constantly needed but useful for the initialization. It can be assumed that the geo-coordinates of the AMFIS base-station as well as a coarse initial value of the sensor altitude and orientation are given. As a consequence, the value of the initial homography H_1 is assigned.

The process works in the same way as described in our previous work [5] and visualized in Figure 4. The intrasensorial registration of neighboring frames of the video sequence – also known as mosaicking – is performed via KLT-tracking algorithm [3] supported by a robust method of outlier rejection (in our case, it is the RANSAC [20]

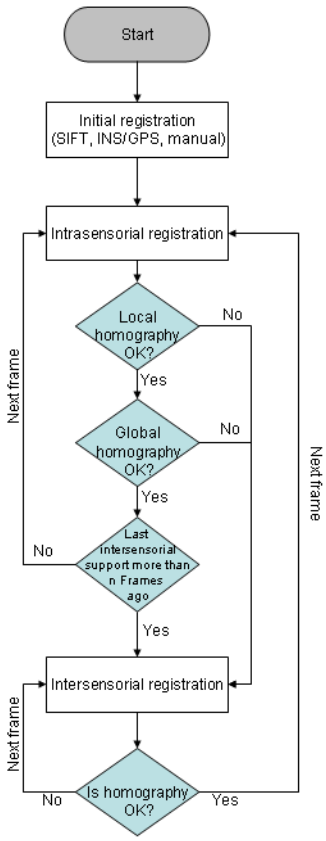


Figure 4. Flow-chart of the Image-based Georeferencing Process, see also [5].

algorithm accelerated by a $T_{1,1}$-test [21]). The quality of the local homography $H_{t-1,t}$ between two subsequent frames I_{t-1} and I_t is estimated by the number and distribution of inliers of $H_{t-1,t}$. The quality of the global homography is given by the average value of the reprojection errors between X and $H_t x_t$. As soon as the quality of either local or global homography is below a threshold, it is rejected and the intersensorial registration procedure between the current frame and the orthophoto has to be performed again.

This intersensorial registration comprises the crucial part of our procedure. The task of automatic matching images having different radiometry and also taken from slightly different viewing directions is known to be challenging. Among numerous methods [1][4][22][2] that we tested for registration, there were two candidates that most closely met our expectations.

First, the SIFT method [4] is invariant against rotations and scaling differences (moderate, until approximately 1:3). These properties, as well as the fact that SIFT is easily parallelizable and implementable on GPUs [23], make it one of the most popular state-on-the-art methods for registration, especially, when combined with a sophisticated software architecture as illustrated in Figure 4.

However, standard descriptors like SIFT do not perform well if corresponding images have a completely different radiometry. Also, when differences in resolution are insuperable for SIFT-descriptors, we follow a different approach. We use Self-Similarity descriptors [2] in combination with an Implicit Shape Model (ISM) in order to find matches between local image regions. We additionally filter the correspondences based on a self-similarity distances densely computed across both images.

Our current implementation robustly handles very large radiometric changes but has a limited scale and rotation invariance. Due to these restrictions and the involved high computational cost, we use this method after the creation of large and accurate mosaics (e.g., in the case of synthetic images described in the following section), when rotation and scaling parameters are approximately known. This also avoids unnecessary registrations.

Several applications of the registration procedure were discussed in [5]. For motion detection, weighted image differences of video frames can be computed and a threshold decision can deliver possible alarms. The advantages of the geo-referencing consists, first of all, of an elegant and efficient possibility to remove false alarms (in the areas, which are known to be parts of buildings, there can be no motion, so the false alarms in these areas are probably 3D-influenced). Secondly, trajectories of different objects are mostly easy to recognize, if they are referenced on a map. For an object reported by different sources, e.g., two UAVs of the swarm, or an UAV and a stationary camera looking from a different direction, the operator has less difficulties deciding whether all reports refer to the same object or several different objects in the case of given geo-coordinates and geo-referenced object trajectories. Finally, additional information, such as the speed of vehicles is easily calculable from their tracks on the maps and the camera frame rate. Results of geo-referenced motion detection are presented in Figure 5.

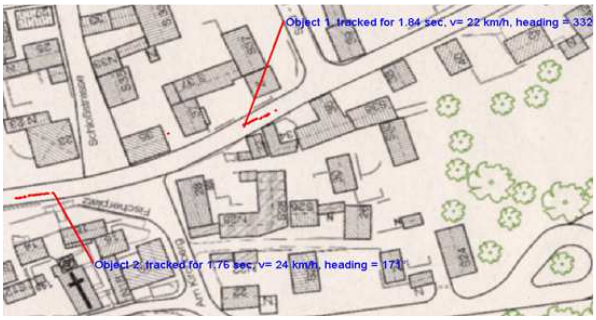


Figure 5. Two trucks have been detected and tracked on a UAV flight. Their velocities and headings can be computed, as in [5].

The other important application is given by annotation of objects of interest into the video by means of the inverse homographies. These objects can be loaded, if needed, from a data-base.

B. 3D reconstruction and geo-referencing of 3D mosaics

The assumption of a negligible spatial depth can either be made if the sensor's altitude is low, if the field of view is extremely narrow (very short focal length) or if the terrain is approximately planar. In other words, in order to achieve a satisfactory resolution of the acquired images and videos in urban terrain, the presence of buildings and vegetation can no longer be interpreted as a disturbing factor for the upcoming computations, even for Nadir views. As a consequence, the 3D structure of the scene must be taken into account for geo-referencing.

For 3D reconstruction, one of numerous approaches for structure-from-motion (SfM) or simultaneous localization and mapping (SLAM) available in the literature can be used. We use an approach of [6] because no additional information is needed here except the video stream itself. Characteristic features are detected [24] and tracked [3] from image to image. Then, fundamental matrices can be retrieved from pairs of images. As a result, a sparse set of 3D points and camera matrices in a projective coordinate system are computed. A step of self-calibration is needed to obtain an (angle- and ratio-preserving) Euclidean reconstruction, illustrated in Figure 6. In order to minimize the geometric error and thus improve reconstruction results, bundle adjustment over camera parameters and coordinates of 3D points can be activated. If the application is time-critical, dense reconstruction can be performed by creating triangular networks [16] from already available points and triangular interpolation [6]. Otherwise, a robust and accurate multi-camera approach [7] recently developed can be applied to extract the depth information from (almost every pixel) of several frames of the sequence.

From this point, we can proceed to geo-referencing of the reconstructed scene. Here, we subdivide this task into two subtasks: the first subtask is a quasi-intrasensorial regi-

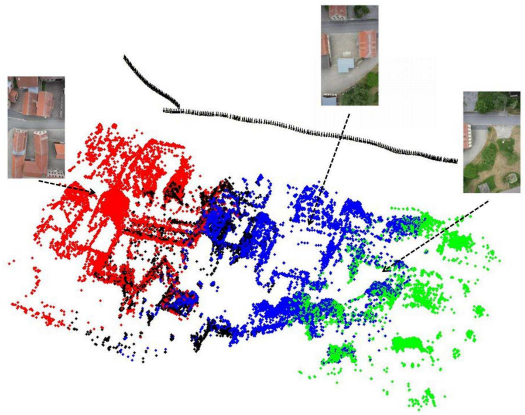


Figure 6. Result of the registration procedure of four sequences (illustrated by different colors) and the camera trajectory depicted by viewing cones. Three example images are shown as well.

stration of different subsequences of the video sequence or different video sequences of a similar appearance. These video sequences can be taken by different UAVs that carry different kinds of cameras and operate, as discussed before, in different parts of the region to be explored. As a consequence, the slightly misleading term of intrasensorial registration is now replaced with quasi-intrasensorial registration. Since the camera trajectory and point coordinates for every reconstruction in some relative, Euclidean coordinate system have different positions, orientations and scales, a registration step is necessary. The second subtask treats registration of the material obtained by videos and our 3D reconstruction procedure onto the orthophoto. A detailed explanation of these subtasks will be provided in the following two subsections.

We assume to be given two sets of camera matrices $\boldsymbol{P}_1$ (= $P_{1,1}$, ..., $P_{1,K}$) and $\boldsymbol{P}_2$ (= $P_{2,1}$, ..., $P_{2,L}$) in homogeneous coordinates that were computed by a SfM approach in different Euclidean coordinate systems (denoted by the first sub-scripts). The desired output is a spatial transformation between these coordinates systems. This transformation H (also called 3D homography) is given by a regular 4×4 matrix in homogeneous coordinates such that the relations

$$P_{1,k} = P_{2,k}H \quad \text{and} \quad P_{1,l}\,H = P_{2,l}$$

hold for $k = 1,\ldots K$ and $l = 1, \ldots, L$. Without loss of generalization, we assume that two images corresponding to cameras $P_{1,K}$ and $P_{2,1}$ cover an overlapping area of the data-set and that point correspondences c_1 and c_2 can be detected in these images by means of a matching operator (e.g. [4]). Given camera matrices $P_{2,1}$, $P_{2,2}$, ..., we now compute, by tracking points c_2 in other images of the second sequence, several 3D points Y in the second coordinate system, see the linear triangulation algorithm [19]. By backward tracking points c_1 in other images of the first workspace and Y, we obtain the set of camera matrices $Q_{1,K}$, $Q_{1,K-1}$,... via camera resection algorithm [19]. For more than one corresponding

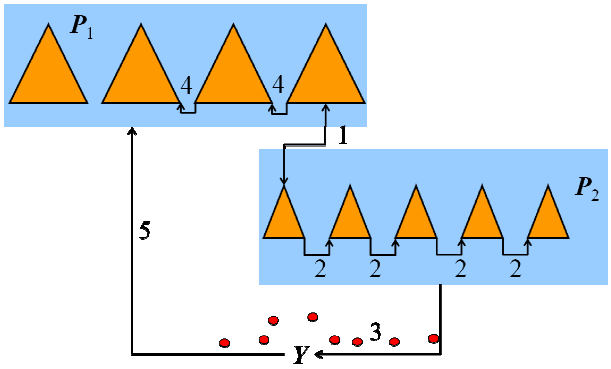


Figure 8. Result of the registration procedure of four sequences (illustrated by different colors) and the camera trajectory depicted by viewing cones. Three example images are shown as well. Registration of two reconstructions: camera locations are depicted by orange viewing cones, 3D points by red circles. Different steps are illustrated by arrows and numbers: 1: Registration of features in two sequences, 2. Tracking of registered features and 3: Triangulation in the second coordinate frame, 4: tracking and 5: obtaining camera parameters of the first reconstruction in the second coordinate frame. The transformation H is obtained from at least two corresponding camera matrices.

cameras $Q_{1,K-n}$ to $P_{1,K-n}$, the initial value of the spatial homography H as a solution of the over-determined system of system up-to-scale

$$\begin{bmatrix} P_{1,K}H \\ P_{1,K-1}H \\ \ldots \end{bmatrix} \cong \begin{bmatrix} Q_{1,K} \\ Q_{1,K-1} \\ \ldots \end{bmatrix}$$

is obtained via Direct Linear Transformation method and refined by means of a geometric error minimization algorithm. The process of the intrasensorial registration is schematically visualized in Figure 8 while Figure 6 shows an example of registering four Euclidean reconstructions for a UAV data-set. From the kink in the camera trajectory in Figure 6, one can see that we are dealing with two different UAV-flights.

We now proceed to the second part of the algorithm. In the case of nadir views, the results of the registration procedure described above can be rasterized into the xy-plane. To do this, we need both to make the z-axis coincide with the physical vertical direction (there are several simple heuristics to do this job) and to obtain the 3D information for a dense set of points. This was done by our approach [7] for

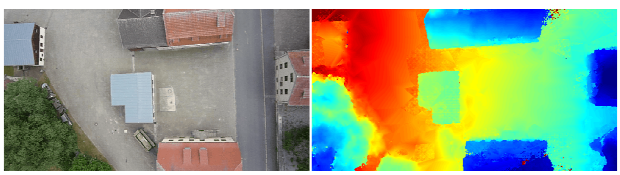


Figure 9. An image and the corresponding depth map obtained with [7]. This is the middle image of Figure 6. Small outliers do not cause any trouble since they can be removed during the rasterization procedure.

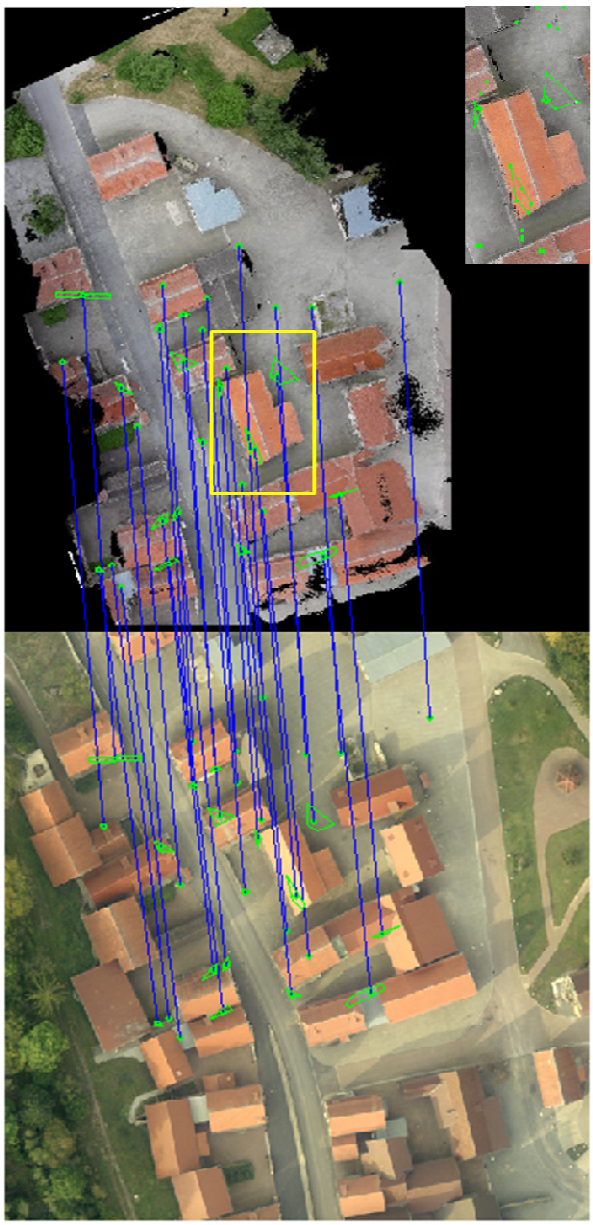


Figure 10. The synthetic image (top) and fragment of the orthophoto (bottom), as well as and clusters (green) of the RANSAC-inliers (connected by blue lines) determined by our extension of the self-similarities algorithm.

computing depth maps. In the future, it will be interesting to investigate to what extent the meshes obtained via triangular interpolation from already available points can compensate for forfeits in the quality of the synthetic image. We show an exemplary depth map in Figure 9 and the synthetic image itself in the top of Figure 10. The result of the registration of the synthetic image to the orthophoto is visualized in Figure 10 and Figure 11. We also mention that the points situated in elevated regions can be optionally identified in the depth maps and excluded from further consideration.

Figure 11. The synthetic image registered onto the orthophoto.

VII. CONCLUSIONS

The described algorithms for the photo flight application were implemented as a software library and are integrated into a geographic information system based on ESRI software [25] specially provided for test purposes. The photo flight tool is an independent software module whereas the logic behind it is interchangeable. 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 theoretically unlimited number of UAVs thereby means a huge advancement in the field of local just-in-time reconnaissance and geo-referencing.

With respect to the image-based geo-referencing without using internal navigation, we showed a robust and autonomous approach that works both in situations of 2D registration (with applications of motion detection and object annotation) as well as 3D registration of Euclidean reconstructions and matching of a synthetic image thus obtained with an orthophoto, even in the case of different radiometry. To create a synthetic image, the assumption of 2.5D surface (terrain skin $z(x, y)$) must hold; as a consequence, our methods work better for almost nadir-views of video frames. In the case of an oblique view with a non-negligible spatial

Figure 12. Gas sensor to detect inflammable gases, Ammonia, Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Chlorine.

depth, algorithms of pose estimation can be applied, but they stay beyond the scope of the work presented here.

Detection of motion in 3D scenes is carried out by an accurate occlusion analysis already on the state of multi view dense depth maps computation while annotation can be generalized from the 2D case.

VIII. FUTURE WORK

Parallel to the work on the photo flight algorithms, a small gas sensor, which can also be carried as a payload by a UAV, was developed in cooperation with an industrial partner (see Figure 12). 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, in addition, contains a photo-ionization detection sensor and a sensor to detect universal inflammable gases. 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. Since the aim of this application differs from the original task of 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.

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.

The approaches of geo-referencing are robust, nearly fully-automatic, and real-time oriented; that is, reconstruction goes along with the video sequence from its beginning to its end and is *not* supposed to be performed after the whole movie has been captured. Due to relatively slow matching algorithms [4] and [2], the 3D reconstruction can only be performed offline at the current state of the implementation, but one of our goals for future work consists of creating a real-time interface for estimation of camera trajectory and a sparse point cloud. An obvious hardware-based acceleration of our algorithm consists of intersensorial registration on GPU parallel to the already extremely fast intrasensorial registration. By better exploitation of initial homography (guided matching) and neighborhood relations, the computing time and memory load for point matching can be drastically reduced.

The only interactive part is given by determination of images with covering overlapping parts of the terrain for quasi-intersensorial registration of Section VI.B. However, in the cases of almost nadir views over the urban terrain and, thus a 2.5D representation of the scenery, it will be possible, in the future, to automate the approach by identifying the search space via *xy*-coordinates.

ACKNOWLEDGMENT

The authors would like to thank the following people for their contributions: Sven Müller, Steffen Burger, Thorsten Ochsenreither and Judy Lee-Wing.

REFERENCES

[1] F. Segor, A. Bürkle, M. Kollmann, and R. Schönbein, “Instantaneous Autonomous Aerial Reconnaissance for Civil Applications - A UAV based approach to support security and rescue forces,” The 6th International Conference on Systems ICONS 2011, January 23-28, 2011, St. Maarten, The Netherlands Antilles, pp.72-76, 2011.

[2] E. Shechtman and M. Irani, “Matching Local Self-Similarities across Images and Videos,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Minneapolis, USA, pp. 1-8, 2007.

[3] B. Lucas and T. Kanade, “An Iterative Image Registration Technique with an Application to Stereo Vision”, Proceedings of 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674-679, 1981.

[4] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal on Computer Vision (IJCV), vol. 60, no. 2, pp. 91-110, 2004.

[5] P. Solbrig, D. Bulatov, J. Meidow, P. Wernerus, and U. Thönnessen, “Online Annotation of Airborne Surveillance and Reconnaissance Videos,” The 11th International Conference on Information Fusion, Köln, Germany, pp. 1131-1138, 2008.

[6] D. Bulatov, “Towards Euclidean Reconstruction from Video Sequences,” Int. Conf. Computer Vision Theory and Applications (2), pp. 476-483, 2008.

[7] D. Bulatov, P. Wernerus, and C. Heipke, “Multi-view Dense Matching Supported by Triangular Meshes”, ISPRS Journal of Photogrammetry and Remote Sensing, vol. 66, no. 6, pp. 907–918, 2011.

[8] Universität der Bundeswehr München, Germany, http://www.unibw.de/lrt13_2/Forschung/Projekte/UAVPF, 18.01.2012.

[9] 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, pp. 196-200, 2009.

[10] 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.

[11] V. Lepetit, F. Moreno-Noguer, and P. Fua: “EPnP: An Accurate O(n) Solution to the PnP Problem”. International Journal of Computer Vision 81(2), pp. 155-166, 2009.

[12] G. P. Penney, J. Weese, J. A. Little, P. Desmedt, D. L. G. Hill, and D. J. Hawkes: A Comparison of Similarity Measures for Use in 2D-3D Medical Image Registration. IEEE Trans. Med. Imaging 17(4), pp. 586-595, 1998.

[13] D. Scharstein and R. Szeliski. "A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms". International Journal of Computer Vision, 47(1), pp.7-42, 2002.

[14] C. Strecha: "Multi-view Stereo as an Inverse Inference Problem", PhD Dissertation, KU Leuven, Belgium, 2007.

[15] A. Bürkle, F. Segor, and M. Kollmann, "Towards Autonomous Micro UAV Swarms," Proceeding of the International Symposium on Unmanned Aerial Vehicles, Dubai, UAE, 2010.

[16] B. N. Delaunay, "Sur la sphere vide," In: Bulletin of Academy of Sciences of the USSR 7, No 6, pp. 793-800, 1934.

[17] D. Hearn and M. P. Baker, "Computer Graphics, C version, 2nd Ed," Prentice Hall, 1997.

[18] D. L. Applegate, R. E. Bixby, V. Chvátal, and W. J. Cook, "The Traveling Salesman Problem. A Computational Study," Princeton University Press, Februar 2007.

[19] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision", Cambridge University Press, 2000.

[20] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", Communications of the ACM, vol. 24, no. 6, pp. 381-395, 1981.

[21] J. Matas and O. Chum, "Randomized RANSAC with Td,d –test", Proceedings of the British Machine Vision Conference (BMVA), vol. 2, pp. 448-457, 2002.

[22] P. A. Viola and W. M. Wells, "Alignment by Maximization of Mutual Information," International Journal of Computer Vision (IJCV), 24(2), pp. 137--154, 1999.

[23] C. Wu, "A GPU Implementation of Scale Invariant Feature Transform (SIFT)," http://cs.unc.edu/~ccwu/siftgpu, 18.01.2012.

[24] C. G. Harris and M. J. Stevens, "A Combined Corner and Edge Detector," Proc. of 4th Alvey Vision Conference, pp. 147–151, 1998.

[25] Esri Enterprise, USA, http://www.esri.com, 18.01.2012.

The Strategic Role of IT as an Antecedent to the IT Sophistication and IT Performance of Manufacturing SMEs

Louis Raymond
Université du Québec à Trois-Rivières
Trois-Rivières, Canada
louis.raymond@uqtr.ca

Anne-Marie Croteau
Concordia University
Montréal, Canada
anne-marie.croteau@concordia.ca

François Bergeron
TELUQ-Université du Québec à Montréal
Québec, Canada
bergeron.francois@teluq.uqam.ca

Abstract — **The business value of IT (information technology) applications for SMEs (small- and medium-sized enterprises) is dependent upon how such applications "fit" with the strategic orientation of these firms. Based on the strategic alignment of IT, this study uses a systemic approach to test the contribution of three predictors of IT performance in an organization: the strategic role of IT as well as the sophistication of the management and the use of IT. A multivariate mediation perspective is used to conceptualize alignment. The results of an empirical investigation of 44 manufacturing SMEs establish an important mediating effect of IT management and IT usage sophistication between the strategic role of IT and IT performance.**

Keywords - IT sophistication; IT performance; SME; strategic alignment; strategic role; e-business applications

I. INTRODUCTION

The current economic context is marked by a considerable expansion of electronic markets. The need for IT (information technology) comes with the ever increasing demand for information and communication. The impact on business is tremendous, especially on SMEs (small- and medium-sized enterprises) who need to invest in systems with the ability to store, process, and generate data stemming from their dealings with various business partners [37]. For many small companies, investing in IT has become necessary to ensure their survival and competitiveness in the face of globalization [27], and especially to enable their innovation capabilities [17]. While IT-based systems have substantially improved the manufacturing process and productivity of SMEs, they have also allowed for more organizational flexibility by removing constraints of time and space, and by reinventing the internal and external mode of business organization [16].

With the opening of the global marketplace, SMEs have been subject to rapid change and great instability. Now, IT can play an important role in a company's performance and its ability to respond effectively to the changing needs of the market, therefore special attention for the small business domain has been demanded of researchers [35]. Given the inherent fragile nature of the SME, IT plays an increasingly strategic role in creating new managerial challenges for these firms [11]. New online competitors easily enter emerging or existing markets, customers are more informed and more demanding because they can compare features and prices of products through the Web, and the changing needs and wants of the market often render recent IT investments obsolete [39].

As a result of substantial IT investments by many manufacturing SMEs, it has become essential to foresee the threats and opportunities that are inherent in these technologies, to discover the mechanisms that manage and drive IT, and to analyze their impact in terms of cost-effectiveness and profitability for such enterprises [24]. The increased strategic nature of the role of IT in the organization may give rise to IT management and IT usage problems that may be not only technical but also strategic and organizational in nature [6]. It has therefore become important for SMEs to understand how their investments in IT, coupled with an increased understanding of their IT management and usage practices, can provide the most value to them [12].

This research thus aims to study the impact of the strategic role of IT upon the IT management sophistication, the IT usage sophistication and the IT performance of manufacturing SMEs by answering the following research question: *To what extent and in what manner do the strategic role of IT, the sophistication of IT management, and the sophistication of IT usage contribute to the realization of business value from IT by manufacturing SMEs?* We first present the theoretical background of the research, followed by the research model, and the method by which 44 French manufacturing SMEs were empirically studied in order to answer the research question. Next, the results are presented and discussed. We further identify the study's implications and limitations, and conclude with future research.

II. THEORETICAL BACKGROUND

The study's theoretical background is founded on the concept of *strategic alignment*, which is at the core of the strategic paradigm in information systems research. First defined in terms of its impact on organizational performance rather than on the attainment of business value from IT, this concept still constitutes a fundamental basis of our understanding of the strategic role of IT and IT performance in organizations. According to Henderson and Venkatraman [20], strategic alignment is founded on the assumption that the firm's realization of business value from information technologies results from a dynamic coherence among the firm's business strategy, organizational infrastructure, IT

strategy, and IT infrastructure. SMEs should thus "align" their IT processes and IT capabilities with their business processes and organizational capabilities. As presented in Figure 1, Henderson and Venkatraman's model is based on a systems perspective of alignment, emphasizing the importance of aligning both internal and external business activities in order to achieve strategic objectives and improve organizational performance.

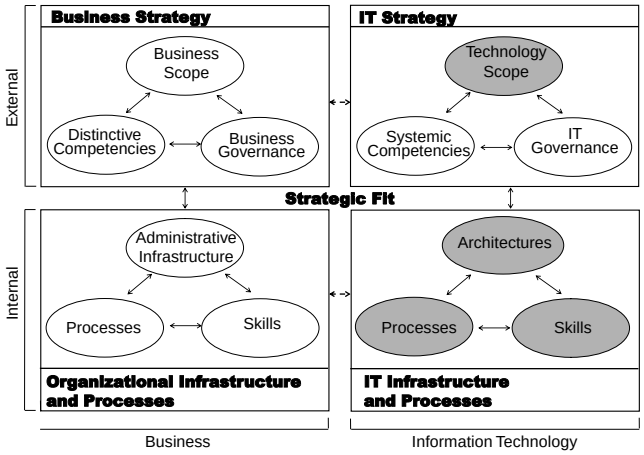


Figure 1. Adaptation of Henderson and Venkatraman's [20] IT alignment framework

As conceptualized above, the strategic alignment of IT has studied in various ways over the past two decades, often by exploring some of its aspects or dimensions in more detail. While it has been demonstrated that coherence between business strategy and IT strategy contributes to both IT performance and organizational performance [8] [13] [25], few studies have taken into account exogenous factors such as the organizational context (business strategy, organizational structure), the environmental context (industry, firm size), and the technological context (technology solutions, IT management), that is, the "TOE" framework emanating from Tornatsky and Fleischer's [42] work. Despite previous empirical studies that have allowed us to better understand the contexts in which strategic alignment contributes to the attainment of business value from IT and to organizational performance, many aspects remain unexplored, including alignment at the technological level [2].

This study proposes a research model for ascertaining in what manner IT "works" in SMEs, that is, in terms of the strategic role of IT, and the sophistication of IT management and IT usage. The attainment of business value from IT, namely IT performance, is seen here as a result of direct or "proximal" strategic alignment of IT [8], while organizational performance would be considered rather as an indirect consequence or "distal" of this alignment. Returning to Figure 1, the shaded sections of Henderson and Venkatraman's model are the basis of the research model used for the current study.

A. Strategic Role of IT

According to Powell and Dent-Micallef [34], the firm's IT capabilities and organizational capabilities must complement one another in a way that creates intrinsic benefits from IT investments and the use of IT. Thus, more emphasis is placed upon optimizing the use and management of IT based on the internal characteristics of the firm as well as its strategic profile, size, in-house IT expertise, as well as its managerial, technological, and functional capabilities. In this regard, certain researchers have explored the idea of an evolution in IT usage within the firm, such Ward, Taylor and Bond [46] who observed that the strategic role of IT is developed over three major periods in order to support the business throughout its growth cycle: 1) a period of developing data processing standards and automating repetitive tasks, thereby improving operational efficiency, 2) a period of managing information systems, designed to improve managerial efficiency by producing relevant and timely information that will be used to better manage and control the firm, and 3) a period of developing strategic information systems that enable the firm to be more competitive in a global economy.

In the evolutionary model of the role of IT proposed by Philip and Booth [33], each organization has specific expectations with regard to IT that are dependent upon its capacity to align these technologies with its strategic objectives. According to this model, information technologies can play five potential roles in the enterprise:

- survival - the goal of IT is to achieve greater control over managerial processes and day-to-day administrative and production tasks in order to improve operational performance and reduce costs;
- resources - IT is used by the firm to procure itself of resources such as materials and services from suppliers, and to provide and deliver products and services to its customers;
- competitive advantage - IT is used strategically to fully exploit the potential of resources in order to gain a competitive advantage, especially by enabling innovation;
- value analysis service - using IT to rethink business processes by reengineering them to improve the firm's competitiveness and flexibility, taking into account the rapid changes in the environment;
- cyberspace - it is through the Internet and the Web that virtual organizations build relationships with suppliers, consumers, and other organizations. This type of structure is meant to be very flexible, innovative, and to provide personalized service.

B. IT Sophistication

The evolution of the strategic role of IT is closely linked to IT sophistication because it reflects the way IT is managed

and used by the company. Now, IT sophistication can be explained by the way IT "falls into line" with the firm's strategic objectives [36]. The concept of IT sophistication and its measurement were first defined and validated by Raymond, Paré and Bergeron [38], to be subsequently used by other researchers [10] [21] [31] [36]. IT sophistication refers to the nature, complexity, and interdependence of the management and use of IT within an organization. IT management sophistication includes managerial and functional sophistication on the one hand, while IT usage sophistication includes informational and technological sophistication on the other hand.

Managerial sophistication takes into account the mechanisms used to plan, monitor, and assess current and future applications [37]. Within the context of the SME, this is demonstrated by the degree of formalism of the company's IT processes and the level of alignment with the organization's goals, including the development of applications and the level of user participation in this development. This dimension may also contain aspects related to the presence of external consultants, the development of IT resources and competencies, and the level of support for – and appropriation of – IT within the firm. Functional sophistication refers to the location and functional autonomy of IT within the organization. In the SME context, this can refer to the presence of a designated manager for IT and by the organizational level at which the IT function is positioned [4].

Informational sophistication refers to the nature, both transactional and managerial, and functional coverage (accounting, finance, HRM, logistics, production, distribution, marketing, sales, customer service) of the applications portfolio [5]. IT may also include information quality and user-system interaction quality. Another aspect of informational sophistication is the degree in which applications are integrated in the SME; this element can be characterized by the implementation of an enterprise system software package (ERP). Technological sophistication reflects the number or variety of technologies used by the SME in several areas such as CAD/CAM, internal networking and external networking, including Internet and Web technologies. It also refers to the measures put in place by the firm for purposes of IT security and confidentiality [38].

C. IT Performance

Assessing the performance of the IT function in an organization is not a simple task [28]. In a process evaluation of IT costs, Keen [23] proposed taking into account various elements such as the technical obsolescence of software, the declining cost of work units and operating software, development flows, and operating costs. Benefits gained from IT remain very complex to identify, specifically in relation to organizational performance [43]. In addition, quantifying benefits from organizational change, improved customer follow-up, or even an improvement of internal and external communication, are a challenge for a number of enterprises.

In regard to assessing the business value of IT, DeLone and McLean [14] [15] developed and validated a model of IT "success" or performance that comprises six dimensions: quality of the system, quality of information, usage, user satisfaction, individual impact of IT use, and organizational impact of IT use. User satisfaction remains however one of the most important measures of success and most recognized in IT, and it has been demonstrated that the quality of the system, the quality of the information output and the usefulness of applications point to, in large part, the satisfaction of users [40].

III. RESEARCH MODEL

As presented in Figure 2, the research model developed for this study is based upon a conceptualization of the strategic alignment of IT proposed by Henderson and Venkatraman [20], more specifically the alignment between the IT strategy and the IT infrastructure and processes that is deemed to have a positive impact upon the performance of IT in manufacturing SMEs. The IT strategy is defined as the strategic role attributed to IT by the SME's leader, whereas the IT infrastructure and processes are as the firm's sophistication in both managing and using IT. Testing this model should help us answer the research question.

As shown in the research model, the strategic role of IT is an independent construct directly related to the dependent construct, i.e., IT performance. The impact of the strategic role of IT will also be felt by the sophistication of IT management and IT usage. This research model aims to explain IT performance in a novel way by focusing on the strategic role of IT while taking into account the sophistication level of IT deployed in manufacturing SMEs. It is for this reason that the IT sophistication concept [38] is mobilized here, that is, IT management sophistication on one hand, and IT usage sophistication on the other hand. The first hypothesis is in line with the main proposition found in previous conceptualizations of the strategic alignment of IT, and is made on the basis of the evolution of information technology's role in organizations [33] [45]. Its distinction and contribution however lie in the choice of IT performance (business value of IT) rather than organizational performance as the outcome of such alignment.

It had been previously noted that the strategic role played by IT in organizations could only be ascertained if one took into account their IT management and usage characteristics. Now the notion of IT sophistication effectively reflects how IT are managed and used within organizations [46]. Hence, the second hypothesis assumes the more strategic the role played by information technology in the organization, the greater the presence of its IT function. Following Henderson and Venkatraman [20], it is presumed that in manufacturing SMEs, the strategic importance of IT will be reflected in the IT resources and capabilities developed by the IT function.

The third hypothesis reflects the premise that users will be more satisfied with the applications implemented and with the quality of information output if the SME's leadership views IT as a strategic necessity or as a source of competitive advantage. Here, the notion of "top-management support" as a determinant of IT performance would take on added importance in small business [40].

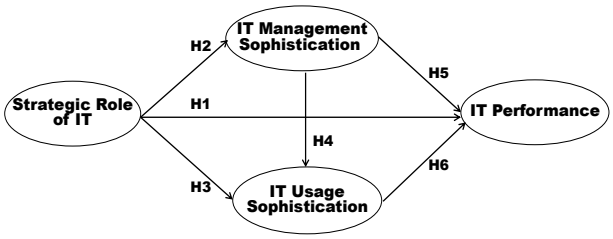


Figure 2. Research model

The fourth hypothesis assumes a certain hierarchy in the evolution of IT, as previously indicated, i.e., this technology must be effectively managed and deployed in the SME if it is to be appropriated and effectively used by employees. This is basically in line with DeLone and McLean's [15] updated IS success model in which system usage and user satisfaction are dependent upon the quality of the system, the information output, and the "service" provided by the IT function [32]. The fifth hypothesis proposes that the performance of IT improves when the sophistication of the management of IT increases [29]. As noted by Philip and Booth [33], "sustainable advantage depends on the ability to manage the IS resources effectively on an ongoing basis". The last hypothesis similarly proposes that IT performance improves with a more sophisticated usage of IT [38], this being in line again with DeLone and Mclean's [14] [15] IS success model.

In summary, the following hypotheses are tested:

H1: The more strategic the role played by IT, the higher the performance of IT.
H2: The more strategic the role played by IT, the greater the IT management sophistication.
H3: The more strategic the role played by IT, the greater the IT usage sophistication.
H4: The greater the IT management sophistication, the greater the IT usage sophistication.
H5: The greater the IT management sophistication, the higher the performance of IT.
H6: The greater the IT usage sophistication, the higher the performance of IT.

IV. METHOD

Secondary data was provided by a database created by a university research center for benchmarking purposes and containing information of 44 French manufacturing SMEs. For the study's purpose, a SME is defined as having between 10 and 299 employees, the median size of the sampled firms being 38 employees. The industrial sectors represented include metals (27%), food and beverage (16%), wood (9%), plastics (9%), textile (7%), minerals (5%), electronics (2%), and others (25%).

A. Data Collection

This database was created in collaboration with business owners that belong to chambers of commerce in Midi-Pyrénées region, by asking the management team and the IT manager to answer a questionnaire on the firm's strategic orientation, practices, and performance with regard to information technology and e-business, broken down by the main business functions of the SME, namely operations and production, sales and marketing, and accounting, finance and HRM. In exchange for this information, the firm was provided with an overall diagnostic of its situation relative to the management and performance of its information technology.

B. Measures

In view of Henderson and Venkatraman's [20] framework on which this research is based, fit or alignment between the strategic role of IT and the sophistication of IT management and usage in the firm is ascertained here from a "fit as mediation" perspective [44]. First, the extent to which IT plays a strategic role in the SME was measured through a self-typing approach based on Venkatraman's [45] and Philip and Booth's [33] stage models, by asking the chief executive to answer the following question (statements were coded from 1 to 4 in order of increasing strategic importance):

Indicate among the following statements the one that best defines your understanding of the strategic role that is assigned to information technology-based applications (ITApps) in your firm (choose one statement)?

1. ITApps should allow us to improve our managerial control and our production monitoring.	❒
2. ITApps should insure greater operational flexibility and better response to our customers' needs.	❒
3. ITApps should facilitate and accelerate the development of new products, and allow us to increase our market share.	❒
4. ITApps should allow us to integrate our business and production processes, and to improve exchanges with our business partners.	❒

The measures of IT management sophistication, in terms of managerial and functional sophistication, and of IT usage sophistication, in terms of informational and technological sophistication, emanate from constructs developed, validated, and used in previous research [31] [38]. IT

performance is measured by the level of attainment of the benefits associated with four types of IT-based applications (accounting-finance-HRM, logistics-production-distribution, marketing-sales-customer service, e-business-Internet-Web), thus following a process-based approach wherein the respondents evaluate the business value of IT for their firm [41] [43]. A list of expected benefits specific to each type of application (e.g. "increase flexibility", improve customer service", "facilitate the recruitment of personnel") is presented to the manager (CEO or CFO, operations manager, sales and marketing manager, and IT manager) who must indicate on a 5-point scale the extent to which the applications implemented contribute to the attainment of these benefits.

V. RESULTS

Descriptive statistics of the research variables are presented in Table I.

TABLE I. DESCRIPTIVE STATISTICS (N = 44)

Variable	mean	s.d.	range
Org. Size (no. of employees)	59	61	10-301
Strategic Role of IT[a]	2.5	1.3	1-4
Functional Sophistication			
designated manager for IT[b]	.727	-	0-1
org. level of the IT function[c]	.546	-	0-1
Managerial Sophistication			
IT development	3.0	0.8	1.0-5.0
IT evaluation	2.8	1.2	0.0-5.0
user participation	2.9	0.9	1.0-5.0
IT resources and competencies	3.4	0.9	1.2-5.0
IT support and appropriation	3.7	0.8	1.7-5.0
external consultants	2.4	1.8	0.0-5.0
Technological Sophistication			
# of uses of IT	4.8	1.8	2-10
# of uses of e-bus/Internet/Web	6.0	3.3	0-15
quality of IT security	4.4	0.9	2.0-6.0
Informational Sophistication			
# of accounting/fin./HRM apps	6.0	2.8	0-11
# of logistics/prod./distrib. apps	7.0	3.1	2-16
# of mark./sales/cust. serv. apps	3.8	2.0	0-7
	2.8	2.5	0-7
# of ERP system modules	3.6	0.8	1.4-4.6
information output quality	3.5	0.9	1.5-5.0
user-system interaction qual.			
IT Performance			
acc./fin./HRM app. benefits	3.3	1.2	0.0-5.0
log./prod./distrib. app. benefits	3.3	0.6	2.0-5.0
mark./sales/serv. app. benefits	3.0	1.1	0.0-5.0
e-bus./Net/Web app. benefits	2.6	0.7	0.0-4.1

[a]1: IT for control (n = 14)
2: IT for flexibility (n = 9)
3: IT for product and market development (n = 4)
4: IT for internal and external integration (n = 17)
[b]1: yes (n = 32)
0: no (n =12)
[c]1: supervised by the chief-executive (n = 24)
0: supervised by another manager (n = 20)

Structural equation modeling was used to validate the research model. To this effect, the PLS technique was chosen for its robustness, more precisely its capacity to handle small samples and formative measurement models in comparison to covariance structure analysis techniques such as Lisrel, EQS and Amos [19].

A. Measurement Model

Given their composite and multidimensional nature, the research constructs are modeled as being "formative" rather than "reflective" [9]. Such a construct is composed of many indicators that each captures a different aspect; hence changes in these indicators bring or "cause" change in their underlying construct [26]. IT management sophistication is thus modeled as a second-order formative construct from two sub-constructs, namely managerial sophistication and functional sophistication. As presented in Table II, each of these sub-constructs is in turn composed of six and two formative measures respectively, a functional sophistication and managerial sophistication score being obtained from the factor scores determined by a principal components analysis. Given that this analysis produced two components for managerial sophistication, a single score was obtained by averaging the two factor scores.

TABLE II. PRINCIPAL COMPONENTS ANALYSIS OF IT MANAGEMENT SOPHISTICATION

factor / indicator	Funct. Soph.	Manag. Soph.[a]	Manag. Soph.[b]
Functional Sophistication			
designated manager for IT	.91	-	-
org. level of the IT function	.91	-	-
Managerial Sophistication			
IT development	-	.79	-
IT evaluation	-	.68	-
user participation	-	.75	-
external consultants	-	.58	-
IT resources & competencies	-	-	.93
IT support & appropriation	-	-	.95

[a]IT management practices
[b]IT management capabilities

The reliability of a formative construct, as opposed to a reflective one, is confirmed by the absence of multicollinearity between its measures or indicators [30]. Formative indicator validity is confirmed by a weight that is significant and not less than 0.1 [22], as confirmed in Figure 3. Discriminant validity of a formative construct is confirmed by it sharing less than 50% variance with any other construct, whereas nomological validity is confirmed when the construct's hypothesized links with other constructs are significantly greater than zero and in the expected direction [1].

In similar fashion, IT usage sophistication is modeled and measured from two sub-constructs, namely

informational sophistication and technological sophistication. As presented in Table III, each sub-construct is in turn composed of six and three indicators respectively. The reliability and validity of the IT usage sophistication construct was similarly confirmed. As to the IT performance construct, it is composed of four measures, that is, the average benefits obtained from each type of IT-based application. One may note again that there is no multicollinearity among these last formative measures, the highest correlation among them being equal to 0.19 ($p > 0.1$), with all four regression weights being greater than 0.1 (see Figure 3), thus showing adequate reliability and validity.

TABLE III. PRINCIPAL COMPONENTS ANALYSIS OF IT USAGE SOPHISTICATION

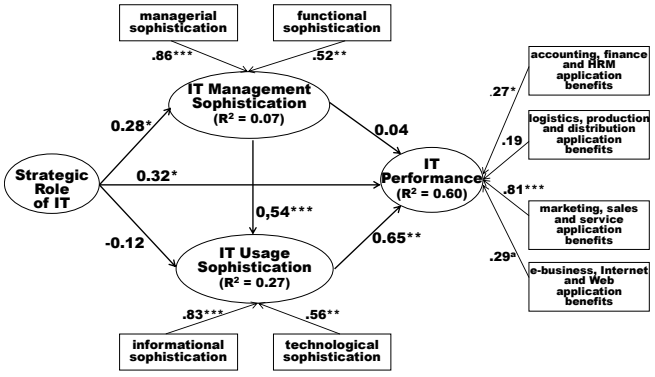
factor / indicator	Techn. Soph.	Inform. Soph.[a]	Inform. Soph.[b]
Technological Sophistication			
uses of IT	.90	-	-
uses of e-bus/Internet/Web	.81	-	-
quality of IT security	.50	-	-
Informational Sophistication			
accounting/fin./HRM apps	-	.78	-
logistics/prod./distrib. apps	-	.60	-
mark./sales/cust. serv. apps	-	.74	-
ERP system modules	-	.69	-
information output quality	-	-	.93
user-system interaction qual.	-	-	.95

[a]extensiveness of IT usage

[b]quality of IT usage

B. Test of the Research Model

The research hypotheses were tested by evaluation the direction, value, and level of significance of the path coefficients estimated by PLS, as presented in Figure 3.



Nota. Significance levels were obtained by bootstrapping. [a]p < 0.1 *: p < 0.05 **: p < 0.01 ***: p < 0.001

Figure 3. Results of testing the research model

A positive and significant path coefficient ($\beta_1 = 0.32$; $p < 0.05$) confirms the first research hypothesis, that is, the more strategic the role played by IT in the manufacturing SME, the greater its IT performance. Moreover, if one removes the effect of IT management and IT usage sophistication upon IT performance, the strategic role of IT still explains 25% of the variance in this same performance. The benefits obtained from marketing, customer service, and e-business applications thus flow directly from a vision of IT as a mean for the SME to develop its products and markets, to integrate its production processes, and to improve exchanges with its business partners.

A positive and significant path coefficient ($\beta_2 = 0.28$; $p < 0.05$) confirms the second hypothesis, that is, the more strategic the role played by IT, the greater the IT management sophistication of the SME. When IT constitutes a strategic necessity or a competitive weapon, when IT is of critical importance for "core" business processes of small manufacturers, these organizations act in a coherent manner by adopting managerial practices that allow them to better manage the development and use of these technologies. These are practices such as planning, designing and evaluating IT-based applications, sustaining and favoring user participation and user appropriation of IT, preserving and developing IT resources and competencies, and seeking outside consultants to overcome internal lacks in this regard. These firms show similar coherence when they place the IT function at a high hierarchical level in the organization and render IT autonomous (with a designated manager), that is, not subordinated to the financial or accounting function as is still often the case in small business.

Due to a negative and non significant path coefficient ($\beta_3 = -0.12$), the third hypothesis could not be confirmed. It stated that the more strategic the role played by IT, the greater the IT usage sophistication of the SME. Thus it seems that the strategic role of IT would be only indirect here, that is, through its effect on IT management sophistication. For instance, seeking internal and external integration of business processes through IT would lead the firm to better plan its use of IT and to dispose of better IT resources and competencies; only then could a more advanced technological infrastructure and applications such as ERP and e-business be implemented.

The fourth research hypothesis is confirmed by a positive and significant path coefficient ($\gamma_4 = 0.54$; $p < 0.001$), relating the firm's IT management sophistication to its IT usage sophistication. This result increases the relevance of a strategic perspective based on IT resources and competencies, namely a resource-based view to explain the level of adoption and assimilation of IT in manufacturing SMEs [3]. Now, firms that have sufficiently developed their IT function and managerial competence and that have access to external resources are those that have adopted and assimilated the greatest number of advanced manufacturing applications, and where system quality and security are best.

Due to a non significant path coefficient ($\beta_5 = 0.04$), the fifth hypothesis could not be confirmed. It stated that the greater the IT management sophistication, the greater the IT performance of the SME. In the absence of a direct effect, better management of IT has nonetheless an indirect effect upon IT performance, that is, through its positive effect on the use of TI (which in turn has a direct effect on performance, as we shall see). This last result is obtained with an estimation of this indirect effect by the product of the two path coefficients ($\beta_4 * \beta_6 = 0.54 * 0.65 = 0.35$; $p < 0.05$).

A strong path coefficient ($\beta_6 = 0.65$; $p < 0.001$) confirms the sixth research hypothesis, that is, the greater the small manufacturer's IT usage sophistication, the greater the performance of its information technology. Advanced applications such as an ERP system, a transactional Web site, videoconferencing, and mobile computing, to the extent that they are effectively assimilated by SMEs, are those that are the most strategic, that is, bring the greatest "value" to these firms in the form of increased competitiveness and competitive advantage. One may recall moreover that this increased assimilation of IT is the result of better management of these technologies. In turn, this better management is the result of a more strategic vision of the role played by IT in the organization.

In total, these three factors combined explain 60% of the variance in the performance of IT. One may note here that the applications that are most affected in terms of performance are the marketing and sales applications, followed by the accounting, finance and HRM applications, and the e-business, Internet and Web applications. This last result tends to underline the more operational rather than strategic nature of the logistics, production, and distribution applications as presently implemented in the sampled manufacturing SMEs.

VI. Discussion and Implications

The results obtained from 44 SMEs show that IT performance is influenced in two ways. First, IT performance is directly affected by the strategic role played by IT in the firm, and especially in the eyes of its leader. Second, IT performance is also influenced indirectly by the strategic role of IT through IT management sophistication, which in turn influences IT usage sophistication. It is then IT usage sophistication that directly contributes and contributes most to IT performance.

This dual contribution of the strategic role of IT to IT performance suggests that the functional sophistication of IT alone is not sufficient to increase IT performance; it is also necessary that IT be adequately used by employees. Thus, to ensure that information technology applications fully meet their strategic role, their development has to be adequately managed. The development and evaluation of these applications should take into account the needs of users, involving them when conducting business process analysis to make the most effective use of resources and competencies, this being done within a structured IT function which reflects the strategic and operational reality of the organization while procuring external resources when necessary.

The strategic role of IT has no direct influence on IT usage sophistication; however it does have an indirect effect through IT management sophistication. This means that that once information technologies are properly deployed, it is possible for users to enhance their strategic role. These results are in line with Westerman's [47] work on the evolution of IT, reiterating that IT should adequately support business operations, making certain that IT-based systems work as and when they are supposed to, that their access is secure, that the information output is accurate, complete and correct, and that all this is done in time and within budget. Users should then be able to learn and appropriate themselves of the various functional applications implemented by the firm, and to assess the quality of information output by these applications in order to make better decisions.

The descriptive results indicate that for all SMEs, the benefits of IT mainly come from accounting/finance/HRM, and logistics/production/ distribution applications. Then come benefits accruing from marketing/sales/customer service applications, and to a lesser extent e-business, Internet and Web applications. This descending order of benefits is consistent with the increasing complexity of the strategic role of IT. Most manufacturing SMEs do not use IT for purposes of internal and external integration of business processes, which is the most strategic role. The IT applications easiest to implement are often the first deployed, and therefore are the first to provide benefits.

In this study, where the benefits are cumulated by the type of applications used, firms that have deployed several types of applications are the ones showing the highest performance from their IT. They are also those who envision the most comprehensive strategic role for information technologies, the more complex and more demanding. In this context, manufacturing SMEs that gain more business value from IT are those that devote a more strategic role to these technologies, manage them in a more sophisticated way, and use them more extensively and intensively.

VII. Limitations and Conclusion

As in any empirical research, this study has some limitations that should be mentioned. Given the nature of the sample, its representativeness in relation to all SMEs limits the scope of the results. As sample firms have chosen to undertake an IT benchmarking exercise, they could differ from the general population in terms of strategic orientation, IT sophistication, and IT performance [6]. The use of perceptual measures for assessing the strategic role and performance of IT may also have induced some respondent cognitive biases, although earlier studies have also resorted to such measures [41].

Notwithstanding its limitations, this study revealed that a strategic vision of the role of IT is critical to the managerial and technological skills developed by manufacturing SME, and to the realization of IT business value from these capabilities. Based on a strategic alignment perspective, future studies could extend the research model by examining whether the role assigned to IT depends on how well it "fits" with the SME's business strategy, structure, and environment. A more complete formative model for measuring IT performance, such as that proposed by Gable, Sedera and Chan [18], could also be used to include, in addition to the organizational impacts, individual impacts, quality of IT-based systems, and quality of information output by these systems.

REFERENCES

[1] Andreev, P., Heatr, T., Maoz, H., and Pliskin, N. (2009), Validating formative partial least squares (PLS) models: Methodological review and empirical illustration, *Proceedings of the Thirtieth International Conference on Information Systems*, Phoenix, Arizona, pp. 1-17.

[2] Bergeron, F., Raymond, L., and Rivard, S. (2004), Ideal patterns of strategic fit and business performance, *Information & Management*, (41:8), pp. 1003-1020.

[3] Bharadwaj, A. (2000), A resource-based perspective on the information technology capability and firm performance: An empirical investigation, *MIS Quarterly*, (24:1), pp. 169-196.

[4] Blili, S. and Raymond, L. (1993), Information technology: Threats and opportunities for SMEs, *International Journal of Information Management*, (13:6), pp. 439-448.

[5] Brown, C.V. (1997), Redesigning the emergence of hybrid IS governance solutions: Evidence from a single case site, *Information Systems Research*, (8:1), pp. 69-94.

[6] Caldeira, M.M. and Ward, J.M. (2003), Using resource-based theory to interpret the successful adoption and use of information systems and technology in manufacturing small and medium-sized enterprises, *European Journal of Information Systems*, (12), pp. 127-141.

[7] Cassell, C., Nadin, S., and Gray, M.O. (2001), The use and effectiveness of benchmarking in SMEs, *Benchmarking: An International Journal*, (8:3), pp. 212-222.

[8] Chan, Y.E., Huff, S.L., Copeland, D.G., and Barclay, D.W. (1997), Business strategic orientation, information systems strategic orientation, and strategic alignment, *Information Systems Research*, (8:2), pp. 125-150.

[9] Chin, W.W. (1998), Issues and opinion on Structural Equation Modeling, *MIS Quarterly*, (22:1), pp. vii-xvi.

[10] Chwelos, P., Benbasat, I., and Dexter, A.S. (2001), Research report: Empirical test of an EDI adoption model, *Information Systems Research*, (12:3), pp. 304-321.

[11] Cragg, P.B. (2002), Benchmarking information technology practices in small firms, *European Journal of Information Systems*, (11:4), pp. 267-282.

[12] Cragg, P.B., Mills, A., and Suraweera, T. (2010), Understanding IT management in SMEs, *Electronic Journal of Information Systems Evaluation*, (13:1), pp. 27-34.

[13] Croteau, A.-M. and Bergeron, F. (2001), An information technology trilogy: Business strategy, technological deployment and organizational performance, *Journal of Strategic Information Systems*, (20:2), pp. 77-99.

[14] DeLone, W.H. and McLean, E.R. (1992), Information systems success: The quest for the dependent variable, *Information Systems Research*, (3), pp. 60-95.

[15] DeLone, W.H. and McLean, E.R. (2003), The DeLone and MacLean model of information systems success: a ten-year update, *Journal of Management Information Systems*, (19:4), pp. 9-30.

[16] Dibrell, C., Davis, P.S., and Craig, J. (2008), Fueling innovation through information technology in SMEs, *Journal of Small Business Management*, (46:2), pp. 203-218.

[17] Dierckx, M.A.F. and Stroeken, J.H.M. (1999), Information technology and innovation in small and medium-sized enterprises, *Technological Forecasting and Social Change*, (60), pp. 149-166.

[18] Gable, G.G., Sedera, D., and Chan, T. (2008), Re-conceptualizing information systems success: The IS-impact measurement model, *Journal of the Association for Information Systems*, (9:7), pp. 377-408.

[19] Gefen, D., Straub, D.W., and Boudreau, M.-C. (2000), Structural equation modeling and regression: guidelines for research practice, *Communications of the AIS*, (4:7), pp.1-76.

[20] Henderson, J.C. and Venkatraman, N. (1999), Strategic alignment: Leveraging information technology for transforming organizations, *IBM Systems Journal*, (38:2&3), pp. 472-484.

[21] Iacovou, C.L., Benbasat, I., and Dexter, A.S. (1995), Electronic data interchange and small organizations: Adoption and impact of technology, *MIS Quarterly*, (19:4), pp. 465-485.

[22] Jahner, S., Leimeister, J.M., Knebel, U., and Krcmar, H. (2008), A cross-cultural comparison of perceived strategic importance of RFID for CIOs in Germany and Italy, *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, IEEE Computer Society, pp. 1-10.

[23] Keen, P.G.W. (1993), *Shaping the future, business design through information technology*, Harvard Business School Press, Cambridge, Massachusetts.

[24] Kohli, R. and Grover, V. (2008), Business value of IT: An essay on expanding research directions to keep up with the times, *Journal of the Association for Information Systems*, (9:1), pp. 23-39.

[25] Kyobe, M. (2008), The influence of strategy-making types on IT alignment in SMEs, *Journal of Systems and Information Technology*, (10:1), pp. 28-35.

[26] MacKenzie, S.B., Podsakoff, P.M., and Jarvis, C.B. (2005), The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions, *Journal of Applied Psychology*, (90:4), pp. 710-730.

[27] Marbert, V.A., Soni, A., and Venkataramanan M.A. (2003), The impact of size on enterprise resource planning (ERP) implementation in the US manufacturing sector, *Omega*, (31), pp. 235-246.

[28] Myers, B. L., Kappelman, L.A., and Prybutok, V.R. (1998), A comprehensive model for assessing the quality and productivity of the information systems function: Toward a theory for information systems assessment, in Garrity, E.J. and Sanders, G.L. (Eds.) *Information Systems Success Measurement*, Idea Group, Hershey, Pennsylvania, pp. 94-121.

[29] Paré, G. and Sicotte, C. (2001), Information technology sophistication in health care: an instrument validation study among Canadian hospitals, *International Journal of Medical Informatics*, (63), pp. 205-223.

[30] Petter, S., Straub, D., and Rai, A. (2007), Specifying formative constructs in information systems research, *MIS Quarterly*, (31:3), pp. 623-656.

[31] Pflughoest K.A., Ramamurthy, K., Soofi, E.S., Yasai-Ardekani, M., and Zahedi, F. (2003), Multiple conceptualizations of small business Web use and benefit, *Decision Sciences*, (34:3), pp. 467-512.

[32] Pitt, L.F., Watson, R.T., and Kavan, C.B. (1995), Service quality: a measure of information systems effectiveness, *MIS Quarterly*, (19:2), pp. 173–185.

[33] Philip, G. and Booth M.E. (2001), A new six 'S' framework on the relationship between the role of information systems (IS) and competencies in 'IS' management, *Journal of Business Research*, (51), pp. 233-247.

[34] Powell, T.C. and Dent-Micallef, A. (1997), Information technology as competitive advantage: The role of human, business, and technology resources, *Strategic Management Journal*, (18:5), pp. 375-405.

[35] Premkumar, G. (2003), A meta-analysis of research on information technology implementation in small business, *Journal of Organizational Computing and Electronic Commerce*, (13:2), pp. 91-121.

[36] Rai, A., Tang, X., Brown, P., and Keil, M. (2006), Assimilation patterns in the use of electronic procurement innovations: A cluster analysis, *Information & Management*, (43), pp. 336-349.

[37] Raymond, L., Croteau, A.-M., and Bergeron, F. (2011), The strategic role of IT: An empirical study of its impact on IT performance in manufacturing SMEs, *2011 Sixth International Conference on Internet and Web Applications and Services (ICIW)*, St. Maarten, The Netherlands Antilles, pp. 89-97.

[38] Raymond, L., Paré, G., and Bergeron, F. (1995), Matching information technology and organizational structure: Implications for performance, *European Journal of Information Systems*, (4:1), pp. 3- 6.

[39] Riemenschneider, C.K. and Mykytyn, P.P. (2000), What small business executives have learned about managing information technology, *Information & Management* (37), pp. 257-269.

[40] Seddon, P.B., Graeser, V., and Willcocks, L.P. (2002), Measuring organizational IS effectiveness: An overview and update of senior management perspectives, *DATA BASE for Advances in Information Systems*, (33:2), pp. 11-28.

[41] Tallon, P.P., Kraemer, K.L., and Gurbaxani, V. (2000), Executives's perceptions of the business value of information technology: A process-oriented approach, *Journal of Management Information Systems*, (16:4), pp. 145-173

[42] Tornatsky, L.G. and Fleischer, M. (1990), *The processes of technological innovation*, Lexington, Massachusetts: Lexington Books.

[43] Uwizeyemungu, S. and Raymond, L. (2009), Exploring an alternative method of evaluating the effects of ERP: A multiple case study, *Journal of Information Technology*, (24:3), pp. 251-268.

[44] Venkatraman, N. (1989), The concept of fit in strategy research: Toward verbal and statistical correspondence, *Academy of Management Review*, (14:3), pp. 423-444.

[45] Venkatraman, N. (1994), IT-enabled business transformation: From automation to business scope redefinition, S*loan Management Review*, (35:2), pp. 73-87.

[46] Ward, J., Taylor, P., and Bond, P. (1996), Evaluation and realisation of IS/IT benefits: An empirical study of current practice, *European Journal of Information Systems*, (4:4), pp. 214-226.

[47] Westerman, G. (2009), IT risk as a language for alignment, *MIS Quarterly Executive*, (8:3), pp. 109-121.

Using Proximity Information between BitTorrent Peers: An Extensive Study of Effects on Internet Traffic Distribution

Peter Danielis, Jan Skodzik, Jens Rohrbeck, Vlado Altmann, Dirk Timmermann
University of Rostock
Institute of Applied Microelectronics and Computer Engineering
18051 Rostock, Germany
Tel./Fax: +49 (381) 498-7272 / -1187251
Email: peter.danielis@uni-rostock.de

Thomas Bahls
Ernst-Moritz-Arndt-University of Greifswald
Institute for Community Medicine
17487 Greifswald, Germany
Tel: +49 (3834) 86-7524
Email: Thomas.Bahls@uni-greifswald.de

Daniel Duchow
Nokia Siemens Networks GmbH & Co. KG
Broadband Access Division
17489 Greifswald, Germany
Tel: +49 (89) 5159-18237
Email: daniel.duchow@nsn.com

***Abstract*—Peer-to-Peer file sharing generates by far the most Internet traffic reaching up to 70 % in some regions of the world. These data volumes pose a significant challenge to Internet Service Providers regarding traffic engineering. Because Peer-to-Peer routing is usually agnostic of the underlying topology, traffic engineering abilities of Internet Service Providers are inhibited and their core networks are overburdened with Peer-to-Peer data. To disburden Internet Service Providers' core networks, a new algorithm for the BitTorrent protocol is proposed in order to improve peer selection. BitTorrent users are provided with accurate information on the hop counts to other BitTorrent users to select physically proximate users. Thereby, the initial Time-To-Live value of outgoing IP packets is copied and inserted as part of the BitTorrent payload. At the packet's destination, the hop count is calculated as the difference between the copied Time-To-Live value and the Time-To-Live value of the IP header. Simulation results for standard and modified BitTorrent implementation are presented highlighting beneficial effects on both the traffic of Internet Service Providers' core networks and BitTorrent users' download performance. The extensive exploration of impacts on Internet traffic includes the observation of traffic loads and peaks in the core networks as well as the number of dropped packets due to congestions. Moreover, as realistic simulations require the consideration of the BitTorrent users' behavior of continuously leaving and entering the BitTorrent network, this article elaborates on the dynamic nature of peers.**

Keywords-Peer-to-Peer; Topology Awareness; BitTorrent; Dynamic User Behavior; Traffic Analysis.

I. Introduction

During the last years, several new flavors of private and commercial use of the Internet have developed. One of these flavors are Peer-to-Peer (P2P) networks, clients, tools, and communities [1]. Today, Internet traffic is dominated by P2P data ranging from 43 % up to 70 % in different regions of the world. This is mainly caused by file sharing applications such as eMule or BitTorrent (BT). Particularly, BT traffic accounts for up to 80 % of P2P traffic, i.e., 56 % of overall Internet traffic [2]. However, there are numerous other application areas for P2P, e.g., P2P-based IPTV like Joost or Zattoo [3], [4].

On the one hand, Internet Service Providers (ISPs) benefit from the P2P hype through an increase of their operating income. This is because P2P applications are one of the main reasons for Internet users to subscribe for a broadband connection [5]. But, on the other hand, the high P2P data volumes pose a significant traffic engineering challenge. Traffic engineering denotes the process of managing traffic flows through the network [6]. This discrepancy between operating income and traffic engineering challenge puts network operators and ISPs into a difficult situation. Other traffic like HTTP is choked down because the network infrastructure is overburdened with P2P data. The main reason is that routing within logical P2P networks does not take the underlying physical Internet topology into account [7]. Usually, an unstructured P2P network overlay—on which this article focuses—is constructed by choosing random peers [8]. Due to this arbitrary procedure, neighborship on the P2P overlay does *not* implicate proximity on the underlying Internet topology at all. This problem is usually denoted as topology mismatching problem between P2P overlays and physical network infrastructures [9]. Thus, two communicating P2P neighbors may be physically far away from each other although the desired content is often available on a physically more proximate peer as well [10]. Communication with physically distant peers uses long data paths, e.g., regarding the hop count. This consumes more bandwidth, which is highly inefficient when the load of the network is already high. It can therefore cause traffic congestions [6]. Congestions are neither good for ISPs nor for users. In contrast, communication with proximate peers consumes less bandwidth and traffic in the core network diminishes. ISPs benefit from a traffic reduction in their core networks as more bandwidth is available for other applications. Moreover, traffic congestions can be reduced as well.

Consequently, this contribution was motivated by the idea of disburdening ISPs' core networks by reducing the physical path lengths, i.e., the hop count. A new algorithm for the BT protocol to use hop count as additional selection criterion for peers (besides their download performance) has been developed. However, the hop count for determining proximity is not part of packets. Therefore, a new approach to provide P2P users—*BT users in particular*—with information on physical hop counts to other BT users is proposed as well. The standard BT implementation is modified such that the initial Time-To-Live value (TTL) of outgoing IP packets is inserted as part of the BT payload. At the packet's destination, the modified BT algorithm calculates the hop count from the inserted initial TTL and the received IP packet's TTL.

However, a BT user primarily wants to download desired content as fast as possible. He is usually not aware of or even not interested in the underlying transport mechanism. Thus, a user would not select the most proximate BT user among all users, which provide the desired content with nearly the same upload speed. However, BT users are assumed to be cooperative by selecting close-by users unless they do not suffer from this decision regarding their performance. Performance is defined as the time required to get desired content and is denoted as the users' Quality of Experience (QoE) in the following. In the best case, it is beneficial for users and ISPs. In the worst case, the performance shall be equal to or not considerably lower than the standard case when not using the hop count.

Briefly summarized, the main contributions of this article are the following:

- Investigations are carried out on how to calculate the hop count and provide it for BT users.
- Improved peer selection for BT is described, where hop count information is used by a modified BT algorithm to select close-by users.
- The integration of the BT protocol in the network simulator ns-2 is detailed as well as the incorporation of the dynamic nature of BT peers.
- Simulation results for standard and modified BT algorithm are presented and (dis-)advantages for both ISPs and users are discussed.

The remainder of this article is organized as follows: Section II contains a comparison of the proposed approach with related work. Section III explains the computation of the hop count from the TTL and how to provide the initial TTL. Section IV addresses improved peer selection for BT. Section V motivates the use of the network simulator ns-2, elaborates on the dynamic nature of BT peers, and shows impacts of improved peer selection on the traffic in an ISP's core network and BT users' performance. The article concludes in Section VI.

II. Comparison with Related Work

The mismatch between logical P2P overlay and the underlying physical topology becomes a serious obstacle for the development of P2P systems. Being aware of the mismatching problem, much scientific literature can be found, which addresses the locality problem in Distributed Hash Table (DHT)-based P2P networks, i.e., structured P2P networks [11], [12]. In structured P2P networks, all peers are organized into an identifier ring and an association between content and location where it is stored is given. Basically, there are three approaches, which have been proposed to exploit network proximity in DHTs. For a detailed discussion of the three approaches, the interested reader is referred to [13].

However, not only structured but also unstructured P2P networks suffer from the mismatching between the logical overlay and the underlying physical topology. Hence, this article focuses on selecting close-by peers in unstructured P2P networks. In contrast to structured P2P networks, in unstructured P2P networks like BT, peers are organized arbitrarily. Thus, neighborship on the P2P overlay does *not* implicate physical proximity on the underlying Internet topology at all.

Many approaches, e.g., [14] and [15], to *construct* unstructured topology-aware overlay networks do exist. These approaches improve performance significantly and avoid unnecessary traffic by exploiting network proximity. However, they require adding structure to unstructured P2P networks following physical network characteristics. Furthermore, traffic overhead is created for maintaining this structure. In contrast to these approaches, the authors' approach does not intervene with the *construction* of unstructured P2P networks. Instead, the BT protocol is slightly modified to provide the hop count and select proximate peers by means of a new BT choking algorithm. Thereby, no modification of the construction algorithm is necessary and no additional packets have to be sent to determine the distance, i.e., the hop count between peers.

Also, there are approaches to *shape* P2P traffic in a more efficient way *with* network support. The IETF has planned to develop a protocol for Application-Layer Traffic Optimization (ALTO). By means of an ALTO server, peers can obtain information "to perform better-than-random initial peer selection" [16]. The Portal for P2P Applications (P4P) project aims at allowing more effective cooperate traffic control between P2P applications and ISPs via dedicated trackers to localize P2P traffic [17]. Moreover, in the course of the Network-Aware P2P-TV Application over Wise Networks (NAPA-WINE) project, the impacts on the underlying transport network shall be minimized when distributing P2P-IPTV datastreams [18]. Thereby, a so-called "network-peer can perform actions to optimize the P2P overlay taking into account the status of the transport network". Liu et al.

[19] suggests to use autonomous system (AS) hop count to achieve locality-awareness in BT-like P2P applications. However, the tracker is required to know the Internet topology and peers have to obtain dynamic distance information by P4P or content distribution networks. As opposed to these approaches, the approach of this article is completely self-sufficient and does not require network support, but does the necessary modifications solely in the BT application.

III. Computing Hop Count for BT

Since hop count information is neither stored in the IP header nor elsewhere, it is necessary to compute it. There are two methods for hop count determination [20]. Either it can be actively measured or it can be calculated by using the TTL. For active measurement, ICMP ECHO packets are used. Although this method mostly results in an accurate hop count, applying it to many hosts in current and prospective P2P scenarios is impractical since enormous traffic overhead is created. Thus, as measuring method in the envisaged P2P use case, it is not favorable to use active measurement. Contrary, the calculation of hop count simply means subtracting the final TTL of a received IP packet from its initial TTL. This is optimal for computing hop counts of many hosts as no extra packets have to be sent. Consequently, this approach is chosen for calculating the hop count. However, the problem regarding hop count calculation is that the initial TTL is not available in IP packets. Thus, it must be made available.

The TTL is part of the IPv4 (further referred to as IP) header [21]. It is used as hop counter. Thus, each router processing a packet decrements the TTL by one. To calculate the hop count from TTL, the initial TTL of an outgoing IP packet is required. Then, this value can be subtracted from the final TTL of the IP header at the packet's destination to get the hop count. As shown in [22], due to the heterogeneity of the Internet, there is *no* unique initial TTL. The initial TTL depends on the operating system (OS).

Consequently, the question is: How to provide the initial TTL? Therefore, a modified BT algorithm is proposed, which inserts the initial TTL into BT messages.

A. Standard (tracker-based) BT algorithm

BT is currently the most widespread P2P network for file sharing. It accounts for about 56 % percent of the overall Internet traffic [2]. Its popularity is mostly due to its high download rates—the main interest of users. BT focuses on high speed rather than on search capabilities.

For each shared file in BT, an own P2P net is created. To search for a file, usually a web site is contacted to get a *.torrent* file. This file contains, among other things, the address of a *tracker* and information about the file to be downloaded. The tracker is contacted to get a list of BT users holding the file (or parts of it—so-called *pieces*, which are further subdivided into sub-pieces). Thereby, all BT users, which are interested in this file, form a so-called *swarm*. Complete downloaders serving the whole files are called *seeds*. Incomplete downloaders are called *leechers*.

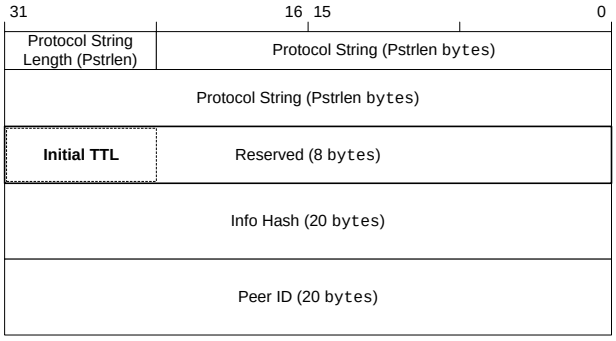


Figure 1. Composition of a BT Handshake message.

BT users start to download pieces in *random* order and change to *rarest first* order after the first piece is completed [23]. Thereby, they follow the *strict priority* of solely requesting sub-pieces of a particular piece before sub-pieces from the next piece. Furthermore, for a quick finish of a download, BT applies the so-called *endgame mode*, which is enabled when a download is almost complete. Then, a BT user simultaneously requests all missing sub-pieces from all BT users he knows.

For selecting a user who may download a piece, BT applies the so-called *choking algorithm* [23]. Put in a nutshell, this is a variant of the tit-for-tat strategy. Only users offering sufficient upload performance are given download in return (they are *unchoked*). The choking algorithm to determine a user that may download pieces is executed periodically because upload performance of users can change quickly. As an exception, each BT user has an *optimistic unchoke* available to unchoke one other user regardless of his upload performance. Furthermore, BT applies an *anti-snubbing* policy as part of the choking algorithm. Thereby, BT users, which have not given a single piece within one minute, are not unchoked except for optimistic unchokes.

Once a BT user has finished his download, he may decide to stay in the network for a while (lingering). During this time span, he only uploads pieces preferring users, to which he has the best upload rates.

B. Including the initial TTL in BT messages

To introduce as little overhead as possible into the BT algorithm and BT traffic, the initial TTL is only included into necessary BT messages. Two types of messages can be distinguished in BT flows: the tracker requests and responses and the messages between BT users. Thereby, messages between BT users are solely exchanged via TCP sockets. One message necessary for BT user interaction is the *Handshake* message (see Figure 1) [24].

A BT Handshake is sent by the initiator of a connection

between two users of a swarm. In return, the recipient of the Handshake message has to respond with a Handshake message himself. In both the Handshake message and the response to it, the modified BT algorithm directly inserts the initial TTL. The authors suggest to use the first byte of the eight *Reserved* bytes since these bytes can be used to change the BT protocol behavior.

C. Providing the TTL in the BT application

As TCP sockets are used for communication, IP header fields like the TTL are not available in applications per se. Therefore, following two questions are answered to clarify how to make it available:

1) How to get the initial TTL of outgoing BT Handshakes? The initial TTL is retrieved via the BSD sockets compatible function *getsockopt*, which is both Windows and Unix-compatible. Thereby, the specific socket option *IP_TTL* is used, which is supported by TCP sockets for outgoing packets.

2) How to get the final TTL of incoming BT Handshakes? Since TCP sockets do not offer support for providing the TTL of incoming packets, the pcap (packet capture) library is used *additionally*. Unix-like OS's apply the pcap implementation *libpcap*, whereby Windows OS's use a port of libpcap called *WinPcap*. libpcap provides a high degree of portability as it supports numerous OS's [25]. Using filters like the Berkeley Packet Filter, already the kernel can be instructed to copy only packets, which match the composition of a BT Handshake, to the BT application. Thus, the kernel buffer is not overfilled with packets, which could lead to high packet loss. Possible alternatives are *raw sockets* offering direct access to the network layer as well. However, raw sockets have turned out to be buggy, unportable, and may not be able to capture TCP packets at all (depending on the OS) and are thus not feasible.

IV. Improved Peer Selection for BitTorrent

The modification of BT for improved peer selection concerns BT's choking algorithm. The standard choking algorithm selects BT users that may download pieces solely depending on their offered upload performance (except *optimistic unchokes*). Users are ranked in an unchoke list such that the user with the highest upload performance (i.e., the highest *service rate*) is on top. Usually, a fixed number of users on top of the list (e.g., 4) may download concurrently. In the modified version, users are no longer ranked only depending on their service rate. Instead, for two users A and B *in another user's unchoke list*, quotients for those users A and B are calculated as

$$Quotient = Service\ Rate/Hop\ Count.$$

The quotient (denoted as *Quot* in the algorithm depicted in Figure 2) determines a user's rank in the unchoke list. If A's quotient is greater than B's, i.e., condition 1 = true

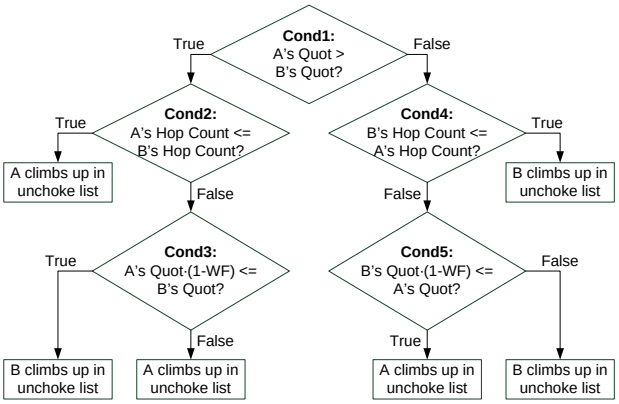


Figure 2. New BT choking algorithm for calculating a user's position in the unchoke list. The hop count is used as additional selection criterion.

(denoted as *Cond1* in Figure 2) and A's hop count is smaller than or equal to B's (Cond2 = True), A climbs up in the unchoke list. However, if A's hop count is greater than B's (Cond2 = False), A's quotient is multiplied (i.e., weighted) by a variable factor (*1 - hop count weighting factor* (WF)) with WF values ranging from 0 to 1. If A's weighted quotient is smaller than or equal to B's quotient (Cond3 = True), B climbs up in the unchoke list because B's hop count is weighted more than A's service rate. Otherwise (Cond3 = False), A climbs up because A's service rate is weighted more than B's hop count. In the else-branch of the algorithm, the analog conditions for B's quotient being greater than A's are stated.

As an exemplification for the algorithm, let us assume a user A with high service rate and high hop count and a user B with moderate service rate but very low hop count. Following the calculation rule for a user's quotient, A is assigned a relatively low quotient compared to B's quotient regardless of A's high service rate. Still, A's quotient be greater than B's quotient in this example (Cond1 = True) although B's hop count be significantly smaller than A's hop count (Cond2 = False). However, B's hop count can be given an even higher weight by assigning an appropriate, i.e., high WF value to let B climb up in the unchoke list (Cond3 = True).

To summarize, WF is used to make a compromise regarding the weighting of a user's service rate and his hop count. A high WF value results in BT users with low hop counts on top of the unchoke list almost regardless of their service rate. Vice versa, a low WF value leads to a higher weight of a user's service rate. Thus, users with high service rate are put more probably on top of the unchoke list nearly irrespective of their hop count.

One approach for the selection of close-by BT users is to let the user decide directly. The alternative approach followed in this article is to integrate an automatic selection mechanism into the BT algorithm.

V. Evaluation of Standard and Modified BT Algorithm

In this section, the choice of the network simulator ns-2 for the evaluation of the authors' proposed approach is motivated. Then, the integration of the BT protocol into ns-2 is described including an elaboration of the dynamic nature of BT users. Finally, simulation results are shown using the network simulator ns-2. Results include a comparison of standard with modified BT algorithm in terms of the number of hops between users, the load of the core network, the number of dropped packets, the maximum throughput in the whole network, and users' QoE.

A. Determination of an appropriate simulator

It is essential to choose an appropriate simulator to achieve realistic and trustful simulation results. There is a wide range of network simulators, especially in the field of P2P simulations. Table I lists the most significant open-source simulators, which could be used for the evaluation of standard and modified BT algorithm and compares them regarding scalability, license, and OS. Most of the simulators run under Linux/Unix or are implemented in Java and are thus platform independent. Node numbers within the range of 10^4 up to 10^6 are supported and hence, simulations of large P2P networks are possible.

Table I
Simulators and their properties [26], [27], [28], [29], [30].

Simulator	Scalability	License	OS
ns-2	$200 - 10^4$	GNU GPL	Linux/Unix
OverSim	10^5	GNU GPL	Linux, Mac, Windows
PlanetSim	10^5	GNU LGPL	Platform independent
PeerfactSim.KOM	$10^5 - 10^6$	GNU GPL	Platform independent
PeerSim	10^6	GNU LGPL	Platform independent

In the following, the choice of ns-2 is motivated by giving an overview of its advantages regarding the consideration of the physical underlay.

1) Advantages of ns-2: Most simulators do not accurately take into account the physical underlay, e.g., to speed up simulations [31].

OverSim supports three different types of underlays, which are the simple, single host, and INET underlay [27]. The INET framework, for example, offers the opportunity of modeling access and backbone networks based on a complete IP protocol stack. However, the accuracy of the simulator is not known and thus, realistic conditions cannot be assumed without further ado. Planetsim is an overlay network simulation framework, which solely provides simple networks like RingNetwork or CircularNetwork that do not consider latency costs [28]. Therefore, it does not provide sufficiently realistic network conditions. The same applies to PeerfactSim.KOM, which uses mathematical and stochastic models to emulate the behavior of a network and does not take into account the physical underlay [29]. PeerSim contains, among other things, an event-based engine, which supports transport layer simulations [30]. Thereby, the transport layer is represented by a special protocol that provides a message sending service and therefore, realistic conditions cannot be assumed without further ado as well.

On the contrary, ns-2 is a complex simulator attempting to model the whole network stack using real network components like network nodes and routers to construct the network [32]. Therefore, an explicit correlation between the logical P2P overlay and physical underlay is given. Thus, it is possible to examine the impact of improved peer selection on the underlay, which is necessary for the evaluation of the authors' proposed approach.

Furthermore, ns-2 contains full implementations of standard protocols like TCP, UDP, and FTP. These protocols are verified and run through continuous validations [33]. Therefore, these protocols are available for the implementation of the BT protocol to the full extent. This is an essential aspect as BT uses TCP for data transmission through the network.

Last but not least, ns-2 is well documented, which alleviates the development and evaluation of the modified BT algorithm [26].

In the authors' opinion, ns-2 is best suited for the envisaged simulations due to the given reasons and has therefore been chosen.

2) ns-2's principle of operation: There are two possibilities to set up and influence simulation runs in ns-2.

First of all, complex networks with hundreds to thousands of nodes and their connections can be defined in a TCL script. Thereby, the physical underlay (network topology) for the BT network is created. Following basic parameters for the simulation can be adjusted:

- Number of nodes
- Topology of network
- Bandwidth of the connection between nodes
- File size to be distributed

Furthermore, trace files can be declared, which record the whole network traffic during the simulation.

Table II
Recorded data in a trace file [26]

Event	Description
+	Packet has been prepared and put into the waiting queue to be sent
-	Packet has been transmitted to destination
h	Packet traveled a hop by passing a node
d	Packet has been dropped

The second and most powerful possibility is to modify the protocols to be simulated directly in their C implementation. Additionally, new protocols can be created and

basic elements like nodes, waiting queues, or links may be redefined. The BT protocol modifications concerns BT's choking algorithm as described in Section IV. When the TCL script starts running, trace files are generated and record the network communication between nodes. Recorded data consists of events (important events are listed in Table II) and serves as the basis of the statistics creation.

An example for traced data is shown in Table III. Thereby, T denotes simulated time in seconds. SOURCE gives the source node from where the data packet is sent and SINK specifies the destination node of the packet. TYPE provides information on the packet type; SIZE states the packet size in bytes.

Table III
EXAMPLE DATA IN A TRACE FILE [26]

EVENT	T [s]	SOURCE	SINK	TYPE	SIZE [Byte]
+	0.000233	12	6	tcp	40
-	0.000234	12	6	tcp	40
h	0.000312	12	10	tcp	40
d	0.000333	3	7	tcp	40

Traced data is finally processed by Perl scripts and sorting algorithms to create desired statistics. The processing has been automated to be able to handle the large amounts of simulation data (several tens of GB per simulation run) efficiently.

B. BitTorrent in ns-2

A BT patch from [34] has been used to implement the BT algorithm in ns-2. This implementation contains almost the full functionality of the BT protocol (see Section III-A) as follows:

- Partitioning of a file into pieces
- Strict priority download strategy
- Random and rarest first download strategy
- Choking algorithm
- Optimistic unchokes
- Upload only: Leave options of peers, which have completed their file download

The endgame mode and the anti-snubbing strategy have not been implemented. However, as the focus is on data transmission efficiency, this simplification is tolerable.

The BT algorithm has been complemented by the dynamic nature of peers, i.e., the BT users' behavior of continuously leaving and entering the BT network. According to [35], BT users follow a Weibull distribution when arriving at the BT network (inter-arrival time) (see Figure 3). A Weibull distribution follows the cumulative distribution function $Weibull(x, \lambda, k) = 1 - e^{-(x/\lambda)^k}$ with scale parameter λ and shape parameter k. For better comparability with [35], the complementary cumulative distribution function (CCDF) $1 - Weibull(x, \lambda, k)$ of the respective distribution is used.

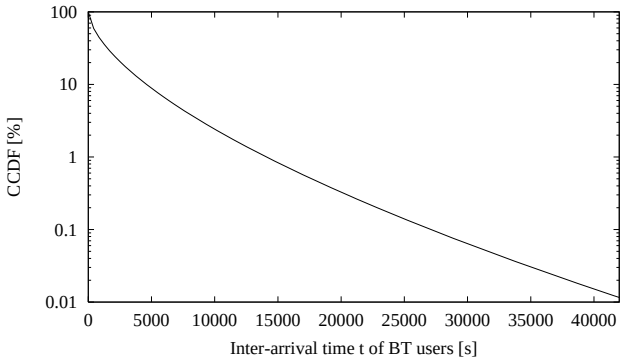


Figure 3. Complementary cumulative Weibull(t, 1200, 0.62) distribution function for the inter-arrival time of BT users. For the y-axis, a base 10 logarithmic scale is used.

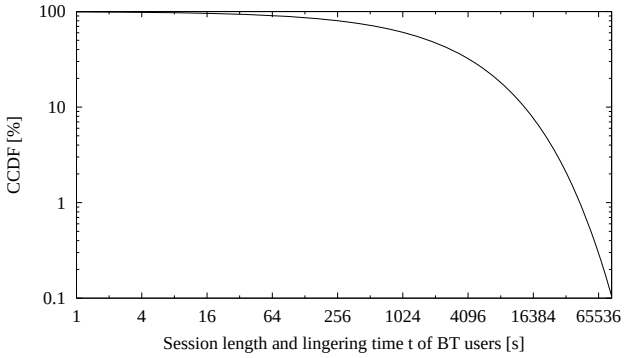


Figure 4. Complementary cumulative Weibull(t, 3302.977, 0.59) distribution function for the session length and lingering time of BT users. For the x-axis, a base 2 logarithmic scale is used; for the y-axis, a base 10 logarithmic scale is used.

Session lengths of BT users, i.e., the time how long they stay in the BT network each time they appear are implemented to follow a Weibull distribution as well (see Figure 4). When BT users leave the BT network after a session, they return after a uniformly distributed time (downtime). For clarity reasons, the uniform distribution is not depicted. Finally, BT users stay in the BT network for a while after the completion of a download (lingering time). The lingering time is modeled with the same Weibull distribution like the session lengths (see Figure 4).

Table IV
ASPECTS OF THE DYNAMIC NATURE OF BT USERS [35]

BT USER BEHAVIOR	DISTRIBUTION	λ	K	T_{max}
Inter-arrival time	Weibull	1200	0.62	700 min
Session length	Weibull	3302.977	0.59	-
Downtime	Uniform	-	-	0.5 d
Lingering time	Weibull	3302.977	0.59	-

Parameter values of the distribution functions are listed in Table IV. Thereby, scale parameter λ and shape parameter k have been set to values, which are reasonable for the

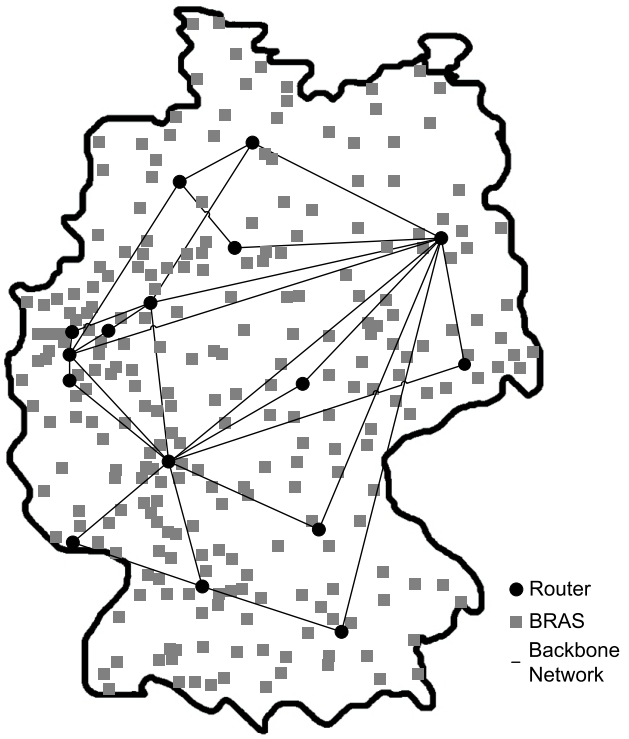


Figure 5. Telefonica's network infrastructure in Germany [36]

simulation of downloading a file of 100 MB in size by 200 users. Moreover, T_{max} has been introduced for the distribution functions of inter-arrival time and downtime to properly restrict their codomains for the simulation.

C. Simulation setup

For the simulation, a topology has been developed, which maps Telefonica's backbone network in Germany [36]. Telefonica possesses one of the most capacious network infrastructures in Germany. The schematic layout of the developed topology is depicted in Figure 5 and consists of

- a backbone (core) network of routers,
- Broadband Remote Access Servers (BRAS's) that are connected to routers, which are linked to DSL Access Multiplexers (DSLAMs),
- and BT users, which are connected to DSLAMs.

In accordance with Telefonica's network infrastructure in Germany, the developed topology comprises 16 routers. The number of BRAS's is set to 4 per router (resulting in 64 BRAS's) and there are 6 DSLAMs per BRAS (resulting in 384 DSLAMs). The number of BT users is set to 200 for the simulations. The routers form a static structure like the one apparent in Figure 5. BRAS's are uniformly distributed around routers and in the same way, DSLAMs are uniformly distributed around BRAS's. Users are randomly connected to DSLAMs. The number of BRAS's, DSLAMs, and users is fixed for all simulations. The bandwidth between routers and between routers and BRAS's has been set to 20 Gbit/s. The bandwidth between DSLAMs and BRAS's is set to 1 Gbit/s. These bandwidths values are common values in practice. Each BT user is assigned a download capacity of 6 Mbit and an upload capacity of 1.5 Mbit, which are reasonable values for an asymmetric Internet access.

At the start of each simulation run, one BT user is a seeder. This seeder stays in the network until each BT user of the swarm has finished his download of a file of 100 MB in size. The maximum number of users that may download concurrently from another user is set to 4.

D. Simulation results

Both standard and modified BT algorithm (BTA) have been simulated on the developed topology. In the simulations, the following values have been determined for varying WF values to compare both algorithms:

- Number of physical hops: Summed up number of physical hops that data has to travel through the network during the simulation.
- Data volume in the core network: Summed up data volume passing the routers of the core network during the simulation.
- Number of packet drops: Absolute number of packet drops due to congestions.
- Maximum throughput in the whole network: Traffic peak in the whole network and its point in time.
- Time necessary for the last user to finish a file download: Time needed until the last BT user has finished the file download during the simulation.

As users are connected to DSLAMs randomly, for each WF value on the x-axis, 10 measurements have been taken. In the diagrams, the mean value of those 10 measurements is depicted on the y-axis for each WF value. For the modified BT, the 95 % confidence interval (CI) is depicted to demonstrate that the measurements' precision is sufficient to draw conclusions.

The calculated mean value for standard BT is independent from WF values and thus constant. Therefore, the CI is not charted for standard BT in the given figures:

- CI for the number of physical hops: 950842
- CI for the data volume in the core network: 214 Mbit
- CI for the number of packet drops: 35785
- CI for the maximum throughput in the whole network: 2 Mbit/s and for its point in time: 29 s
- CI for the time necessary for the last user to finish a file download: 46.5 min

Number of physical hops: As apparent from Figure 6, the number of hops decreases when applying the modified BTA. For WF = 0, the simulation results show a minor increase of the number of hops by 0.02 % compared to the standard BTA. This is due to the fact that hop count is considered by the algorithm in Figure 2 but the users' service rate

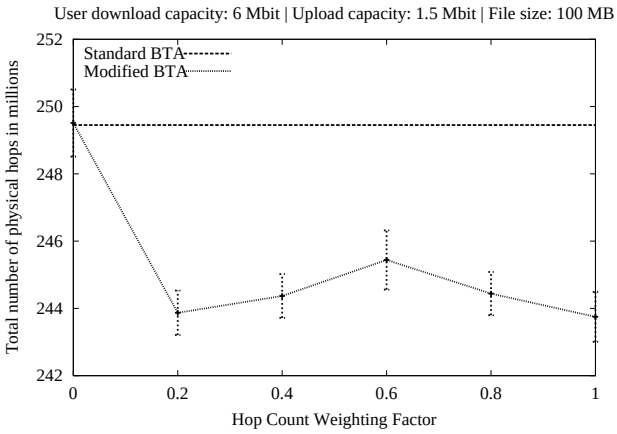


Figure 6. Number of hops for varying WF values. The result for standard BT is independent from WF values and therefore constant.

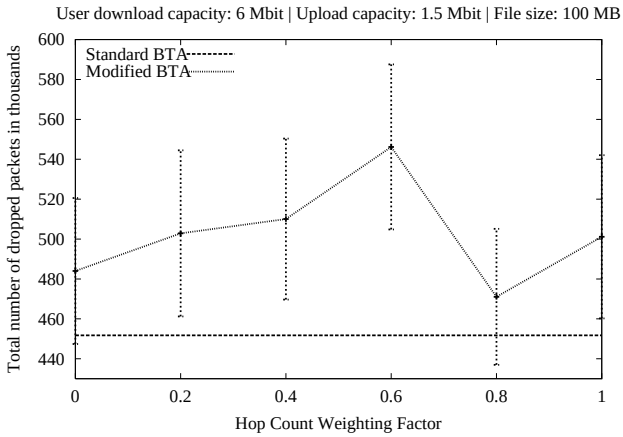


Figure 7. Number of dropped packets for varying WF values. The result for standard BT is independent from WF values and therefore constant.

Table V
DATA VOLUME IN THE CORE NETWORK FOR VARYING WF VALUES. THE RESULT FOR STANDARD BT IS INDEPENDENT FROM WF VALUES AND THEREFORE CONSTANT.

WF	Standard BTA	Modified BTA		
	Data volume [Gbit]	Data volume [Gbit]	Data volume reduction [%]	CI [Mbit]
0	14.43	14.22	1.5	240
0.2		12.13	15.9	175
0.4		12.29	14.8	168
0.6		12.23	15.2	179
0.8		13.01	9.8	291
1.0		11.98	17.0	142

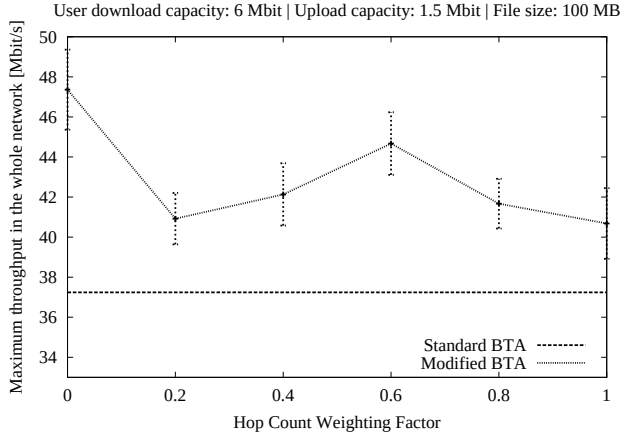


Figure 8. Maximum throughput in the whole network for varying WF values. The result for standard BT is independent from WF values and therefore constant.

dominates as peer selection criterion. Thereby, the degrees of freedom are limited and the number of hops increases. Please remember, that hop count is *always* considered by the modified BTA and an increasing WF value solely *boosts* the hop count's influence. For each other WF value except WF = 0, the total number of hops decreases. In fact, for WF = 1, the highest reduction of 2.3 % has been achieved.

Data volume in the core network: This relatively slight decrease in the number of hops results in a significant lower load of the core network for the modified BT variant. Table V illustrates this fact, showing a reduction of the core load by 17 % for WF = 1.

Number of packet drops: In Figure 7, the number of dropped packets is shown. The number of dropped packets increases in case of the modified BT algorithm with an increase ranging from 7 to 21 %. However, the load in the core network is lower as traffic is pushed away from the core network to the network's edges. Thereby, packets may be dropped at BRAS's due to high data volumes exchanged by close-by users.

Maximum throughput in the whole network: In Figure 8, the maximum throughput in the whole network is depicted. It is considerably higher in case of the modified BT algorithm with an increase ranging from 9 % for WF = 1 to 27 % for WF = 0. Thereby, the users' upload service rate dominates as peer selection criterion for WF = 0 allowing for the highest increase in terms of maximum throughput still benefiting from traffic localization compared to the standard BTA. Thus, the modified BTA enables an increased maximum throughput allowing for a higher utilization of available bandwidth and a better exploitation of the network infrastructure's capabilities, respectively.

Thereby, the point in time of the traffic peak is between minute 15 and 17 after simulation start and shows only minimal differences (between 2 and 7 %) between both algorithms (see Figure 9). At this point in time, the majority of BT users (approximately 60 %) has entered the network as the inter-arrival time has been modeled by a Weibull(t, 1200, 0.62) distribution function (see Figure 3 and Table IV).

Time necessary for the last user to finish a file download: Furthermore, the simulations show that the time

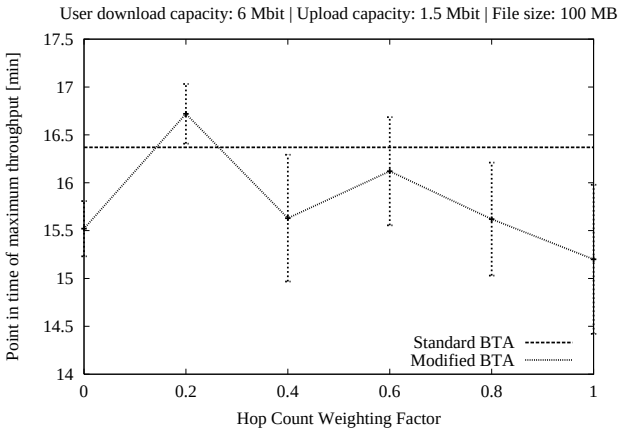


Figure 9. Point in time of maximum throughput for varying WF values. The result for standard BT is independent from WF values and therefore constant.

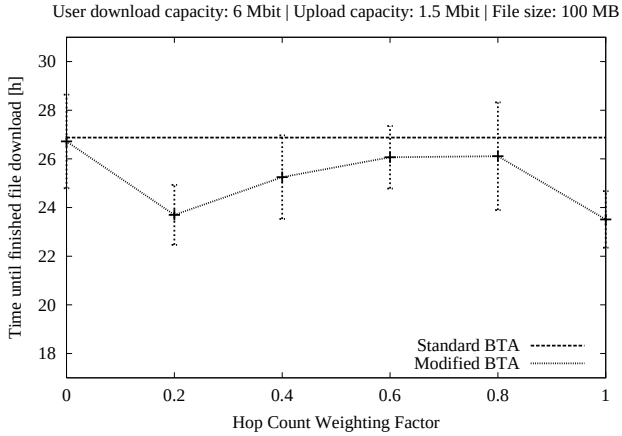


Figure 10. Time until the last BT user has downloaded the file for varying WF values. The result for standard BT is independent from WF values and therefore constant.

necessary until the last BT user has downloaded the complete file is considerably lower if the modified BTA is applied (see Figure 10). In fact, time is even decreased by up to 13 % for WF = 1. This decrease of time involves a reduction of the core load by 17 % and a decrease of the number of hops by 2.3 %. Thus, the most favorable WF value regarding BT users' QoE and the core network's load is obviously WF = 1, which weights the hop count highest.

VI. Conclusion

This article proposes a new choking algorithm for the BT protocol to preferably select physically close-by BT users in order to disburden ISPs' core networks. The selection criterion is the hop count. It is calculated from the difference of the initial TTL value of a packet's IP header and the TTL value at the packet's destination. As the initial TTL is not directly available, it is inserted into BT Handshake messages by the modified BT algorithm. Thereby, the modified BT protocol does neither require a modification of the construction algorithm for the BT network nor does it require sending additional packets to determine the hop count.

The simulations carried out for the BT algorithm clearly show that ISPs benefit from a modified BT using the hop count as additional selection criterion for download partners. The load of the ISP's core network is alleviated by up to 17 %. Thereby, traffic is localized (the number of hops is reduced by up to 2.3 %). By pushing traffic away from the core network to the network's edges, packets may be dropped there. Due to high data volumes exchanged by close-by users, this results in an increase of dropped packets ranging from 7 to 21 %. However, this is tolerable as the focus is on a decreased load in the core network to free bandwidth for traffic caused by other applications. The modified BTA enables an increased maximum throughput in the whole network of 9 to 27 % allowing for a higher utilization of available bandwidth. Moreover, the simulation show that users' QoE increases as time for the last BT user to finish a download is decreased by up to 13 %.

The utilization of hop count for the consideration of physical proximity is a generic approach. Future work will therefore focus on providing the hop count for further P2P file sharing protocols such as eMule's unstructured eDonkey2000 network. Currently, the mechanism has been implemented for IPv4 but IPv6 will be the dominating protocol in the prospective Internet. Thus, future work will be concerned with the adaptation of the new mechanism to IPv6 environments.

Acknowledgement

The authors would like to thank the Broadband Access Division of Nokia Siemens Networks GmbH & Co. KG in Greifswald, Germany for their inspiration and continued support in this project. This work is partly granted by Nokia Siemens Networks.

References

[1] P. Danielis, J. Skodzik, D. Timmermann, T. Bahls, and D. Duchow, "Impacts of Improved Peer Selection on Internet Traffic in BitTorrent Networks," in *International Conference on Internet Monitoring and Protection (ICIMP 2011)*, 2011, pp. 8–13.

[2] H. Schulze, K. Mochalski (ipoque), "Internet Study 2008/2009," 2009.

[3] "Joost Free Online TV," 2007. [Online]. Available: http://www.joost.com/

[4] "Zattoo TV to Go," 2007. [Online]. Available: http://zattoo.com/

[5] T. Mennecke, "DSL Broadband Providers Perform Balancing Act," 2005.

[6] X. Xiao and L. Ni, "Internet QoS: A Big Picture," vol. 13. IEEE Network Magazine, 1999, pp. 8–18.

[7] V. Aggarwal, S. Bender, A. Feldmann, and A. Wichmann, "Methodology for Estimating Network Distances of Gnutella Neighbors." GI Jahrestagung (2), 2004, pp. 219–223.

[8] R. Steinmetz and K. Wehrle, *P2P Systems and Applications, Springer Lecture Notes in Computer Science*. Springer-Verlag Berlin Heidelberg, 2005.

[9] H. Wan, N. Ishikawa, and J. Hjelm, "Autonomous Topology Optimization for Unstructured Peer-to-Peer Networks." IC-PADS, 2005, pp. 488–494.

[10] A. Rasti, D. Stutzbach, and R. Rejaie, "On the Long-term Evolution of the Two-Tier Gnutella Overlay," no. 4146697. INFOCOM, 2006.

[11] B. Y. Zhao, Y. Duan, L. Huang, A. D. Joseph, and J. D. Kubiatowicz, "Brocade: Landmark Routing on Overlay Networks," in *Proceedings of 1st International Workshop on Peer-to-Peer Systems (IPTPS)*, 2002, pp. 34–44.

[12] N. J. A. Harvey, M. B. Jones, S. Saroiu, M. Theimer, and A. Wolman, "SkipNet: A Scalable Overlay Network with Practical Locality Properties," in *Proceedings of the 4th conference on USENIX Symposium on Internet Technologies and Systems (USITS)*, 2003.

[13] M. Castro, P. Druschel, Y. Hu, and A. Rowstron, "Exploiting Network Proximity in Distributed Hash Tables," in *International Workshop on Future Directions in Distributed Computing (FuDiCo)*, 2002, pp. 52–55.

[14] S. Merugu and E. Zegura, "Adding Structure to Unstructured Peer-to-Peer Networks: The Use of Small-World Graphs." JPDC, 2005, pp. 142–153.

[15] Y. Liu, X. Liu, L. Xiao, L.M.Ni, and X. Zhang, "Location-Aware Topology Matching in P2P Systems." INFOCOM, 2004, pp. 2220–2230.

[16] IETF, "Application-Layer Traffic Optimization (alto)," 2009. [Online]. Available: http://datatracker.ietf.org/wg/alto/charter/

[17] H. Xie, Y. R. Yang, A. Krishnamurthy, and Y. L. A. Silberschatz, "P4P: Provider Portal for Applications." ACM SIGCOMM, 2008, pp. 351–362.

[18] E. Leonardi, M. Mellia, A. Horvath, L. Muscariello, S. Niccolini, and D. Rossi, "Building a cooperative P2P-TV application over a wise network: The approach of the European FP-7 strep NAPA-WINE." IEEE Communications Magazine 46 (4), 2008, pp. 20+22.

[19] B. Liu, Y. Cao, Y. Cui, Y. Lu, and Y. Xue, "Locality Analysis of BitTorrent-Like Peer-to-Peer Systems." 7th IEEE CCNC, 2010, pp. 1–5.

[20] K. Fujii and S. Goto, "Correlation between Hop Count and Packet Transfer Time." APAN/IWS, 2000.

[21] Information Sciences Institute, University of Southern California, "Internet Protocol Specification," RFC 791, September 1981.

[22] Swiss Education & Research Network (SWITCH), "Default TTL Values in TCP/IP," 2002.

[23] B. Cohen, "Incentives Build Robustness in BitTorrent." First Workshop on the Economics of Peer-to-Peer Systems, June 2003.

[24] "Bittorrent Protocol Specification v1.0," 2009. [Online]. Available: http://wiki.theory.org/BitTorrentSpecification

[25] T. Al-Harbash, "Raw IP Networking FAQ," 1999. [Online]. Available: http://www.ntua.gr/rin/rawfaq.html

[26] K. Fall and K. Varadhan, "The ns manual (the vint project)," 2008. [Online]. Available: http://www.isi.edu/nsnam/ns/doc/ns_doc.pdf

[27] I. Baumgart, B. Heep, and S. Krause, "OverSim: A flexible overlay network simulation framework," in *Proceedings of 10th IEEE Global Internet Symposium (GI '07) in conjunction with IEEE INFOCOM 2007*, 2007, pp. 79–84.

[28] P. García, C. Pairot, R. Mondéjar, J. Pujol, H. Tejedor, and R. Rallo, "Planetsim: A new overlay network simulation framework," in *Proceedings of the 19th IEEE International Conference on Automated Software Engineering (ASE). Workshop on Software Engineering and Middleware (SEM)*, 2004, pp. 123–136.

[29] D. Stingl, C. Gro, J. Rueckert, L. Nobach, A. Kovacevic, and R. Steinmetz, "Peerfactsim.kom: A simulation framework for peer-to-peer systems," in *Proceedings of the 2011 International Conference on High Performance Computing & Simulation (HPCS 2011)*, 2011, pp. 577–584.

[30] A. Montresor and M. Jelasity, "PeerSim: A scalable P2P simulator," in *Proceedings of the 9th Int. Conference on Peer-to-Peer (P2P'09)*, 2009, pp. 99–100.

[31] M. Baker and R. Lakhoo, "Peer-to-Peer Simulators," ACET, University of Reading, Tech. Rep., 2007.

[32] "The network simulator - ns-2," 2012. [Online]. Available: http://www.isi.edu/nsnam/ns/

[33] "The network simulator ns-2: Validation tests," 2008. [Online]. Available: http://www.isi.edu/nsnam/ns/ns-tests.html

[34] K. Eger, T. Hofeld, A. Binzenhofer, and G. Kunzmann, "Efficient Simulation of Large-Scale P2P Networks: Packet-level vs. Flow-level Simulations." UPGRADE-CN'07, 2007, pp. 9–16.

[35] D. Stutzbach and R. Rejaie, "Understanding Churn in Peer-to-Peer Networks." ACM SIGCOMM Internet Measurement Conference, 2006, pp. 189–202.

[36] O2/Telefonica, "Unser Netz," 2011. [Online]. Available: http://www.o2online.de/business/unternehmen/hosting/rechenzentren/netz/

www.iariajournals.org

International Journal On Advances in Intelligent Systems
ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS
issn: 1942-2679

International Journal On Advances in Internet Technology
ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING
issn: 1942-2652

International Journal On Advances in Life Sciences
eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO
issn: 1942-2660

International Journal On Advances in Networks and Services
ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION
issn: 1942-2644

International Journal On Advances in Security
ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS
issn: 1942-2636

International Journal On Advances in Software
ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS
issn: 1942-2628

International Journal On Advances in Systems and Measurements
ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL
issn: 1942-261x

International Journal On Advances in Telecommunications
AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA
issn: 1942-2601