International Journal on Advances in Systems and Measurements

2019 vol. 12 nr. 1&2

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 12, Number 1 & 2, 2019

Jonathan Blackledge, Dublin Institute of Technology, UK
Dario Bottazzi, Laboratori Guglielmo Marconi, Italy
Diletta Romana Cacciagrano, University of Camerino, Italy
Javier Calpe, Analog Devices and University of Valencia, Spain
Jaime Calvo-Gallego, University of Salamanca, Spain
Maria-Dolores Cano Baños, Universidad Politécnica de Cartagena,Spain
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Vítor Carvalho, Minho University & IPCA, Portugal
Irinela Chilibon, National Institute of Research and Development for Optoelectronics, Romania
Soolyeon Cho, North Carolina State University, USA
Hugo Coll Ferri, Polytechnic University of Valencia, Spain
Denis Collange, Orange Labs, France
Noelia Correia, Universidade do Algarve, Portugal
Pierre-Jean Cottinet, INSA de Lyon - LGEF, France
Paulo Estevao Cruvinel, Embrapa Instrumentation Centre - São Carlos, Brazil
Marc Daumas, University of Perpignan, France
Jianguo Ding, University of Luxembourg, Luxembourg
António Dourado, University of Coimbra, Portugal
Daniela Dragomirescu, LAAS-CNRS / University of Toulouse, France
Matthew Dunlop, Virginia Tech, USA
Mohamed Eltoweissy, Pacific Northwest National Laboratory / Virginia Tech, USA
Paulo Felisberto, LARSyS, University of Algarve, Portugal
Javad Foroughi, University of Wollongong, Australia
Miguel Franklin de Castro, Federal University of Ceará, Brazil
Mounir Gaidi, Centre de Recherches et des Technologies de l'Energie (CRTEn), Tunisie
Eva Gescheidtova, Brno University of Technology, Czech Republic
Tejas R. Gandhi, Virtua Health-Marlton, USA
Teodor Ghetiu, University of York, UK
Franca Giannini, IMATI - Consiglio Nazionale delle Ricerche - Genova, Italy
Gonçalo Gomes, Nokia Siemens Networks, Portugal
Luis Gomes, Universidade Nova Lisboa, Portugal
Antonio Luis Gomes Valente, University of Trás-os-Montes and Alto Douro, Portugal
Diego Gonzalez Aguilera, University of Salamanca - Avila, Spain
Genady Grabarnik,CUNY - New York, USA
Craig Grimes, Nanjing University of Technology, PR China
Stefanos Gritzalis, University of the Aegean, Greece
Richard Gunstone, Bournemouth University, UK
Jianlin Guo, Mitsubishi Electric Research Laboratories, USA
Mohammad Hammoudeh, Manchester Metropolitan University, UK
Petr Hanáček, Brno University of Technology, Czech Republic
Go Hasegawa, Osaka University, Japan
Henning Heuer, Fraunhofer Institut Zerstörungsfreie Prüfverfahren (FhG-IZFP-D), Germany
Paloma R. Horche, Universidad Politécnica de Madrid, Spain
Vincent Huang, Ericsson Research, Sweden
Friedrich Hülsmann, Gottfried Wilhelm Leibniz Bibliothek - Hannover, Germany
Travis Humble, Oak Ridge National Laboratory, USA

Florentin Ipate, University of Pitesti, Romania
Imad Jawhar, United Arab Emirates University, UAE
Terje Jensen, Telenor Group Industrial Development, Norway
Liudi Jiang, University of Southampton, UK
Kenneth B. Kent, University of New Brunswick, Canada
Fotis Kerasiotis, University of Patras, Greece
Andrei Khrennikov, Linnaeus University, Sweden
Alexander Klaus, Fraunhofer Institute for Experimental Software Engineering (IESE), Germany
Andrew Kusiak, The University of Iowa, USA
Vladimir Laukhin, Institució Catalana de Recerca i Estudis Avançats (ICREA) / Institut de Ciencia de Materials de Barcelona (ICMAB-CSIC), Spain
Kevin Lee, Murdoch University, Australia
Wolfgang Leister, Norsk Regnesentral (Norwegian Computing Center), Norway
Andreas Löf, University of Waikato, New Zealand
Jerzy P. Lukaszewicz, Nicholas Copernicus University - Torun, Poland
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Stefano Mariani, Politecnico di Milano, Italy
Paulo Martins Pedro, Chaminade University, USA / Unicamp, Brazil
Don McNickle, University of Canterbury, New Zealand
Mahmoud Meribout, The Petroleum Institute - Abu Dhabi, UAE
Luca Mesin, Politecnico di Torino, Italy
Marco Mevius, HTWG Konstanz, Germany
Marek Miskowicz, AGH University of Science and Technology, Poland
Jean-Henry Morin, University of Geneva, Switzerland
Fabrice Mourlin, Paris 12th University, France
Adrian Muscat, University of Malta, Malta
Mahmuda Naznin, Bangladesh University of Engineering and Technology, Bangladesh
George Oikonomou, University of Bristol, UK
Arnaldo S. R. Oliveira, Universidade de Aveiro-DETI / Instituto de Telecomunicações, Portugal
Aida Omerovic, SINTEF ICT, Norway
Victor Ovchinnikov, Aalto University, Finland
Telhat Özdoğan, Recep Tayyip Erdogan University, Turkey
Gurkan Ozhan, Middle East Technical University, Turkey
Constantin Paleologu, University Politehnica of Bucharest, Romania
Matteo G A Paris, Universita` degli Studi di Milano,Italy
Vittorio M.N. Passaro, Politecnico di Bari, Italy
Giuseppe Patanè, CNR-IMATI, Italy
Marek Penhaker, VSB- Technical University of Ostrava, Czech Republic
Juho Perälä, Bitfactor Oy, Finland
Florian Pinel, T.J.Watson Research Center, IBM, USA
Ana-Catalina Plesa, German Aerospace Center, Germany
Miodrag Potkonjak, University of California - Los Angeles, USA
Alessandro Pozzebon, University of Siena, Italy
Vladimir Privman, Clarkson University, USA
Mohammed Rajabali Nejad, Universiteit Twente, the Netherlands

Konandur Rajanna, Indian Institute of Science, India
Nageswara Rao, Oak Ridge National Laboratory, USA
Stefan Rass, Universität Klagenfurt, Austria
Candid Reig, University of Valencia, Spain
Teresa Restivo, University of Porto, Portugal
Leon Reznik, Rochester Institute of Technology, USA
Gerasimos Rigatos, Harper-Adams University College, UK
Luis Roa Oppliger, Universidad de Concepción, Chile
Ivan Rodero, Rutgers University - Piscataway, USA
Lorenzo Rubio Arjona, Universitat Politècnica de València, Spain
Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität Münster / North-German Supercomputing Alliance, Germany
Subhash Saini, NASA, USA
Mikko Sallinen, University of Oulu, Finland
Christian Schanes, Vienna University of Technology, Austria
Rainer Schönbein, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), Germany
Cristina Seceleanu, Mälardalen University, Sweden
Guodong Shao, National Institute of Standards and Technology (NIST), USA
Dongwan Shin, New Mexico Tech, USA
Larisa Shwartz, T.J. Watson Research Center, IBM, USA
Simone Silvestri, University of Rome "La Sapienza", Italy
Diglio A. Simoni, RTI International, USA
Radosveta Sokullu, Ege University, Turkey
Junho Song, Sunnybrook Health Science Centre - Toronto, Canada
Leonel Sousa, INESC-ID/IST, TU-Lisbon, Portugal
Arvind K. Srivastav, NanoSonix Inc., USA
Grigore Stamatescu, University Politehnica of Bucharest, Romania
Raluca-Ioana Stefan-van Staden, National Institute of Research for Electrochemistry and Condensed Matter, Romania
Pavel Šteffan, Brno University of Technology, Czech Republic
Chelakara S. Subramanian, Florida Institute of Technology, USA
Sofiene Tahar, Concordia University, Canada
Muhammad Tariq, Waseda University, Japan
Roald Taymanov, D.I.Mendeleyev Institute for Metrology, St.Petersburg, Russia
Francesco Tiezzi, IMT Institute for Advanced Studies Lucca, Italy
Wilfried Uhring, University of Strasbourg // CNRS, France
Guillaume Valadon, French Network and Information and Security Agency, France
Eloisa Vargiu, Barcelona Digital - Barcelona, Spain
Miroslav Velev, Aries Design Automation, USA
Dario Vieira, EFREI, France
Stephen White, University of Huddersfield, UK
Shengnan Wu, American Airlines, USA
Qingsong Xu, University of Macau, Macau, China
Xiaodong Xu, Beijing University of Posts & Telecommunications, China
Ravi M. Yadahalli, PES Institute of Technology and Management, India
Yanyan (Linda) Yang, University of Portsmouth, UK

International Journal on Advances in Systems and Measurements
Volume 12, Numbers 1 & 2, 2019

CONTENTS

Accelerating OpenMP Applications Through Parallel Hardware Architecture

Atakan Doğan, İsmail San
Department of Electrical and Electronics Engineering
Eskişehir Technical University
Eskişehir, Turkey
email: atdogan@eskisehir.edu.tr,
email: isan@eskisehir.edu.tr

Kemal Ebcioğlu
Global Supercomputing Corporation
Yorktown Heights, NY, USA
email: kemal.ebcioglu@global-supercomputing.com

***Abstract*—It is a well-known fact that application-specific hardware has both performance and power advantages as compared to general-purpose CPUs and GPUs. Furthermore, in order to improve the computing performance leveraging available parallelism in software and hardware, high-level parallel programming paradigms, such as OpenMP and OpenCL, have been viable choices for designing application-specific hardware. In this study, an application-specific parallel hardware architecture with a specialized memory hierarchy is proposed for a class of fork-join applications that can be modeled by an OpenMP program. Furthermore, three different case studies are provided to show how this model can be employed for the hardware acceleration of such applications.**

Keywords-OpenMP applications; high-level synthesis; application-specific hardware; NoCs; system-on-chip.

I. INTRODUCTION

This article is based on our previous paper [1] and extends it in several dimension, which at least includes a revised parallel hardware architecture model.

The OpenMP Application Programming Interface is a well-established standard for parallel programming on shared-memory multiprocessors. OpenMP has adopted the fork-join model of parallel execution. According to this model, an OpenMP program begins as a single thread of execution, called an *initial thread*. When any thread encounters an OpenMP *parallel* construct, a team of master and slave threads is created to execute the code enclosed by the construct (this corresponds to the *fork*). At the end of the construct, only the master thread continues, while all slave threads are terminated (this corresponds to the *join*) [2][3].

In the literature, there are several approaches that attempt to generate parallel hardware from OpenMP applications. These studies may be broadly grouped into three classes: (i) OpenMP-based pure hardware-based acceleration [1][4], (ii) OpenMP-based system-on-chip design with a soft processor and a number of hardware accelerators [5][6][7][8], (iii) OpenMP-based device offloading [9][10].

In [4], OpenMP parallel directive and a few worksharing and synchronization directives are first translated to synthesizable VHDL, and then from VHDL to FPGA hardware. In [4], each OpenMP thread is implemented by a finite state machine. A crucial limitation of this study is that there is no memory hierarchy. That is, an OpenMP hardware thread is only enabled to access on-chip memory resources, which clearly hampers to provide a scalable shared memory system in an efficient manner.

Any task specified by OpenMP task directive is converted into a custom hardware unit that carries out the work within that particular task [5][6]. These hardware units are then combined together to form an accelerator component. Finally, a system-on-chip is created based on a Nios II soft-core processor and a number of such accelerator components. However, this created system-on-chip is not equipped with memory hierarchy. Furthermore, [5] and [6] neither provide any details about synchronization, nor support any nested parallelism.

In [7], on the other hand, the system-on-chip with a MIPS soft-core processor has a memory hierarchy that is composed of a local memory per accelerator unit and a shared L1 data cache, both of which are implemented on on-chip Block RAMs, and off-chip DDR memory. Two special-purpose IP cores, *hardware mutex core* and *hardware barrier core*, are further defined in order to support several OpenMP synchronization directives as well. In addition, [7] features two-level nested parallelism.

Different from the single MIPS processor in [7], the system-on-chip in [8] includes one or more Microblaze soft-core processors. An OpenMP thread can be run on either a processor or a hardware subsystem in [8]. Furthermore, the system-on-chip of [8] instantiates an application-specific synchronization network based on the Shared Memory Process Network model of the related OpenMP application.

The surveyed approaches [4][5][6][7][8] so far do not leverage the OpenMP `target` directive for creating hardware accelerators. The OpenMP `target` directive [2], on the other hand, enables programmers to mark regions of an application that should be offloaded to an FPGA (or GPU or DSP) device. Additionally, the data mapping clauses of the OpenMP `target` directive help programmers specify what and how data should be mapped to the target device. Two different tool chains are introduced in [9][10]. These tool chains aim to offload OpenMP-based applications annotated with the OpenMP `target` directive to FPGA-based hardware accelerators.

A few High Level Synthesis (HLS) tools, such as [11][12] have support to produce parallel hardware from OpenCL. Finally, fork-join like hardware constructs that are automatically generated from single-threaded sequential code using compiler dependence analysis is described in [13]. The present work focuses on converting explicitly parallel OpenMP programs to parallel hardware, as opposed to converting single threaded sequential programs as in [13], to parallel hardware.

This study makes the following contributions to the literature as compared to [1], [4]-[10]: (i) A parallel hardware architecture with explicit support for the OpenMP synchronization directives is introduced. That is, it provides specialized components and networks for the hardware implementations of the OpenMP barrier, atomic, and critical directives, which are not found in [1]. (ii) The memory hierarchy with L1 and L2 caches is presented in detail to support the OpenMP memory model. In [1], however, there are a few untouched crucial issues related to OpenMP memory model. In the other studies [4]-[10], either there is no data cache, or there is a single data cache shared by all hardware threads. (iii) The model proposed here features dynamic scheduling of OpenMP for-loop iterations, while [1] supports only static scheduling. (iv) Finally, the nested parallelism in [1] is refined here to make it fully conform to the OpenMP semantics.

The rest of the paper is organized as follows: Section II summarizes a few features of OpenMP pertaining to this study. Section III introduces the proposed parallel hardware architecture. Section IV shows how this architecture provides support for the fork-join applications using three different case studies. Finally, Section V concludes the paper.

II. PARALLEL PROGRAMMING IN OPENMP

OpenMP is briefly introduced in this section to show how it can be used to express parallelism in applications. The execution model of OpenMP is based on the creation and management of threads, which requires the execution of at least one parallel region. In order to better explain the execution model, consider the following OpenMP code fragment:

```
void main_prog () {
   ......
   sequential_part-1
   ......
   #pragma omp parallel {
      ......
      parallel_region-1
      ......
   }
   ......
   sequential_part-2
   ......
}
```

According to the semantics of OpenMP, an initial thread starts with executing `sequential_part-1`. The sequential execution of the initial thread continues until it encounters `#pragma omp paralel`, which results in spawning (forking) a team consisting of itself (master thread) and additional other slave threads. Each thread in the team executes an implicit task that will be generated by the code according to `parallel_region-1`. At the end of the parallel construct, there is always an implicit barrier. Once all threads reach to this implicit barrier point, only the master thread continues its execution with `sequential_part-2`, while all slave threads are terminated, which corresponds to a join event [2][3]. There are a few points to emphasize related to the `parallel` directive:

- Any part of a program that is not enclosed by a parallel construct will be executed serially, including OpenMP worksharing constructs.
- The work of a parallel region will not be distributed among the threads in a team unless a worksharing construct is used.
- Although a parallel region is executed by all threads in the team, each thread is allowed to follow a different path of execution.

OpenMP allows any number of parallel constructs to be specified in a single program. For example, right after the end of `sequential_part-2`, there could be another `#pragma omp paralel` that encloses `parallel_region-2`. It is possible in OpenMP that each parallel region can be executed by a different number of threads.

OpenMP also supports *nested parallelism* that enables a parallel region to be nested within another one. For example, `parallel_region-1` above can include a second `#pragma omp parallel` with an additional `parallel_region-2` nested inside `parallel region-1`. Any thread of `parallel region-1` that encounters this nested parallel construct can start a new team of threads and become the master of its own team.

A. Worksharing

A worksharing construct distributes the execution of the related worksharing region among the members of the team that encounters it. Each thread executes a portion of the worksharing region in the context of its implicit task. A worksharing region has no barrier upon entry, but an implied barrier upon exit, unless a `nowait` clause is specified. Note also that a worksharing construct does not launch any new threads and it is effective only in a parallel region [2][3].

The `#pragma omp for` directive is the most important worksharing construct of OpenMP since loops are the most common source of parallelism in many applications. Here is an example OpenMP code fragment with the for directive:

```
#pragma omp parallel num_threads(4)
{
 #pragma omp for schedule(static) {
   for (i=0; i<1000; i++)
     a[i]=(b[i]+b[i+1])/2.0;
 }
}
```

The for directive causes the iterations of the loop immediately following it to be distributed across the threads and executed in parallel. The most relevant clause supported by the for directive is `schedule`, which determines how the iteration space should be distributed among the team of threads. The schedule clause accepts one of the five different scheduling choices, namely *static*, *dynamic*, *guided*, *runtime*, and *auto*. Thus, the user, the compiler, or the runtime is allowed to decide about the load balancing of threads for achieving the best application performance. In the case of static scheduling, for example, iterations are equally divided among threads as specified by the OpenMP standard. As a result, in the example given above, each thread will be assigned a task composed of 250 i-loop iterations. In the case of dynamic and guided scheduling, however, a thread is assigned a new chunk of iterations only if it completes the execution of the current task and is ready for the next one.

The `#pragma omp sections` directive allows a set of structured code blocks (e.g., several independent subroutines) to be executed in parallel by a team of threads, where each thread executes one code block at a time, and each code block will be executed exactly once. Note that all threads must finish their corresponding sections before any thread can proceed [2][3].

The `#pragma omp single` directive specifies that the associated structured block must be executed by only one of the encountering threads among in the team, while the other threads wait at an implicit barrier at the end of the single construct if the barrier is not eliminated by a `nowait` clause [2][3].

`#pragma omp paralel for` and `#pragma omp paralel sections` are parallel worksharing constructs that can be used when a parallel region is composed of only one worksharing construct. That is, the worksharing region includes all the code in the parallel region.

B. Synchronization

OpenMP does not guarantee atomicity when accessing and/or modifying shared data by multiple threads running in parallel. Consequently, the user is responsible for avoiding data race conditions among multiple threads. In order to make it easier for the user to orchestrate the access to shared data by multiple threads, OpenMP supports a few synchronization constructs, such as *critical*, *atomic*, and *barrier* [2][3].

The `#pragma omp critical` directive restricts the associated critical region of an application to be executed atomically by a single thread at a time. Suppose that a thread is currently executing inside a critical region. When another thread reaches that same critical region and attempts to execute it, it will be blocked at least until the first thread exits that critical region.

In contrast to the critical construct, the `#pragma omp atomic` directive provides that a single memory location is accessed atomically by multiple threads without interference. The atomic construct is similar to the atomic read-modify-write types of instructions in an instruction set architecture.

In OpenMP parallelism model, there are both implicit and explicit barriers. Remember that there is an implicit barrier at the entry to or exit from parallel regions and at the end of worksharing regions without the `nowait` clause. OpenMP further allows users to explicitly add a barrier to its parallel application by means of `#pragma omp barrier` directive, which ensures that no thread of a team is allowed to proceed beyond a barrier until all threads in the team have reached that point.

C. Memory Model

OpenMP is based on the relaxed-consistency shared memory model. According to this model, there is a *global shared memory* which any thread may read from or write to data; each OpenMP thread is allowed to have a *local*, *temporary view* of the global shared memory that is accessible to only the reads and writes from that thread [2][3]. Here are more details about the OpenMP memory model:

- A thread's temporary view of memory is not required to be consistent with the shared memory at all times.
- A read from a variable by a thread may not reflect all prior writes from other threads to this variable.
- A write to a variable by a thread is not immediately observable by another one.
- Both reads and writes by a thread may be completed with respect to only that thread's temporary view of memory without any access to shared memory.
- All modifications to the shared data objects by a thread must be written (flushed) back to the shared memory at the synchronization points of the program.

In order to make a thread's temporary view of memory consistent with the global shared memory, OpenMP provides users with `#pragma omp flush` directive. Executing the flush directive causes to write the whole thread-visible data state of the program, as defined by the base language, back to memory and then invalidate it in its temporary view.

III. PARALLEL HARDWARE ARCHITECTURE

Motivated by related studies in the literature, a generic parallel hardware architecture that can be instantiated by an OpenMP program for a class of fork-join parallel applications is proposed in this study and illustrated in Figure 1.

Inside an FPGA (Field Programmable Gate Array) or ASIC (Application Specific Integrated Circuit) chip in Figure 1, there are a few types of components, which include

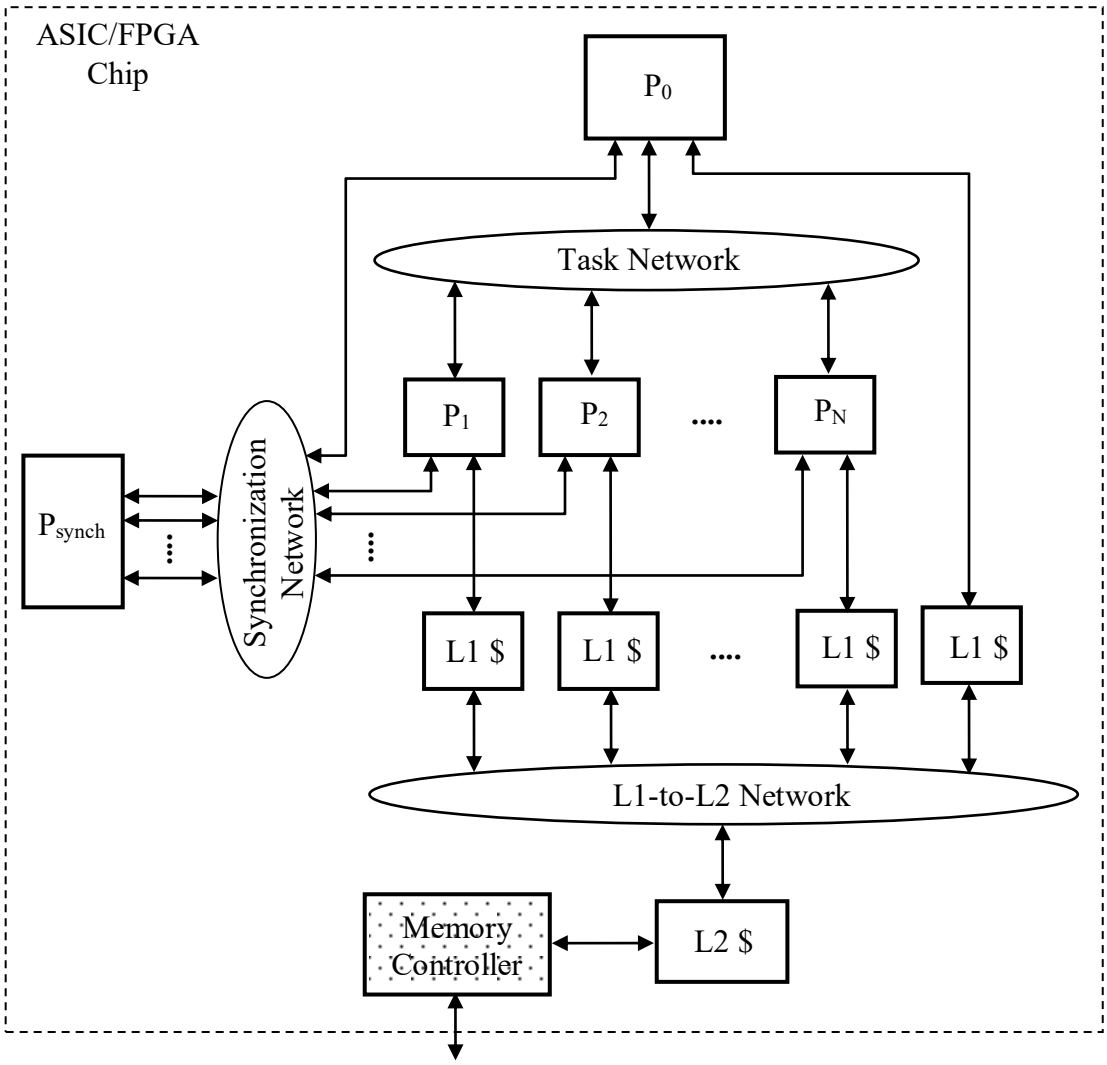


Figure 1. A parallel hardware architecture for fork-join applications

hardware threads, L1 caches (L1 $), a single L2 cache (L2 $), and interconnection networks. A two-way arrow in Figure 1 represents a bidirectional message communication port with sending FIFO (First-In First-Out) and receiving FIFO interfaces, where it can be either master or slave port. That is, a master port has a master sending FIFO to send requests and a master receiving FIFO to receive the corresponding responses; a slave port has a slave receiving FIFO to receive requests and a slave sending FIFO to send the related responses. Thus, a master port of a component is connected to a slave port of another one.

A. Hardware Threads

A hardware thread component is a finite state machine that performs either coordination (P_0 in Figure 1) or computation (P_i, $i>0$), or in some cases, both.

P_0 is the master hardware thread that coordinates/synchronizes the execution of a parallel application among the slave hardware threads P_i, $1 \leq i \leq N$. In Figure 1, P_0 has the following bidirectional ports: a master port to the task network, a master port to its L1 cache, and a master port to the synchronization network, if P_0 also contains a critical or atomic region.

P_0 implements the parallel directive as follows: P_0 forks a team of M slave threads by sending a `start request` with all initial input parameters to every slave thread of the current parallel region through its master send port to the task network. Depending on the number of threads that will be used in different parallel regions, the number of slave threads M is in general a varying number ($M \leq N$ or $M > N$ are both possible), which basically corresponds to `omp_set/get_num_threads` function of OpenMP. In order to perform a join event, P_0 waits until it receives a `finish response` over the task network from each one of the M slave threads of the team at the end of the parallel region. The `finish response` message will be received in the master receive port of P_0. Note that P_0 relies on the number of `finish response` messages received being equal to the number of `start request` messages sent, in order to implement the implicit barrier required at the end of the `parallel` directive.

According to [2], each worksharing region and barrier region must be encountered by all threads in a team or by none at all; the sequence of worksharing regions and barrier

regions encountered must be the same for every thread in a team. Otherwise, such an OpenMP program is considered to be non-conforming, which will lead to unspecified behavior. For a conforming OpenMP program, P_0 plays a role during the implementation of both implicit and explicit barriers (`#pragma omp barrier`).

P_i, $1 \leq i \leq N$, are a team of slave hardware threads that really implements the execution of a parallel application. Each slave thread in Figure 1 has the following bidirectional ports: a slave port to the task network, a master port to its L1 cache, and a master port to the synchronization network. P_i will be designed based on the following principles:

- P_i will be initially in an idle state, waiting for a `start request` message from its master hardware thread P_0.
- Upon receiving the `start request`, each slave thread immediately starts executing its associated task, which is generally realized by a deeply pipelined datapath with a finite state machine.
- The execution of a task may require one or more implicit or explicit barrier requests and one or more other synchronization operations. The details of those synchronization operations will be given shortly.
- A slave task is coupled with a private L1 cache through which it accesses its private and/or shared variables, which will be fetched from the L2 cache on demand.
- Once the computation of a slave thread is completed, the slave thread sends a `finish response` message to P_0 and starts waiting for its next `start request` message.

The static scheduling of any worksharing for loop inside a parallel region will be predetermined at compile time, by means of assigning an approximately equal chunk of loop iterations to each of the slave hardware threads.

In the case of dynamic scheduling of worksharing for a loop, however, P_0 will dynamically assign multiple chunks of loop iterations to slave threads one after another. Such dynamic scheduling can be done with a *load balancing* task network which sends `start request` messages from P_0 to *any* slave thread unit $P_1,...P_N$ that is currently free (whose slave receive port FIFO is not full). For example, implementing the task network as a 1-dimensional or 2-dimensional torus network through which task start requests sent by P_0 travel until they encounter a currently free slave thread, can accomplish such load balancing [13]. When the number M of dynamic tasks is greater than the number N of (physical) slave threads, and the task network is flooded with start requests from P_0, so that it can no longer accept further messages, P_0 will stall until the slave threads complete some of their ongoing work, so that the task network is able to accept start requests again.

On the other hand, *n-level* nested parallelism is possible by enhancing the hardware in Figure 1 as follows:

- A second level nested parallelism can be created by replacing each slave thread P_i by a copy of the part of Figure 1 consisting of the components P_0, the *task_network*, slave threads $P_1,...,P_N$ and their attached L1 caches $L1_1$,…, $L1_N$, while preserving the interconnections between these components. Components in this copy will be renamed respectively as P_i-P_0, P_i-*task_network,* P_i-P_1 … P_i-$P_{N'}$ and P_i-$L1_1$ … Pi-$L1_{N'}$ (the number of slave threads N' within the copy may be different than the original number of slave threads N). In this case, P_i-P_0 (which is the specific component that replaced P_i) will remain connected to the original *task_network* with a slave port and will remain connected to the original L1 cache $L1_i$ of P_i with a master port and will also have an additional master port attached to the new P_i-*task_network*. P_i-P_0 will perform its own computation work. It will also send subthread `start request` messages to the pool of subthread hardware units P_i-P_1 … P_i-$P_{N'}$ and will wait for all invoked subthreads to send back `finish response` messages over the new P_i-*task_network*, before P_i-P_0 itself finally sends back a `finish response` message back to P_0. The components P_i-P_0, P_i-P_1, …, P_i-$P_{N'}$ will also be connected to the synchronization network, so that a subthread running on them can execute critical or atomic regions. Finally, the P_i-$L1_j$ caches for $j \geq 1$ will be directly connected with master ports to the original L1-to-L2 network as well.
- One can repeat the previous step for each level of the nested parallelism. In case the resulting hardware does not fit on a single chip, the hardware can be partitioned into hardware modules interconnected by a scalable network and a semi-reconfigurable ASIC "union module", which can act as any of the partitions based on configuration parameters, can be created, to reduce ASIC NRE costs, as described in [13].

B. Synchronization in Hardware

As explained above, the hardware thread P_0 is used for the realization of an implicit barrier. An OpenMP explicit barrier `#pragma omp barrier` can also be implemented either with any barrier network known in the literature (e.g., for the simpler case where the number of tasks equals the number of physical hardware slave threads), or by using the mechanism already given in Figure 1. That is, a hardware thread P_0' can spawn and wait for completion of task parts before an explicit barrier and then a hardware thread P_0'' can spawn and wait for completion of task parts after the explicit barrier, and a top level hardware thread P_0 can invoke P_0', wait for its completion, and then invoke P_0'', therefore accomplishing the desired barrier synchronization.

In order to implement both critical and atomic directives of OpenMP, on the other hand, a specialized synchronization hardware unit P_{synch} is included in Figure 1. P_{synch} has a number of bidirectional slave ports to the synchronization network, where the number of slave ports n_{synch} depends on the unique synchronization identifiers in application. Note that $n_{synch} = 0$ if P_{synch} is not needed in an application without any synchronization requirement at all; $n_{synch} \leq N$ and $n_{synch} > N$ if there are less than or equal to or more unique synchronization identifiers than the number of slave threads, respectively.

Suppose that every `critical` or `atomic` directive in the application program text is assigned a synchronization identifier *synchID* between 0 and n_{synch}-1. In general, the mapping from atomic or critical constructs to *synchID*'s will be a many-to-one function. However, to enhance concurrency, two distinct critical or atomic constructs in the program text which are known not to access overlapping memory areas can be assigned different synchronization identifiers either by compiler dependence analysis or expressly by the programmer, when the programmer uses different *name* parameters in the related `critical` constructs. In order to get exclusive access to a contended structured block uniquely identified by its *synchID*, a slave thread with *threadID* is first required to send a `synchronization request` with *threadID* and *synchID* to P_{synch} through its master port to the synchronization network. The synchronization network will then route any synchronization request message by using its *synchID* as the destination network output port. P_{synch} will be receiving simultaneous multiple synchronization request messages through its slave ports to the synchronization network. In order to respond these synchronization requests as soon as possible, P_{synch} can also be designed as a parallel hardware architecture as follows:

- A finite state machine with a synchronization status register R_{synch} is assigned to manage each slave port of P_{synch}.
- Initially, R_{synch} is NULL and no thread is granted for the contended structured block.
- If there is a `synchronization request` with *synchID*, one of the following choices is applied:
  - If $R_{synch} = NULL$, let $R_{synch} = threadID$, which corresponds to an acquire event.
  - If $R_{synch} \neq NULL$ and $R_{synch} = threadID$, let $R_{synch} = NULL$, which corresponds to a release event.
  - If $R_{synch} \neq NULL$ and $R_{synch} \neq threadID$, keep R_{synch} unmodified, which ensures that only one thread is granted at any time.
- P_{synch} sends a `synchronization response` with the current value of R_{synch} to the sender thread of the respective synchronization request.

Upon receiving a synchronization response, a slave thread checks if its thread ID is equal to the R_{synch} field of the received response message. If they are equal, it means that its exclusive access has been granted by R_{synch}. At the end of the synchronization point, such a thread must send another `synchronization request` with *threadID* and *synchID* in order to end the period of its exclusive access. Otherwise, a slave thread whose request has been rejected is expected to retry after waiting for a random amount of time.

C. Memory Hierarchy

A two-level on-chip memory hierarchy as shown in Figure 1 is proposed to support the parallel hardware acceleration.

Each hardware thread in Figure 1 is associated with a dedicated, private L1 cache (L1 $) where it keeps its temporary view of the global shared memory. L1 cache has a slave port to its hardware thread and a master port to L1-to-L2 network. L1 cache is a write-back cache that supports conventional *load* and *store* requests coming from slave hardware threads. Furthermore, L1 cache has the following main features:

- L1 cache does not implement any cache coherence protocol, therefore its hardware is simplified.
- L1 cache has a *dirty bit* for each byte of a cache line in order to overcome a false sharing[1] error [15].
- In order to maintain coherence between L1 caches and globally shared L2 cache according to the respective memory model, L1 cache supports `flush_list` and `flush_all` requests.
- A `flush_list address_list request` forces L1 cache to send the cache lines containing any of the given addresses along with their line dirty bits to the L2 cache, invalidate these lines, and return an acknowledgement. Note that the address list may include a single address or multiple addresses.
- A `flush_all request` forces L1 cache to send all dirty cache lines together with their line dirty bits to the L2 cache, invalidate all cache lines, and return an acknowledgement.

The L2 cache is a write-back cache that receives *line read* and *line write* requests from L1 caches and responds to these requests accordingly. All initial and final data of the parallel application are assumed to be kept in the L2 cache. Furthermore, according to Figure 1, the L2 cache state data is held in an on-chip memory, whereas the application data are kept in an off-chip memory accessed through a memory controller. Note that L2 cache has a slave port to L1-to-L2 network and master port to its memory controller.

According to the semantics of OpenMP programs, in addition to the explicit flushes due to the flush directive, a flush operation is implied at several locations in the program as well. The implicit flushes per OpenMP requirement are supported by the proposed hardware accelerator with the help of L1 and L2 caches as follows [2]:

- *Entry to a parallel region*: Before starting the parallel region, P_0 sends a `flush_all request` to its L1 cache. Upon receiving the related acknowledgement, P_0 starts to send a `start request` to each slave thread.

[1] With non-coherent caches, a false sharing error may occur even when two slave threads access non-overlapping memory areas. Assume that a line in L2 contains two data items *a* and *b*. L1 Cache A loads the initial contents of the line and stores *a* into the line. L1 Cache B loads the initial contents of the line and stores *b* into the line. If L1 Cache A flushes the line last, it will incorrectly store the old stale value of *b*. Similarly, if L1 Cache B flushes the line last, it will incorrectly store the stale value of *a*. However, when only the bytes corresponding to the dirty bits of a line are stored back (only *a* from the line from L1 cache A and only *b* from the line from L1 cache B) into L2, this false sharing error is eliminated. Dirty bits per byte can be replaced by dirty bits per 4-byte (or 8-byte) word, when a compiler can determine that a group of L1 caches is accessed with only word accesses.

- *Exit from a parallel region*: Just before a slave thread ends (i.e., reaches the implicit barrier ending the parallel region), it sends a `flush_all request` to its L1 cache. After receiving the respective acknowledgement, the slave thread sends a `finish response` to P_0.
- *Explicit or implicit barrier region*: An explicit barrier can be implemented using an implicit one as described in Section III B first paragraph. Therefore, the L1 cache of a slave thread is flushed just before it reaches any kind of barrier.
- *Exit from a worksharing region without the nowait clause*: Remember that there is an implicit barrier at the end of a worksharing region if there is no `nowait` clause. Thus, this case will also be implemented as an implicit barrier.
- *Entry to an atomic region*: After a slave thread acquires the exclusive right for updating a single shared variable through `synchronization response`, it first sends a `flush_list request` including only the shared variable address to its L1 cache. Upon the acknowledgement of this request, it sends a load request to the cache to obtain the latest value of this variable.
- *Exit from an atomic region*: This will be achieved by first sending a `flush_list request` including the related shared variable to L1 cache, followed by a `synchronization request` with *threadID* and *synchID* to finish its atomic operation.
- *Entry to or exit from a critical region*: These two events are the same as the entry to or exit from an atomic region, except the `flush_list request` will include multiple shared variables instead of a single one.

D. Interconnection Network

In Figure 1, there are three different interconnection networks, namely *task network*, *synchronization network*, and *L1-to-L2 network*. Each of these networks is a packet-based network-on-chip network (NoC) [14] that interconnects various components of the architecture as shown in the figure. For example, the task network can be realized by a 1-to-N forward and N-to-1 backward butterfly networks, whereas the L1-to-L2 network can be implemented by a N-to-1 forward and a 1-to-N backward butterfly networks.

IV. CASE STUDIES

In this section, how different OpenMP code fragments can be compiled into application-specific parallel hardware architectures with respect to Figure 1 will be demonstrated.

A. Matrix-Vector Multiplication

The first case study considers the matrix-vector multiplication of y = A×x, where A is an n×m matrix, x and y denote m×1 and n×1 vectors, respectively. The parallel implementation of the matrix-vector multiplication in OpenMP is given below:

```
#pragma omp parallel num_threads(N) \
   default(shared) private (i,j)
{
 #pragma omp for schedule(static) {
   for (i=0; i<n; i++) {
     y[i] = 0.0;
     for (j=0; j<m; j++)
       y[i] += A[i*m+j]*x[j];
   }
 }
}
```

In Figure 1, this matrix-vector computation will be carried out as follows:

- Each hardware thread P_i, $1 \leq i \leq N$, starts its computation upon receiving a start request from P_0. Note that in this case the number of physical hardware slave threads N is specified by the clause `num_threads(N)`.
- Since OpenMP is directed to assign the iterations of the i-loop to threads in an equal fashion due to the clause `schedule(static)`, each P_i, $1 \leq i \leq N$, computes n/N vector elements y[i], where computing y[i]=A[i,:]×x requires a complete row A[i,:] of the matrix A and the whole vector x.
- The L1 cache ($L1_i$) directly attached to every P_i will be loaded with n/N rows of the matrix and the vector x from the L2 cache on demand during the computation.
- Each P_i computes its n/N part of the y vector and stores this part into its L1 cache.
- At the end of the computation, each P_i sends a flush_all request to $L1_i$ so that all dirty lines of of the y vector in $L1_i$ are written back to the L2 cache.
- Each P_i waits for a flush acknowledgement from $L1_i$, and then sends a `finish response` to P_0. Once P_0 receives N finish responses, the matrix-vector multiplication is completed.
- According to the semantics of OpenMP, there are two implicit barriers, one of which is for the end of the parallel directive, and the other one is for the end of the worksharing for directive. However, a single implicit barrier would be enough for the correct execution of the algorithm. That is why only one implicit barrier that corresponds to the end of the parallel directive is implemented.

Note that the synchronization networks and P_{synch} will not be needed for this example. Thus, the matrix-vector multiplication is realized as a single fork-join paradigm.

B. Vector Inner-Product

The second case study considers the vector inner-product of r = b×x, where b is a 1×n row vector, x denotes an n×1 column vector, and r is a resulting scalar value.

```
r=0.0;
#pragma omp parallel num_threads(N) \
   default(shared)
{
 #pragma omp for reduction(+:r)
   for (i=0; i<n; i++)
     r += b[i]*x[i];
}
```

The parallelization of the vector inner-product can be accomplished within the framework of Figure 1 as follows:

- Upon receiving a start request from P_0, each P_i, $1 \le i \le N$, computes a partial sum scalar value r_i by means of multiplying its exclusive part of n/N elements of vectors b and x, and then performing n/N sums, in pipelined fashion.
- Since each thread needs n/N elements of both vectors, $L1_i$ is loaded with n/N columns of b and n/N rows of x from the L2 cache.
- After the computation of the local r_i is over, each P_i, $1 \le i \le N$, sends a special finish response with the local value of r_i of r (`finish_reduction response`) to P_0.
- Note that P_0 initially sets as r=0.0. For each received `finish_reduction response`, P_0 updates the global value of r with the local one. After P_0 receives N finish responses, P_0 stores the final reduction sum r in cache $L1_0$.
- Finally, P_0 sends a flush request to $L1_0$. With the reception of the respective flush acknowledgement from $L1_0$, the vector inner-product computation is finished.

Once again the synchronization network and P_{synch} component in Figure 1 will not be needed for case B either. Thus, the implementation of a vector-inner product requires a fork-join type of parallel execution with a final reduction operation.

C. Gauss-Seidel Algorithm

Finally, the Gauss-Seidel algorithm is used to iteratively solve differential equations, which based on the finite difference method. A baseline OpenMP implementation of the Gauss-Seidel algorithm [16] is provided in this section:

The parallel implementation of the Gauss-Seidel algorithm is supported by Figure 1 as follows:

- P_0 executes the do-while loop as long as the loop condition is true. During each iteration of the loop, P_0 forks N slave threads in order to simply update the u[i][j] matrix and calculate the new value of dmax.
- Each P_i, $1 \le i \le N$, starts its computation upon receiving a start request from P_0. Each slave thread is assigned a task to update n/N rows of u[i][j] and computes its dmaxL value based on new u[i][j] matrix values.
- $L1_i$ is loaded with the respective n/N rows of the u[i][j] matrix from the L2 cache on demand during computation.
- At the end of each iteration of the i-loop, all N slave threads will contend for the critical section, which requires sending/receiving synchronization requests/responses through the synchronization network.
- P_{synch} will grant access to only one of the slave threads at a given time to update the shared variable dmax. Meanwhile, each slave thread reads the latest value of dmax from the main memory upon entering the critical section and writes the updated value of dmax back to main memory before exiting the critical section.
- At the end of the computation, each P_i sends a `flush_all` request to $L1_i$ so that all dirty lines of u[i][j] in $L1_i$ are written back to the L2 cache. After receiving its flush acknowledgement from $L1_i$, slave thread sends a `finish response` to P_0.
- Once P_0 receives N finish responses, P_0 checks if the loop condition is true. If it is true, it will repeat the loop as explained above. Otherwise, the Gauss-Seidel algorithm has converged, and the algorithm is completed.

```
do {
 dmax=0.0;
 #pragma omp parallel num_threads(N)\
   default(shared)
 {
  #pragma omp for private(temp,d,dmaxL){
    for(i=1; i<n+1; i++) {
      dmaxL=0.0;
      for(j=1; j<n+1; j++) {
        temp=u[i][j];
        u[i][j]= ....
        d=fabs(temp-u[i][j]);
        if(dmaxL<d) dmaxL=d;
      }
      #pragma omp critical
        if (dmax<dmaxL) dmax=dmaxL;
    }
  }
 }
} while (dmax>eps);
```

V. CONCLUSIONS

A parallel hardware architecture for a class of parallel applications that can be modeled by a fork-join programming model adopted by OpenMP is introduced. Its features are further highlighted on three different case studies. The proposed parallel hardware architecture has several important features implemented purely on hardware that are not typically supported by other studies in the literature, such as an L1 data cache for each hardware thread, n-level nested parallelism, and dynamic scheduling of worksharing for loops.

Future work involves devising a compiler to generate such parallel hardware from regular OpenMP applications; measuring and reporting the performance that can be attainable by the generated parallel hardware using a set of benchmark OpenMP applications, and making this compiler to support most of OpenMP constructs.

REFERENCES

[1] A. Doğan, İ. San, and K.Ebcioğlu, "A parallel hardware architecture for fork-join parallel applications," The Eighth International Conference on Advanced Communications and Computations, (INFOCOMP 2018), IARIA Press, July 2018, pp. 57-59.

[2] OpenMP Application Programming Interface, Version 5.0, November 2018.

[3] B. Chapman, G. Jost, R. van der Pas, Using OpenMP Portable Shared Memory Parallel Programming. London, UK: The MIT Press, 2008.

[4] Y. Y. Leow, C. Y. Ng, and W.F. Wong, "Generating hardware from OpenMP programs," IEEE International Conference on Field Programmable Technology, (FPT 2006), IEEE Press, Dec. 2006, pp. 73-80, doi: 10.1109/FPT.2006.270297.

[5] A. Podobas, "Accelerating Parallel Computations with OpenMP-driven System-on-Chip Generation for FPGAs," IEEE 8th International Symposium on Embedded Multicore/Manycore SoCs, IEEE Press, Sept. 2014, pp 149-156, doi: 10.1109/MCSoC.2014.30.

[6] A. Podobas and M. Brorsson, "Empowering OpenMP with automatically generated hardware," International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS), IEEE Press, Jul. 2016, pp. 201-205, doi: 10.1109/SAMOS.2016.7818354.

[7] J. Choi, St. Brown, and J. Anderson, "From software threads to parallel hardware in high-level synthesis for FPGAs," International Conference on Field-Programmable Technology (FPT'13), IEEE Press, Dec. 2013, pp. 270-277, doi: 10.1109/FPT.2013.6718365.

[8] A. Cilardo, L. Gallo, and N. Mazzocca, "Design space exploration for high-level synthesis of multi-threaded applications," Journal of Systems Architecture, vol. 59, pp. 1171-1183, Nov. 2013, doi: 10.1016/j.sysarc.2013.08.005.

[9] L. Sommer, J. Korinth, and A. Koch, "OpenMP device offloading to FPGA accelerators," 2017 IEEE 28th International Conference on Application-specific Systems, Architectures and Processors (ASAP 2017), IEEE Press, Jul. 2017, pp. 201-205, doi: 10.1109/ASAP.2017.7995280.

[10] D. Cabrera, X. Martorell, G. Gaydadjiev, E. Ayguade, D. J.-Gonzalez, "OpenMP extensions for FPGA Accelerators," International Symposium on Systems, Architectures, Modeling, and Simulation, IEEE Press, Jul. 2009, pp. 17-24, doi: 10.1109/ICSAMOS.2009.5289237.

[11] Xilinx SDAccel Programmers Guide. [Online]. Available fromhttps://www.xilinx.com/support/documentation/sw_manuals/xilinx2018_3/ug1277-sdaccel-programmers-guide.pdf 2019/02/22.

[12] Intel® FPGA SDK for OpenCLTM Pro Edition Programming Guide [Online]. Available from https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/hb/opencl-sdk/aocl_programming_guide.pdf 2019/02/22.

[13] K. Ebcioglu, E. Kultursay, and M. T. Kandemir, "Method and system for converting a single-threaded software program into an application-specific supercomputer," US patent 8,966,457, filed 2011/11/15 issued 2015/02/24.

[14] T. Bjerregaard and S. Mahadevan, "A survey of research and practices of network-on-chip," ACM Computing Surveys, vol. 38, pp. Jun. 2006, doi: 10.1145/1132952.1132953.

[15] E. Kultursay, K. Ebcioglu, "Storage Unsharing", US patent 8,825,982, filed 2011/06/09 issued 2014/09/02.

[16] Parallel Methods for Partial Differential Equations [Online]. Available from http://www.hpcc.unn.ru/mskurs/ENG/PPT/pp12.pdf 2019/02/22.

Santiago de Compostela Hikers and Facebook: Digital Identities and Social Representations

Christian Bourret and Joumana Boustany
DICEN IDF
Université Paris Est (UPEM)
Marne-la-Vallée - France
e-mails: {christian.bourret, joumana.boustany}@u-pem.fr

Abstract— **The Compostela Ways correspond to a strong and heavy tradition in the Middle Ages. But it was quasi-forgotten during three centuries. In the last forty years, they have constituted a highly publicised phenomenon of our ultra-modern society. In this paper, we analyse digital identities, motivations and social representations of the hikers from information and communication approach with a focus on meaning and interactions. We concentrate on the digital aspects of the Compostela Ways, especially identities, traces and interactions on social networks as a new perspective to this social phenomenon. We analyse the importance of this communication media for the hikers as a specific manner to interact and give meaning to their trip and all their life in our individualist and consumerist society.**

Keywords – digital identities; social representations; digital society; social media; Facebook; information; communication, situations; hikers, Santiago de Compostela.

I. INTRODUCTION

With the ascent of social media, the community of Santiago de Compostela hikers adopted this mediation tools: Facebook, Twitter, Instagram, etc. as it was the case previously with websites. The aim of this research is to study the digital identities and social representations of the Compostela hikers by analysing their Facebook posts. In a previous paper, published in the proceedings of HUSO 2018 Conference [1], we presented the first results. This paper consequently extends the previous results to an analysis of interactions and feelings on social networks especially on Facebook.

In fact, Santiago de Compostela Pilgrim Way is a highly publicised phenomenon and it is part of a long tradition that goes back to the 10th century [2]. It reached its peak in the twelfth and thirteenth centuries, times of affirmation of Latin Christianity in Western Europe and the Reconquest of Muslim powers in Spain. In this context, monasteries federations such as Cluny and Citeaux, Military Orders (Templars, Hospitallers, etc.) played an important role. After a long period of lethargy, for more than three centuries (1650 – 1980), the interest in Santiago de Compostela Ways has increased for the past forty years. In 1982, John Paul II was the first Pope to go to Santiago de Compostela. Since this visit, the number of people who obtained the "compostela" has increased significantly: from 2491 hikers in 1985 to 7274 in 1991, 277,854 in 2016, 301,036 in 2017 and 327,378 in 2018 from 177 countries. The Compostela is an official certificate given by Santiago's archbishopric to those pilgrims who did at least the last 100 km on foot or horseback, or the last 200 km by bicycle. For this purpose, pilgrims have to collect the stamps on the "Credencial del Peregrino" from the places they pass through to certify that they have been there. We use the word hikers, rather than pilgrims, which has a religious connotation, not shared by all the travellers. These hikers come from Spain (44.03%) Italy (8.25%), Germany (7.73%), USA (5.68%), Portugal (4.40%), France (2.68%), etc. [3]. As shown in the last available statistics, Santiago de Compostela is currently trendy all over the world. In addition to the growth rate of visitors and hikers, this interest is also shown by the number of publications: books, films, newspapers, etc. in various countries and languages.

In this paper, which is the second step in a larger project, we will focus on the digital identity of hikers in the specific issue of the "trace human", a concept defined by B. Galinon-Mélenec: "The 'trace human' would identify the Human of the 21st century, leaving everywhere traces of his passage and activities, likely tracked by merchants, watched to the detail by observers of all kinds, punished for any deviation from the norm... risking to raise legitimate concerns about the respect of privacy, the respect of individual freedoms and of ethics" [4]. The virtual community of Compostela hikers, by communicating on social media, keeps traces of their experience, but also of their life, their thought, their feeling and of the villages, towns and cities they visit. These traces allow creating the digital identity of the members of this community. Since the invention of the Internet, this issue has been subject to many publications. For Stutzman "The social network community fosters a more subjective and holistic disclosure of identity information" [5] even though it has been demonstrated that "In cyberspace the economies of interaction, communication, and coordination are different than when people meet face-to-face" [6].

Below, we explained the method used (in Section II) to study the social representation and the digital identities of Compostela hikers. Then we focused on the theoretical foundations of this study (in Section III) and explained why Compostela ways constitute an interesting field to track interactions and digital identities (in Section IV). Afterward, we presented the results for both quantitative and qualitative methods (in Section V) before discussing them (in Section VI) and concluding (in Section VII).

II. METHOD

In this paper, we study the presence of the Compostela hikers on Facebook. We address this issue through an interdisciplinary lens with a focus on the social representation theory. In fact "social representations provide criteria for evaluating the social environment that enable determination, justification or legitimization of certain behaviors" [7]. Social representations specify a number of communicative mechanisms explaining how ideas are communicated and transformed into what is perceived as common sense and allows the understanding and interpretation of the digital identity of the Facebook Compostela Pilgrim.

As a first step, we focused our study only on francophone Facebook pages considered as a public space. We excluded personal accounts as they belong to the private space as well as Facebook Groups that belong to both private and public spaces with closed and public groups. Our choice is also motivated by the characteristic of a Facebook page. It lets the page owner engage with people on Facebook as anyone can follow a page to get the public updates, even those who are not friends on Facebook. To assess our issue we used a mixed method "to achieve a systematic understanding of both the magnitude and frequency of the phenomena (quantitative) under study and the context, meaning, and motivation of those phenomena (qualitative)" [8].

The search for the keyword *Compostelle* gave 75 francophone pages on Facebook, the first one dating back to 2008 and being still active. Two of the accounts publishing these pages were commercial. As they do not allow any understanding dealing with our issue, we decided to exclude them. Since then, the number of pages has grown every year (see Figure 1) also showing the growing interest in the Compostela Ways phenomenon.

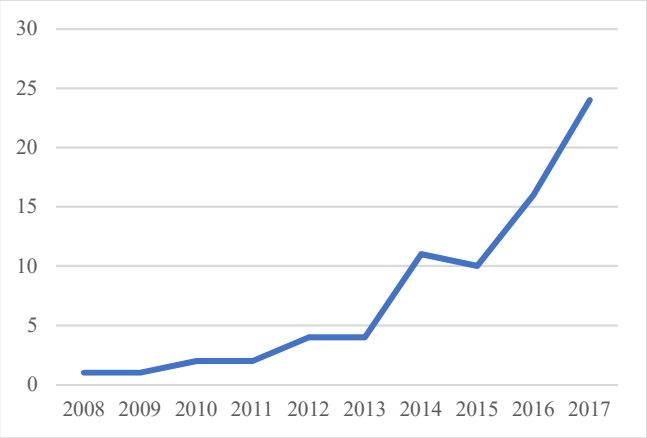


Figure 1. Number of Facebook pages created per year.

Facebook allows classifying pages in different types, more than half (42 of 75) of the pages are designated by their author as a page that brings together a community (cf. Table I). When they create their account, people are trying to federate and create a community to be able to share their interests and feelings.

TABLE I. TYPE OF FACEBOOK PAGES.

Type	*Number*
Non-profit organisations (NPO)	7
Websites & blogs	13
Communities	42
Travel agencies, Guides	6
Others (Books, Films, Sports, Events, etc.)	7

To conduct the data collection and analysis, we employed different tools.

The online tool *Likealyzer* developed by Meltwater allowed us to evaluate the activity of the selected pages. It provides data according to 5 criteria:

- *Frontpage,* which gives the first impression concerning the Facebook page.
- *About page*, which should contain milestones that give context of the page and contact information.
- *Activity,* which gives the information type (text, photos, or videos), the number of posts per day, events, etc.
- *Response,* which measures the interaction with visitors.
- *Engagement,* which relates to people talking or liking the page.

The Facebook application *Netvizz*, that allows investigating different aspects:

- *Who*? Which explore users' profiles, their relations (friendship patterns and interactions), and the larger social spaces emerging through pages.
- *What?* pages that allow for an investigation into posts, in particular concerning media types and audience engagement.
- *Where?* For all outputs containing information about users, interface language is provided in a comprehensive way, because users do not have the possibility to prevent applications from receiving this information. While interface language is certainly not a perfect stand-in for the locality, it allows engaging the question of geography in interesting ways.
- *When?* Temporal data is limited to pages, but here, a timestamp for each post and comment is provided, allowing for investigating page and user activity over time. [9]

As the *Where* aspect of Netvizz is based on the interface language, this criterion has not been accurate for our study. In fact, as we focused on French Facebook pages, it was more accurate to take into consideration the location listed in the posts than the language criterion.

Netvizz allowed us to collect 5,782 posts for the year 2018. In fact, as the number of Facebook accounts pages is growing, we decided to focus on the year 2018 which has the most important number of posts related to our issue (cf. Fig. 1).

The qualitative analysis took into consideration 15 accounts with 3,286 posts. This sample represents 56.83 % of the total number of posts for 2018. The qualitative analysis has been done by using Nvivo 12.

III. A HIGHLY PUBLICISED PHENOMENON ILLUSTRATING THE CONTRADICTIONS OF OUR SOCIETY

A. A privileged field to capture social representations and identity

In this study, the "Compostela phenomenon" has been considered in a double perspective: search for meanings and analysis of the interactions between actors: "hikers", inhabitants and the role of information and communication technologies and social networks in the new sociability of the Compostela Ways.

We also have a position of "engaged" or committed researchers [10] because one of the authors accomplished the Santiago de Compostela walk in 2011 and continues his observations and discussions with other walkers, interviews with Tourism Office members, with people in charge of Compostela Walkers' Associations, by specific documentary resources and by websites and social media networks.

The Compostela Ways constitute a privileged field to outline two key concepts in Human Sciences: social representations and identities. As defined by D. Jodelet [11], "social representations constitute an 'ordinary knowledge'. They describe, explain and recommend. They provide a method to interpret the reality, controlling our environment and driving us in society". For J.-C. Ruano-Borbalan, [12] "we interpret the world continuously through the representations that the brain accumulates… they constitute reference mind systems to understand the world around us. The social representations constitute a key concept for Human Sciences that allow the interpretation of the mechanism of intelligence, the ideologies and mentalities." The proposed framework for this study subscribes to the perspectives drawn by these two authors.

B. The modern individualism crisis: a multiple and burst identity questioning the sense of existence

Modernity, which has gradually been affirmed since the beginning of the sixteenth century in the West, largely corresponded to the development of individualism at the expense of collective institutions and traditions as trade unions, religious organisations and even States.

Having become actors of their lives in a secularised society [13], the human being has become responsible for their successes, as well as failures [14]. In a "communicating society, but where people meet less and less" [15] the individuals of social networks correspond to a "connected individualism" [16], but where the ultra-connection does not prevent hard loneliness. The "Compostela Phenomenon" has progressively developed since the 1980s in this context of a crisis of meaning, of "tiredness of being oneself" and of "connected individualism" with a lot of ambivalence. This ambivalence corresponds to those of contemporary individualism and of all our consumerist society. Our work tries to illustrate these situations and ambivalent feelings.

IV. TRACKS FOR INTERACTIONS AND DIGITAL IDENTITY ON SOCIAL NETWORKS

The Camino de Santiago – the Way of St. James – has always been a search for identity and meaning. In the Middle Ages, pilgrims belonged to various religious brotherhoods, benefiting from their assistance system (such as the organisation of groups to avoid the high level of insecurity of the ways: robberies and also murders) or a specific religious cult in a chapel dedicated to Saint James in their parish church, and also annual meetings, particularly on July 25th (St James day). Communities on Facebook reflect some of these two dimensions in the digital area that we can measure by the number of likes on each page.

The Compostela Ways may both favor a collective approach or a search for loneliness, for the inner self, but staying always connected through their mobile Internet. For the collective approach, we can cite the Facebook page "*Chemin de Compostelle*" [Way of Compostela] that gives "All the information to enjoy the experience of Compostela Ways: advantages of each way from France, where to stay, monuments, landscapes…" [17] and for the inner self, we are referring to "*Mon chemin à Compostelle*" [My way to Compostela] where we can read in the About page "I feel like at the crossroads and I have to find the one that is mine. I am looking for myself, walking to find myself ..." [18].

Compared to similar religious phenomena like Virgins of Fatima in Portugal or Lourdes devotees in France, the presence of Compostela hikers is significant on Facebook, but compared to the number of Francophone hikers who achieved the trip and get the Compostela, it is not so high.

These virtual communities vary between 35,444 followers for the largest one [17], and 89 followers for the smallest one [19]. These figures have been updated in February, 2019. The aims of both pages are to help those who make the way to Santiago de Compostela by giving advice, provide addresses, express feelings and emotions, and to surround themselves with their impressions and pictures. As these figures show, like brotherhoods and chapels in the past, social media allows hikers to brotherhoods and chapels.

Today motivations have become secularised. NPOs and communities replaced brotherhoods of the past. Interactivity is the magic word. This is done in two ways, by commenting on a post or by posting a message on the page if this function is available. This was the case for 67 pages from the selected corpus and 45 had a response rate higher than 90% that may reflect the desire to share with others. In general, hikers share their feelings, thoughts, progression, experience, photos, videos, etc. and the community reacts, reassures or encourages. Sometimes they thank hikers for the shared information. NPOs provide assistance and advice.

Today, social networks have become essential to share information about the Compostela Ways and to provide assistance, as the objectives of the pages suggest.

Commercial Facebook pages (around 8%) have the same weaknesses as personal pages: contact information is missing and there are not many followers (There are 2,163 followers for the oldest page created in 2012 [20]. The analysis of these pages shows an obvious lack of professionalism: they are not

very active (less than 30% of interaction), the *About* page is not really enticing. If this weakness is acceptable for a personal page, it is less acceptable for a professional one.

To be successful, the communication on FB pages should be optimised in order to increase user engagement. Posting too little or too much information can damage engagement and interaction with followers. The Compostela hikers post less than one message per day except for three pages that have an average of 6.6 [21], 2.7 [22] and 2 [17] posts per day.

Facebook pages are a part of the identity of the connected Compostela hikers. All of them have a Frontpage in accordance with the standards. However, most of them do not depict the full context that helps to engage people, and they do not give information about their identity, even though, hikers leave enough traces to be able to capture some traits of their identity.

The privileged means of communication remains the text, with the photo being widely used as well and the video, but to a lesser extent. Around 50% of the pages do not use photos, and 73% do not use videos.

Many hikers (22%) did not post any information on their pages. Often, the FB page accompanies the walker on his/her journey and dies shortly after the end of the adventure.

Numerous pilgrim walkers exchange on Facebook about preparation or keep in contact after the trip on the ways. Most of the associations linked to Compostela (Saint-Jean-Pied-de-Port, Pyrenean Piedmont, etc.), as well as the more official institutions (Archbishopric of Compostela) have their own Facebook page, sometimes in several languages. They provide above all information, statistics, advice, addresses. In the past, their forums were not, in our opinion, very interactive; today, with Facebook, this is no longer the case. Nevertheless, this digital dimension of the pilgrimage becomes more and more important, especially social networks.

V. Different Motivations and Social Representations

According to Santiago's Archbishopric [3], the motivations of the hikers obtaining the "Compostela" were religious and others 156,720 (47.87%), religious 140,037 (42.78), non-religious 30,621 (9.35%). The religious fact is not the only motivation for the majority of hikers. They are mainly "cultural hikers" more than pilgrims in the strict sense of the term, that is to say, with mixed religious and cultural motivations and often the challenge of a personal experience to better understand oneself in interaction with others.

Motivations may be extremely diverse. It is often a willingness to review a turning point in one's life: divorce, bereavement, retirement or entering the active life for younger adults, especially Spanish. There are also various dimensions of a group trip or an individual trip: to be able to meet others and oneself, walking on a known and valued road linked to the past, traditions and cultural heritage.

More generally, motivation is above all a search for meaning, in the ambient materialism of the consumer society. In addition, the motivation is a search for authenticity. We can cite the overworked executive director who forsakes their role, their social and hierarchical positioning for a period. Some may walk the Way in response to a wish or for a sick person (intention).

There may also be cultural motivations: the Compostela Ways are a magnificent book of art history: Romanesque, Spanish Renaissance ("*Plateresco*" in Spanish) or Baroque, discovery of diverse landscapes, contact with nature, etc. Alternatively, there are also historical and traditional motivations: to walk the routes of thousands of pilgrims who have travelled these paths for more than a millennium.

Up to a certain point, the Compostela Ways put the hikers, regardless of their social standing, on an equal footing, dealing with the challenges of a long-distance walk. However, some clues can be significantly revealing. There are those who sleep in the overcrowded refuges, those who prefer private inns and monasteries (more expensive but more comfortable), the various categories of hotels, those who carry all their necessities in their bags, those who have support cars ("*coches de apoyo*" in Spanish) from family's support or have received the services of transport companies step by step, etc.

The values of shared meaning and the quality of relationships that fostered the rise of Christianity twenty centuries ago and that of the Compostela Ways in the Middle Ages are again reflected in the success of the Compostela Ways. They create shared meaning and some solidarity, often emphasised, of people who meet new people. They help each other and learn to walk together towards the same goal and sometimes stay together (couples form on the ways).

The quality of relationships depends on the search for a certain authenticity. The language barrier is often easily overcome, as English has become the common language for the majority of hikers. For P. Nadal [23], "The Way... is not a simple walking way... [it is] an initiatory journey to the inner self that would change the perception of many things... the disconnection from the superfluous, the communion of the body with nature... For us, the Compostela Ways may be assimilated to a "semiotic machine" for the construction of a meaning of existence. It is a point to develop in future works.

Do these motivations emerge from the Facebook page's posts? Can we follow the steps of the hikers on the Compostela ways?

To be able to answer these questions, we carried out a qualitative analysis on 15 pages from the 75 selected pages. We have excluded:

- Pages with more than half of the posts containing links. They cause redundancy and they do not add anything new to what is demonstrated in the other pages.
- Pages with commercial offers. The aim of these pages is to sell services and does not contain any information about feelings and experiences of walkers.
- Pages with fewer than 50 posts all over the year 2018. These accounts are not very active.

Only text posts have been analysed, the other form of publications (text, photo, video and link) will be covered at a later stage.

As mentioned previously, we extract data by using Datavizz Facebook application. This data-driven approach enabled us to create a database of 3,336 units extracted from the 15 selected accounts that have been analysed with Nvivo 12.

A. Facebook Pages Types

When a user creates his own pages on Facebook, he has to classify his page among different categories. The remained 15 pages have been classified as follows: Three Blogs, 8 community, 2 NPO, 1 Website, culture and society and 1 touristic webpage and local attractions. In fact, a close analysis of the content of the pages allowed us to suggest this classification:

- pages that serve as synchronous diary. For some walkers, they share the preparation of the trip, their experience during the trip which village they crossed, what churches they visited and express their feelings and emotions. We can cite "Aller à Compostelle" [Go to Compostelle] [24] or "Mon chemin de Compostelle" [My way to Compostelle] [25]. They continue to publish after the trip sharing their feelings and impressions as well as information about Compostela.
- pages that serve as asynchronous diary. The posts serve to share memories, photos, readings, thoughts… For example, in *Compostelle à l'infini* [Compostela endlessly] [26] Florence [the owner of the page] explained in the *about page* why she is writing after the trip: "I do not know how to do it other than in a delayed mode. Doing it in live mode or online it is not for me. I cannot imagine myself doing 20 to 30 km per stage and make a daily report publishing it "instantly" on the web."
- mixed pages "*Pèlerins de Compostelle"* [Compostela Pilgrims] [27] managed by Fabienne Bodan who wrote many guide books on Compostela where we can read information about her trip, her personal life like the death of her father. She also shares impersonal information concerning books, events, articles or broadcasting about the Santiago ways as well as information concerning other hikers

> *Hubert, 82 years old and his walking companion Thierry from the Châlonnais both visually impaired started their trip to Santiago de Compostela. (1st of May, 2018).*

- a real community page that regroup people with a common characteristic interest living together within a larger society. These pages serve to share testimonies and photos of the Santiago Ways. For example "Radiocamino" [28] gathering pilgrims from Belgium that also meet in real life as they organize events and share activities even though this page is not categorised as a community but as a "website, culture and society."
- pages belonging to associations and promoting their activities (exhibitions, film screenings, conferences, etc.). They relay the information of some walkers and report the publication of an article or the release of a video or a broadcast about Santiago de Compostela. We can mention the page of "*Compostelle 2000"* [29] which, since 1998, has provided assistance to pilgrims and hikers on their way to Santiago of Compostela and the "*Agence des chemins de Compostelle"* [30] created in 1990 which informs the public and implements actions of cultural, educational and tourist valorisation of the old ways of pilgrimages towards Santiago de Compostela or the local association which aims to promote the way in a specific region, e.g., "Compostelle Loire Atlantique" [31].
- photo album pages where more than 80% of publications are photo tokens all the way going to Santiago de Compostela with a short comment on the place where the photo has been taken. It is the case of "Chemin Saint-Jacques de Compostelle" [Santiago de Compostela Way] [22]
- pages that could be equated to touristic guides. We can cite *"Chemin de Compostelle"*[Way of Compostela] [17]. It gives advice and information concerning the Ways of Santiago de Compostela: what could be seen or visited, the cities, the monuments, what and where to eat, gastronomy, local products, festivals, traditions, stories and legends… The stated purpose is to help walkers or pilgrims to decide which way to choose and what the best period is to walk way.

In this classification, we split on purpose pages belonging to an association and community pages. In the latter the words used to describe activities are more familiar and warmer:

> *26 participants came this Monday to Anderlecht for the first Pilgrim's get-together in Brussels. Thank you for the exchanges, the sharing, the testimonies, the meetings, the listening, the advice... Friendships were woven tonight: nothing can make me happier! See you soon on the Belgian paths and at the next get-together :-)(11/12/2018) [28]*

B. The most important is not to reach Compostela, but the Way by itself

The content analysis of selected Facebook pages demonstrates that Compostela ways are not usual ways. Many hikers write about its impressive aspect, they spoke about their feelings and emotions before, during and after the trip.

> *We left without any certainty. We have stamped our credentials. Day after day we walked without never really imagining our arrival in Saint Jacques de Compostela. (9/11/2018) [24].*

Compostela ways are a kind of mix between a personal challenge, a search of oneself and an interest in arts and history. They are a way of spiritual quest, but not only in the

religious meaning. People who do the trip try to take stock with their life when appreciating encounter and exchanges in an original way to travel as tourists.

1) A way of spiritual quest, a way to take stock of their life

Diverse posts have highlighted reasons that push a walker to involve themselves in this adventure, and who go all the way despite the difficulties. The Way is seen like a way of redemption and renewing of the self. As evidence, we can cite two different walkers who shared and commented a newspaper article [32] about three prisoners who, by making the Way, "volunteered to reflect on themselves; they expect listening, friendship and they are searching for a new meaning of their lives and to think about their future." These comments reveal their opinion and what they think about the Way:

We never tired of saying Compostela is a way for all and also a path of freedom for prisoners! (17/12/2018) [33].

The paths of Compostela as a way of redemption for the prisoners? (19/12/2018) [27]

The Way is also seen as a means to bounce back after suffering from a tragic event in life. It could be the death of a beloved person, or after a critical illness. We can highlight the post of this pilgrim who lost her father and who shared a photo of a heart on the floor. She wrote:

You were a big-hearted dad, I am offering you this heart, collected in the ways, to go with your so brutal flight to another world. Thank you for what you did for many people during your 82 years in this life. You kept your legendary dynamism until this serious stroke took you in 48 hours. I will keep you in my heart forever? (14/6/2018) [27]

Other hikers reported the story of "Julie [who] left all by herself, on the road to Arles, with a camera and some objects that belonged to her deceased mother to make a film and its mourning on the way." (12/3/2018) [27]

But for the hikers, walking the ways of Compostela is more than that:

To walk is also to share. One's doubts, joys, sorrows, and life. (3/3/2018) [27]

2) A Way of Encounters

Our form of life requires a faith that stimulates us to walk the Way. It can be qualified as a way of encounters and dialogue. This is true, as it is a theme that often comes up in many posts. Interactions with others and sometimes interactions with oneself are always reported at each stage of the trip.

By publishing on Facebook, the hikers try to establish a relationship of trust with their followers. They post messages such as "have a good day", "have a good week", "best wishes", etc. They share their experience, their difficulties. They describe each stage of their trip and their followers' comment and encourage them.

Day 12 Sainte Eutrope. Another long, but beautiful stage through the fields and vineyards. A favourable weather at the departure from Mazeray. Marvelled by the beauty of Romanesque churches, the lantern for the dead. (21/7/2018) [34]

In their post, the word Ultreia is quite recurrent. "The word 'Ultreia' (also 'ultrella' or 'ultreya') comes from Latin and it means 'beyond'. Ultreia is another pilgrim salute, like the more popular 'Buen Camino!'. While 'Buen Camino' literally means 'have a good journey, a good Camino', the meaning of 'Ultreia!' goes a bit deeper, implying encouragement to keep going, reaching 'beyond', heading onwards.

It is also believed medieval pilgrims used to greet each other with 'Ultreia, Suseia, Santiago', meaning something like 'beyond, upwards, Santiago'. Other sources suggest 'Ultreia' was used in the same way as 'Hallelujah', once pilgrims finally reached Santiago de Compostela." [35] By ending their post with this word, do hikers encourage themselves in their trip or does the repetitive aspect of this word suggest that hikers use it as an identity affiliation?

Ultreia
Every morning we walk the Way,
Every morning we go further.
Day after day, St Jacques calls us,
It is the call of Compostela
Ultreia! Ultreïa! E sus eia Deus adjuva nos !. (2/2/2018) [34]

The other kinds of interactions are the real-life encounters and exchanges on the Way. Often, the hikers report their encounters with other pilgrims, with *hospitaleras* and *hospitaleros*. They anonymously mention the confidences of the people they meet, such as the one who confides the illness of his wife, or the pilgrim whose heel hurts and so that he cannot finish his stage. The Way promotes sharing as expressed by these hikers.

After 2 hours' walking, I came across an open barn and house. At the entrance, a table is set with fruits, boiled eggs, coffee, dried fruits… En donativo (we give what we want). A great place run by cool young people." (10/6/2018) [36]

I have had wonderful encounters, people who stood by me when I went through rough patches. The Way is a great family with a classless society, we are all here for the same reason, reach the end." (8/10/2018) [25]

Talking about social classes, much the same can be said when we look at the housing system described by the walkers. In general, they are using cottages, camping and tents. Hotels were rarely mentioned at least by the walkers.

3) An obsessive Way full of happiness and suffering

Many hikers, in their posts, seem to be obsessed by the Way. Months before, they start to write about their preparation for the walk as we can see:

Well, there I cannot back off. The camino del-norte is calling me. Departure from Chartres on September 3. The 4th, I will walk in the footsteps of my fellow pilgrims." (8/7/2018) [25]

When they start walking, they express their happiness as well as their doubts:

Sometimes simply some tiny thing makes us happy. Starting our Camino from Madrid, it is with joy that we found our first yellow arrow, so symbolic of a new adventure in the ways." (6/3/2018) [27]

The way is a long-drawn-out process, but this does not prevent the walkers from enjoying it.

D2: Shelter of Orisson – Roncevaux, 16 km. At the beginning the sky was clear, but not for a long time. The difference in level is less important today than yesterday. Cold rain, snow, fog, but I ARRIVED I am so happy. A lot of emotions invaded me today: the freedom to be in this beautiful nature, in the middle of the beeches, the desire sometimes to cry, I do not know why?" (14/5/2018) [36]

But they are also flooded with doubt:

When you walk alone, greater is the desire to drop out… but you are here to test your limits even if each step counts." (2/4/2018) [34]

Walking the Way of Santiago de Compostela is more than a hike: the body and the spirit are put to the test. This adventure can change a life.

Do not worry! Everything will be alright! The shoes or the backpack do not hike. These are just simple tools. Only you are going to hike in Santiago Not even your legs but it is your brain your mind, your brain, your mind: 20% physical effort 80% morale. The most important thing is to manage the pain, the fatigue to keep the morale." (25/9/2019) [24]

When the walkers end the journey, it is not really the end. Walkers are always attracted by the Camino Ways. They have the "Camino blues" and express it:

This way is not just any way. It becomes a way of excellence, a quest for the absolute. If it is difficult to set out for this journey, how many people think or dream about it... It is also difficult to go back, find the nature, the forest, these plains as far as the eye can see, sown with corn or sunflowers, get up early with a smile and put on the boots. As soon as you get back, you feel the urge to go away from the noise, the pollution, to find that inner peace to the rhythm of my steps and my stick [mon bourdon in the original text]". Precisely, being down in the dumps [le bourdon in the original text]… let us talk about it or not. This morning, I do not take the Way, so I will not go further. Ultreia" (1/8/2018) [34]

The Camino can change a life. This could be summarised by the "about" pages of "*Compostelle à l'infini*" [Compostela endlessly] page:

This page does not aim to "unpack" everything about why or how. The Way is peculiar to everyone, and I consider that I have no advice to give. The comments that accompany my photos are only my feelings, in view of my sensitivity, my experience. I recognize it, the way has changed my life. Or rather, it transformed me. Due to a long exterior path, and therefore interior." [26]

These posts, extracted from different Facebook pages, reminded us of a quote by Lao Tzu, a Chinese philosopher from the 5th century B.C.: "There is no way to Happiness, Happiness is the way" and in our case the different ways of Compostela seem to be the Happiness.

C. Hikers identity from traces left on social networks

Facebook users are said to use "authentic identities" throughout the site's documentation. Normally, we can identify nominatively people as well as their age and sometimes their address or at least the city where they live. Most of the time people share their photo, their opinion, their way of thinking, which allows the reader to reconstitute the digital identities of the walker. Many indicators could be helpful.

A first clue could be the way hikers express themselves.

In the verbatim concerning the camino blues, we kept the word *bourdon* as this pilgrim is playing with words. In fact, *bourdon* has a double meaning in French. *Bourdon* is the walking stick of the pilgrim which was supposed to chase out infidels and devils, while "*avoir le bourdon*" can be translated as being down in the dumps.

From this linguistic detail, we can infer a humoristic trait of this pilgrim's character.

There are also those who use a literary style with many quotations:

Christian Bobin. Life is a gift of which I untie the strings every morning when I wake up (24/12/2018) [26]

Walking is not about saving time but losing it with elegance. Auguste Le Breton. (11/7/2018) [34]

And those who use a telegraphic style:

Cultural moment. Church of St Thibault 16th 17th century at château Porcien. 12/3/2018 [37]

The vocabulary used is also very revealing: for example, "the way of the cross", "path of faith", "way of life" ... denote their religious culture.

Other pilgrims express their faith and religious practices openly:

A 12 o'clock pilgrims' benediction at the cathedral. Again, I cannot help crying on hearing the crystalline voice of the nun who sings... I already think of another Camino with my love. (15/6/2018) [36]

Or when they write about the votive candles:

I lighted 3 candles in Santiago de Compostela Cathedral: one for my family to be protected, another for my friends asking for the same, one for our dear dead people wishing peace for their relatives. (9/10/2018) [25]

The way they mentioned a church is also very informative. Many hikers name it by the city name while others use the patron saint to whom the church is dedicated to designate it. It is a further indication of a religious culture.

In general, we do not enter a church just to admire the statues. For many pilgrims these churches are seen as places of local pilgrimage.

But above all, by the traces left in these posts, backtracking the journey of a tripper becomes easy from a geographical point of view.

Day 10 From Questembert Halls passing by Saint Clair fountain. We need shells to get to Malansac (10/4/2018) [34]

All along the way of Compostela, many frames or sculptures of Saint James shells indicate the Way. That is why this tripper is mentioning that he needs many shells to get to his destination.

The towns and villages mentioned in the posts cover almost all the ways from France to Compostela.

As demonstrated by D. Cardon [38] the process of identity construction has found privileged spaces in the social-network services to deploy. It is the case of the Compostela hikers. The Internet offers multiple social environments in which to perform representations of social identity, Facebook is one of them. It is a publication tool offering people original formats for narrating their personal identity. But whatever the reason to set out for this journey and the identity of the walker it still remains a great life achievement:

Tourist or pilgrim walker or Christian hikers or atheist ... does it matter? (23/3/2018) [24]

VI. Discussion: typology of hikers and internal and external way

C. Bourret [2] proposed a typology of the people met in the Compostela Ways that may be extended to users of Social Network Services:

- Authentic pilgrims (with main religious motivations),
- Semi-pilgrims or walkers' pilgrims in different groups including those called by Spanish, "*turistigrinos*", a mix of tourists and pilgrims,
- hiker-pilgrims, above all for the pleasure of the walk and its interactions,
- sportsmen or sportswomen, often walkers but also cyclists or riders, in search of physical experience and exceeding their limits.
- Cultural walkers, cyclists or riders very interested in various monuments and cultural heritage,
- minimalists, only walking a few kilometres to collect the precious stamp on the "credential" to finally obtain the precious "Compostela" as the others.
- Strictly tourists.

There are always different degrees of involvement or participation: from a few days (with special organized trips, particularly in May) to more than 2 months, but almost always

in one direction, rarely going back on the same ways, instead they use cars, coaches, trains or planes to return home.

Actually, two aspects of the Way coexist for and in each tripper in a "walking situation towards Compostela".

First, the outer dimension, the most visible one: the hikers walking, interacting, living, meeting people on the Compostela Ways. It corresponds to the walking act by itself, visible and tangible: the way with the places crossed, the difficulties, the encounters, etc.

But there is also the whole inner dimension of the ways [39], invisible, intangible. It is about the thoughts, the feelings, the internal emotions that we tried to catch.

We can draw a similar analogy with work situations, where we talk about the invisible part of the human being work. Only a visible part of human activity at work is observed and analysed. Information and Communication sciences as well as Management sciences are studying this hidden part of the whole dimension of feelings, emotions, states of mind, etc. [40]. Trying to make all visible all this invisible and unformulated part of the Compostela Ways experience. It is the challenge that we tried to tackle and to construct the digital identity of the hikers.

Most often, hikers indicate in a neutral way that they have visited a church. They rarely indicate whether they prayed or not. Is this a form of self-censorship in our secularised and highly critical society about values and the religious manifestations, especially in France? But do they walk on the Compostela Ways only by chance? Just for the physical challenge or to roll out and enjoy a beautiful art history book in the middle of striking landscapes? The search for the meaning of existence is often formulated. That of the spiritual dimension, particularly religious, is much less avowed because probably censored in our secularised society.

According to the pilgrim's posts, the places mentioned are very different. Beside the well-known places (Saintes, Rocamadour, Roncevaux, Fromista, Leon, Sahagun, etc.) they mentioned different other places and often little known that have individually marked each pilgrim. This corresponds well to our constructivist approach as defined by P. Watzlawick [41], according to which each one builds his own reality: each pilgrim builds his own path. In fact, "The way is for everyone, but everyone makes his own way," as explained by A. Etchegoyen interviewed in a documentary film on Compostelle [42]. The idea that everyone builds his own way is also present in Machado's poem mentioned in many guide books to Compostela: "Traveller, there is no path, you make the path by walking". The fact that everyone constructs his way leads us to the constructivism theory as explained by Edgar Morin [43].

Finally, there are three different parts of the Camino, totally complementary, corresponding to a progressive process and the evolution of the experience and representations of the hikers. First, the preparation of the Camino: a few months before or even sometimes for several years before the departure. Second, the Camino properly: walking on the path for one or more periods, which may take several years. And thirdly, the after Camino which is still in it, but too often forgotten as expressed by JM Maroquin, former priest of San Juan de Ortega: "When you are back home, consider that you will always remain on the way, and that you will always be there, because it is a way which does not know the end." [44]

VII. CONCLUSION

In this paper we studied the interactions of the Compostela hikers by analysing their Facebook posts. This analysis brought us to comprehend the identities and social representations of the hikers. The Compostela Ways are a very revealing ambivalence and brings into question our society. Compostela hikers always return transformed by their participation in the Compostela Ways and by their interactions with other people. In future works, we would like to try to consider the evolution of the representations and the identities of voluntary "pilgrims", at the beginning and at the end of the "pilgrimage" and thus the changes produced by their experience. As formulated by J.M. Marorquin, the "*Camino* will always be part of the walker's mind."

The Compostela Ways are a particularly favourable ground to meet others but also to find oneself, constitutive of the widened "thought," central in the new humanism advocated by L. Ferry [45], who tries to answer the question of the sense of existence, which is at the heart of the crisis of contemporary individualism.

The Compostela phenomenon is a good way to investigate one's identity, and more specifically the digital identity, with all the traces left on social media. Through the queries of the "trace human", we go back to the eternal question of life's meaning and of our presence on Earth. Humans do not escape their fate, which is to try to understand (or not) the meaning of their lives and about their passage on this Earth, regardless of the communication medium or device they use. The identity and existence questions remain.

As the *Camino* always continues, we are only at the beginning of our work to better exploit all the collected data. We want to further compare our results with deeper content analysis and new interviews with people who have walked on the Camino. With the idea to better understand the "invisible" part of the path that takes place inside each tripper. Always in the idea of "informational tracking" [46] approach, we would also want to try to follow the evolution of the digital identity of the hikers all along their way and, if possible, after, for a few years. This is another big challenge.

Acknowledgement: Authors are listed in alphabetical order; they contributed equally to this paper.

VIII. REFERENCES

[1] C. Bourret and J. Boustany, "Identities, Motivations, Social Representations in Information and Communication Situations and Digital Society: The case of Santiago de Compostela Trippers," Proceedings of *The Fourth International Conference on Human and Social Analytics,* D. J. Folds and J. O. Berndt, Eds., HUSO 2018, Venice, Italy, June 24-28 2018, pp. 11-15, ISBN: 978-1-61208-648-4.

[2] C. Bourret, "Cultural Heritage, Tourism, Identity and Sustainable Development of Territories: the case of the Compostela Ways," *4th Multidisciplinary Scientific*

Conference on Social Sciences and Arts, Hofburg Congress Center Vienna, I, pp. 3–14, 2017.

[3] Oficina del Peregrino, "Informe estadistico [Statistical Report]" 2018, https://oficinadelperegrino.com/wp-content/uploads/2016/02/peregrinaciones2018.pdf [accessed: 2009-05-12].

[4] B. Galinon-Mélénec, "L'"Homme trace", arguments" [The trace man, arguments]," in *L'Homme trace: Perspectives anthropologiques des traces contemporaines [The trace man: Anthropological Perspectives of contemporary traces],* B. Galinon-Mélénec, Ed., CNRS, Paris, 2011.

[5] F. Stutzman, "An evaluation of identity-sharing behavior in social network communities," *International Journal of Performance Arts and Digital Media*, vol. 3, no. 1, 2006.

[6] M. A. Smith and P. Kollock, eds., *Communities in cyberspace*, Routledge, London, New-York, 1999.

[7] P. Rateau, P. Moliner, and J.-C. Abric, "Social Representation Theory," in *Handbook of Theories of Social Psychology,* P. van Lange, A. Kruglanski, and E. Higgins, Eds., pp. 477–497, SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, 2012.

[8] S. L. Schensul, J. J. Schensul, and M. D. LeCompte, *Initiating ethnographic research*: *A mixed methods approach / Stephen L. Schensul, Jean J. Schensul, and Margaret D. LeCompte*, AltaMira; Towcester: Oxford Publicity Partnership [distributor], Lanham, Md., 2013.

[9] R. Bernhard, "Studying Facebook via Data Extraction: The Netvizz Application," in *Proceedings of the 5th Annual ACM Web Science Conference,* H. Davis, Ed., ACM, New York, NY, 2013.

[10] F. Bernard, "Organiser la communication d'action et d'utilité sociétales. Le paradigme de la communication engageante [Organize the communication of action and societal usefulness. The paradigm of engaging communication]," *Communication et organisation,* no. 29, pp. 64–83, 2006.

[11] D. Jodelet, "Les représentations sociales: Regard sur la connaissance ordinaire, [The Social Representations: Issue of the ordinary knowledge]," *Sciences humaines,* no. 27, pp. 16–18, 1993.

[12] J.-C. Ruano-Borbalan, "La représentation: une notion clef des sciences humaines [The representation: a key concept of humanities and social sciences]," *Sciences humaines,* no. 27, pp. 16–18, 1993.

[13] M. Gauchet, *Le désenchantement du monde : une histoire politique de la religion [The disenchantment of the world: a political history of religion]*, Gallimard, Paris, 1985.

[14] A. Ehrenberg, The weariness of the self: diagnosing the history of depression in the contemporary age, McGill-Queen's University Press, Montreal, Ithaca, 2010.

[15] P. Breton, *L'utopie de la communication : Le mythe du village planétaire [The communication utopia: the myth of the global village]*, Editions La Découverte, Paris, 1997.

[16] P. Flichy, "La société de communication [The communication society," *Cahiers français,* no. 326, pp. 65–69, 2005.

[17] "Chemin de Compostelle [Way of Compostela]," https://www.facebook.com/Chemin-de-Compostelle-801524849973655/ [accessed: 2009-05-12].

[18] "Mon chemin à Compostelle [My Way to Compostela]," https://www.facebook.com/m.et.m.compostelle/ [accessed: 2009-05-12].

[19] "Compostelle 37 [Compostela 37]," https://www.facebook.com/pg/Compostelle-37-213463845797788/about/ [accessed: 2009-05-12]

[20] "Randonnées St Jacques de Compostelle [Hiking Santiago de Compostela]," https://www.facebook.com/RandonneesCompostelle/ [accessed: 2009-05-12].

[21] "Compostelle40 [Compostela40]," https://www.facebook.com/compostelle40/ [accessed: 2009-05-12].

[22] "Chemin Saint-Jacques de Compostelle (Camino de Santiago) [Santiago de Compostela Way]," https://www.facebook.com/CheminSaintJacquesdeCompostelle/ [accessed: 2009-05-12].

[23] P. Nadal, *El Camino de Santiago a pie [The Camino de Santiago on foot]*, El País Aguilar, Madrid, 2008.

[24] "Aller à Compostelle [Go to Compostelle]," https://www.facebook.com/conpostel/ [accessed: 2009-05-12].

[25] "Mon chemin de Compostelle [My Way to Compostela]," https://www.facebook.com/Mon-chemin-de-compostelle-1640938529565405/ [accessed: 2009-05-12].

[26] "Compostelle à l'infini [Compostela endlessly]," https://www.facebook.com/compostellealinfini/ [accessed: 2009-05-12].

[27] "Pèlerins de Compostelle [Compostela Pilgrims]," https://www.facebook.com/pelerinsdecompostelle/ [accessed: 2009-05-12].

[28] "RadioCamino : Les chemins vers Compostelle [RadioCamino: ways to Compostela]," https://www.facebook.com/radiocamino/ [accessed: 2009-05-12].

[29] "Compostelle 2000 chemins [Compostela 2000 ways]," https://www.facebook.com/Compostelle-2000-chemins-920303578025278/ [accessed: 2009-05-12].

[30] "Agence des chemins de Compostelle [Compostela Ways Agency]," https://www.facebook.com/Agence-des-chemins-de-Compostelle-253008271473959/ [accessed: 2009-05-12].

[31] "Compostelle Loire-Atlantique," https://www.facebook.com/Compostelle-Loire-Atlantique-261149230933607/ [accessed: 2009-05-12].

[32] "Des pèlerins pas comme les autres vers Compostelle [Pilgrims like no other towards Compostela]," *Midi libre*, 2018-12-17.

[33] "Les Guides Lepère [Lepère guides]," https://www.facebook.com/compostelle/ [accessed: 2009-05-12].

[34] "Le chemin de Compostelle [The Compostela way]," https://www.facebook.com/caminho2018/ [accessed: 2009-05-12].

[35] "What does the word 'Ultreia' mean? - CaminoWays.com," https://caminoways.com/what-does-ultreia-mean.

[36] "Compostelle 2018 [Compostela 2018]," https://www.facebook.com/Compostelle-2018-173174803234427/ [accessed: 2009-05-12].

[37] "De Sagarmãtha 2008 à Compostelle 2018 [From Sagarmãtha 2008 to Compostela 2018]," https://www.facebook.com/De-Sagarm%C3%A3tha-2008-%C3%A0-Compostelle-2018-163456424287950/ [accessed: 2009-05-12].

[38] D. Cardon, "L'identité comme stratégie relationnelle: Identity as a Relational Strategy," *Hermes, La Revue*, n° 53, no. 1, pp. 61–66, 2009.

[39] J. M. Rodríguez Olaizola, *Peregrinar por fuera y por dentro : guía interior para peregrinos y caminantes [Pilgrimage on the outside and inside]*, Sal Terrae, Santander, 2009.

[40] A. Damasio, *Descartes' Error : Emotion, Reason and the Human Brain*, Vintage Digital, London, 2008.

[41] P. Watzlawick, *The Invented Reality: How Do We Know What We Believe We Know? (Contributions to Constructivism)*, W. W. Norton, 1984.

[42] D. Pourajeau, *Sur les chemins de Compostelle*, 2018-11-24.

[43] E. Morin, C. Atias, and Moigne, Jean Louis le, *Science et conscience de la complexité: échanges avec Edgar Morin [Science and conscience of the complexity: discussions with E. Morin]*, Librairie de L'Université, Aix-en-Provence, 1984.

[44] Y. Boëlle and P. Huchet, *365 jours sur les chemins de Compostelle [365 days on the ways of Compostela]*, Éditions "Ouest-France", Rennes, 2016.

[45] L. Ferry, *Apprendre à vivre : traité de philosophie à l'usage des jeunes générations [Learning to live: a treatise on philosophy for younger generations]*, Flammarion, Paris, 2015.

[46] S. Leleu-Merviel, *Informational tracking*, ISTE Ltd/John Wiley and Sons Inc, Hoboken NJ, 2018.

A Hardware/Software Framework for Multiantenna Receivers

Janos Buttgereit, Erik Volpert, Horst Hartmann,
Dirk Fischer, Götz C. Kappen
NTLab
University of Applied Science Münster
Münster, Germany
Email: goetz.kappen@fh-muenster.de

Tobias Gemmeke
IDS
RWTH Aachen University
Aachen, Germany
Email: gemmeke@ids.rwth-aachen.de

***Abstract*—Multiantenna receivers play a significant role if security and spectral efficiency are critical issues for given applications. This could be during the setup of large wireless sensor networks in the Internet of Things (IoT) as well as for applications which suffer from interfering signals. This paper presents a framework to setup and evaluate real-time hardware/software demonstrators, based on flexible Software Defined Radio (SDR) hardware, low-cost multiantenna-arrays and a modular software architecture. The main purpose of this framework is to evaluate cost-benefit parameters (i.e., required processing power, logic resources vs. performance of the multiantenna algorithms) of the overall multiantenna receiver (i.e., antenna, analog and digital signal processing). Therefore, size and power consumption as well as miniaturization of the demonstrator are not considered at this time. To motivate software functions and high-level software architecture, this paper gives a theoretical background of multiantenna receivers and associated algorithms. A highly adaptable and modular C++-based framework has been developed that realizes all relevant low level and high level signal processing tasks (e.g., ADC-data transfer, online system calibration, Direction of Arrival (DoA) estimation and interferer suppression), as well as graphical visualization of the spatial spectrum. The multithread-based realization of the demonstrator ensures high performance and a convenient user experience. First measurements of the whole system (i.e., low-cost antenna, C++-based high level and low level signal processing, as well as graphical visualization using a host PC) in a real-world environment, proof functional correctness while demonstrating real-time capability of the overall system. This paper gives a qualitative overview of the required effort to change the center frequency or the type of modulation. Also, the paper shows the requirements to perform a change in the application domain, e.g., switching from DoA-estimation or interferer suppression.**

***Index Terms*—Multiantenna Systems; Wireless Sensor Networks; Spectral Efficiency; Software-defined-radio; IoT; GNSS; DECT.**

I. Introduction

Wireless systems have gained a fundamental importance in our everyday life. Since the number of transmitters increases rapidly, spectral efficiency and tolerance of interfering signals will be main issues for wireless systems in the upcoming years. In the following paragraphs three widely used systems (i.e., IoT, Global Navigation Satellite Systems (GNSSs), and mobile phones based on the Digital Enhanced Cordless Telecommunication (DECT) standard), are described to show their importance and how these systems can benefit from multiantenna technology. Multiantenna receivers are able to significantly improve spectral efficiency by using digital beamforming techniques. Interferer suppression can be realized by nulling techniques in the spatial domain [1]. Finally, the DoA of signals and interferers can be estimated, which can be used to increase received signal strengths and improve the security of the communication channel by digital post-processing in the spatial domain [2], [3].

Figure 1 shows a simple stack of a wireless sensor node, featuring sensor or general data sink/source, preprocessing, and analog and digital multiantenna processing. For the IoT-case the multiantenna approach is used to communicate within the sensor network, transmit sensor data or receive configuration data. Thus, the lowest layer is the data source in most of the cases. For a GNSS receiver the multiantenna receives the signals from the satellites and generates position, navigation and timing (PNT) information, which can be used by another entity. Hence the lowest layer in Figure 1 will be described by a data sink or a memory unit.

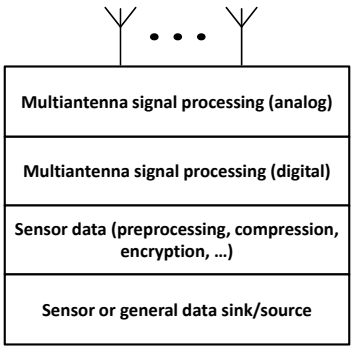


Fig. 1. Simple model of an IoT-sensor node.

The major drawbacks of multiantenna transceivers are the increased amount of required digital signal processing, as well as the complexity of algorithms and software-code. Therefore, a clear code structure, as well as efficiency, flexibility and reusability of the code play a central role, when realizing the

digital signal processing layers of the receiver. Also, for the sensor node, special care must be taken during the realization of the analog part and the data transfer to the digital domain. Especially, interferer suppression and DoA-estimation rely on coherent signal reception and processing. This paper focuses on the two upper layers (i.e., digital and analog multiantenna signal processing) shown Figure 1 and the antenna array.

Since, during the design and evaluation process, flexibility is the key challenge, a flexible and modular SDR-approach is adopted to implement receivers for various systems. A generic antenna design has been used to complete the signal processing chain. Thus, the proposed overall system features a high degree of flexibility and can be easily adapted to different receiver standards and frequency bands (e.g., IoT, DECT, GPS) described in the following subsections.

A. IoT

During the next years the number of IoT-nodes will continue to increase rapidly [4] [5]. At the same time, the complexity of IoT-nodes extends over a wide range starting with simple sensor nodes, used for temperature or humidity measurements, to fully integrated embedded systems which are able to control processes and act autonomously. Figure 2 shows the exponential increase of IoT-nodes starting from 1992 and gives a forecast of the number of devices in 2025 [5]. The figure also shows typical achievements and wireless applications over time.

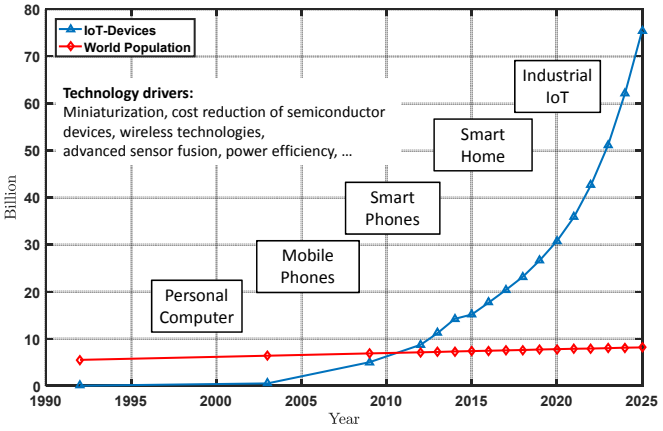


Fig. 2. IoT-Roadmap (based on: [4]–[6]).

Additionally, the world population is given for the same years and it can be seen that with the year 2011, the number of IoT-devices per person is greater than one. The dominant drivers of this evolution are miniaturization, cost reduction and increased power efficiency of semiconductor and sensor devices - and the overarching digitalization of our daily lives. Most IoT-based sensor nodes exchange data adopting wireless standards, suitable for required short or long-range communication. Thus, since the spectrum is a limited resource, spectral efficiency will play a crucial role during IoT-transceiver development. Moreover, communication security and resistance against interfering signals will be further design objectives, as they are already today in nearly all other wireless systems [7]. For all examples in this paper, dealing with IoT, a receiver setup for the 2.4 GHz ISM-band is assumed.

B. Satellite Navigation Receivers

While the first subsection deals with IoT receivers, the focus in this section lies on GNSS receivers. The most popular GNSS is the American Navstar GPS operational since 1995 [8]. Today, there are other systems available as the Russian GLONASS [9], the Chinese Beidou [10] or the European Galileo [11] based on the same idea as Navstar GPS [12]. All these systems assume that the satellite positions are known, based on the transmitted orbital data and the time of transmission. During the last decades the importance and dependency of the every day life on GNSSs has become clearly visible. Today, the power grid systems, access and industrial control, banking operations and communication systems rely heavily on GPS to provide worldwide position, navigation, and timing information. At the same time, the number of applications and sold GNSS receivers increases rapidly [13]. Also, GNSSs have become part of the critical infrastructure (see [14] and [15]), which confirms the importance of this technology.

Nevertheless, GNSSs are very vulnerable to interfering signals [16]. These signals might be transmitted unintentionally (e.g., spurious frequencies from digital television (DVBT) or from distance measurement equipment (DME) used at airports to guide incoming airplanes). The reason for the high susceptability is the very low signal power of GNSS signals received at the earth's surface. Figure 3 shows the power spectral density (PSD) for the GPS signal, the noise floor and an interfering signal. The GPS signal level is about 20 dB below the noise floor since these signals (as the signals of all other existing GNSSs) are transmitted as spread-spectrum signals. That means they are composed of a carrier signal, a Pseudo Random Noise (PRN) sequence and a data signal. To separate signals from different satellites a Code Division Multiple Access (CDMA) based scheme is adopted.

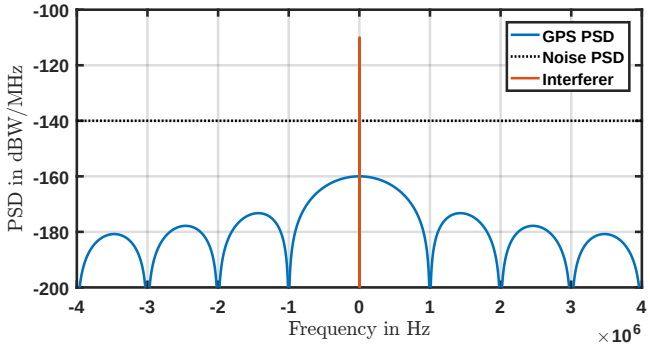


Fig. 3. Comparison of Signal and Noise PSD.

The PRN sequence is multiplied with the navigation data (which for GPS has a data rate of only 50 Hz) and therefore the overall power of the signal is distributed over a large frequency range with a main lobe bandwidth of about 2 MHz (see Figure 3). Thus, a standard GPS receiver placed near the earth's surface receives the incoming signal with only $\approx$ -158 dBW. To recover the navigation message, which includes

the orbital data, the receiver correlates the incoming signal with the known and satellite specific PRN code and thus de-spreads the signal. This correlation process compresses the signal power in a frequency band of about 100 Hz and yields about 40 dB of correlation gain. After this correlation process the 50 Hz navigation message can be recovered [17]. Afterwards, the time of transmission and orbital data can be decoded. With this data the receiver calculates the satellite position at time of transmission. Finally, the receiver determines its current position, time and velocity based on the distance measurements to at least four satellites.

While the dependency on satellite navigation increases, the disruption of GNSS signals has also become more and more obvious during the past years. Simple GNSS jammers transmit a signal with a significantly higher power in the GNSS frequency band and therefore prevent successful signal reception [17]. Figure 3 shows the PSD of GPS, Noise and the interfering signal before the despreading process. In the time domain only the interfering signal and noise is visible. After the despreading process the GPS data signal has a bandwith of about 50 Hz and a PSD maximum value of -120 dBW/MHz. At the same time the power of the CW-interferer is spread to a bandwith of about 2 MHz and the maximum value of the PSD is about 40 dB lower. It can be seen that standard GNSS receivers therefore have a certain resistance to interfering signals as long as the interferer-to-noise ratio (INR) is lower than $\approx$ 40 dB [17]. Nevertheless, interfering signals at the earth's surface easily exceed this budget because of the very weak GNSS signal. More complex architectures called GNSS-spoofers counterfeit the GNSS signal so that the user position can be faked [18]. Both types of interfering signals can be identified and suppressed based on spatial signal processing techniques using SDR-based receivers and the multiantenna design approach.

C. DECT Receivers

Mobile phones based on the DECT-standard are very widespread. In Europe DECT operates in the frequency range from 1881 MHz to about 1897 MHz and therefore directly below a frequency range assigned to the Global System for Mobile Communications (i.e., GSM). Nevertheless, since a large number of handsets and base stations are equipped with low cost oscillators, the proposed center frequency is changing over the time and thus can act as an interferer for GSM communication systems.

DoA-estimators, built based on the approach presented in this paper, significantly simplify the detection of DECT transceivers with this malfunction. The system presented is a 3D-DoA estimator and therefore replaces the standard approach using directional antennas to find interfering DECT base stations. Since the transmitted data is irrelevant for the detection of DECT transceivers, only the lowest layer of the DECT protocol should be considered for DoA-estimation. This layer is responsible for the realization of transmission channels over the radio medium.

The DECT protocol uses a combination of frequency division multiple access (FDMA), time division duplexing (TDD) and space division multiple access (SDMA) to realize several connections at the same time [19]. The DECT center frequencies for the FDMA realization can be calculated using:

$$f_c = 1897.34 \text{ MHz} - a \cdot 1.728 \text{ MHz with } a = 0, 1, ..., 9. \quad (1)$$

Furthermore, DECT defines 24 time slots, which together form a frame, shown in Figure 4. The first 12 time slots realize the downlink, the last 12 the Uplink. Each time slot includes up to 480 bits. However, part of the time available for transmission is used as guard time. Most bits are spread up to a so-called synchronization field S or data field D. In addition, a Z-field can be appended to the data field. This Z repeats the last 4 bits from to detect collisions between two channels. Figure 4 shows a full slot, which is one of the possible frame structures.

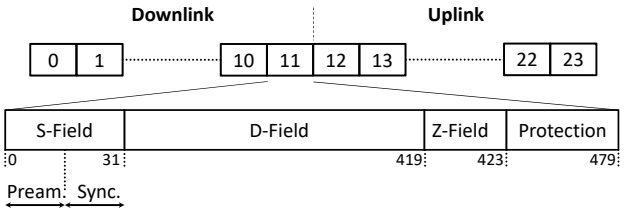


Fig. 4. DECT protocol full slot.

This lower layer allows a 3D DoA-estimation based on a spatial covariance matrix and the algorithms described in Section IV.

D. Organization of the paper

The rest of the paper is organized as follows. Section II gives a general discussion of the problem from the application's and user's point of view. It can be seen that the DoA-estimation is a crucial part during beamforming and interferer suppression, as well as during the process assessing information about the current environment. For a mathematical description, Section III defines the signal model and presents the simulation and receiver test environment. Section IV briefly introduces the spatial processing algorithms for DoA estimation and interferer suppression. Afterwards, Section V gives an in-depth description of the hardware used throughout the paper. The central part of the presented receiver is the SDR, which allows to select various frequency bands and to define sampling frequency and receiver bandwidth. Additionally, this section provides a high level overview of C++-based receiver software (low-level and high-level Digital Signal Processing (DSP)) and Graphical User Interface (GUI) programming, as well as a description of the various external and internal interfaces of the system, while details of the receiver software are discussed in Section VI. The final part of Section V presents some details of the low-cost antenna design and setup. Section VI is devoted to the software-realization of the receiver and gives implementation details of the main blocks of the receiver software (e.g., recording of the incoming frontend samples, calculation of the covariance matrix, DoA-estimation and visualization of the time plot and the DoA-spectrum. Special

emphasis lies on the thread-based realization to ensure real-time performance, portability and flexibility). Therefore, this section deals with three central points:

- Parallel realization of the receiver software tasks.
- Object oriented programming to ensure flexibility and cope with large code-complexity.
- Cross-platform realization of the software-code.

In Section VII-A and Section VII-B, the used measurement setup and measurement results are presented to show the potential of the overall receiver hardware/software-concept. Section VIII shows the steps required to change the application domain or frequency band. Finally, Section IX summarizes the paper and lists the intended optimization steps of the receiver hardware/software (i.e., miniaturization, introduction of new algorithms, introduction of new applications).

II. Problem Definition

Spatial signal processing and a flexible hardware/software architecture for real-time implementation of this processing are mandatory to solve the problems discussed earlier.

IoT-node networks suffer from the operation of a large amount of nodes in close vicinity and indoor environment. Interference and spoofing are the main problems for GNSS receivers. As discussed, for the DECT-application the search for transmitters using a wrong transmission frequency poses the relevant task. Therefore, as discussed in the introductory part this leads to the following problems:

- Interference,
- Spoofing and
- Multipath (especially in an indoor environment).

While multiantenna concepts are able to mitigate these problems, hardware and software development is time-consuming, and power consumption of the sensor node is always a critical issue [20]. Therefore, the neuralgic task is to perform a cost-benefit-analysis (e.g., minimal power consumption vs. meeting application defined development time and performance) in short time.

To quickly develop and evaluate a multiantenna systems (e.g., IoT-nodes, multi-antenna GNSS receivers, and DECT localization devices) its performance has to be assessed efficiently validating that user requirements are met. This includes the quantitative assessment of different DoA estimation or interferer suppression algorithms as well as low cost antenna setups for various real-world signal situations under real-time conditions.

Thus, the first step is to develop a modular PC-application that uses SDR-hardware as input source, runs various estimation algorithms and visualizes their results in real-time using a GUI. This application acts as a proof-of-concept demonstrator and shall help to judge performance of the algorithms and antenna arrays under various circumstances and trigger critical corner cases to ultimately develop better or cost-effective algorithms and arrays. This research focused exploration phase is followed by a design and realization phase of the low-cost and low-power sensor, analog and digital signal processing hardware (cf. Figure 1).

III. Signal model and simulation

This section introduces the signal model and the algorithms used for DoA-estimation (Capon-Beamformer and Multiple Signal Classification (MUSIC) algorithm [21]) and interferer suppression (MVDR [2] and subspaced based). Additionally, the simulation setup, as well as simulation results are presented.

A. Signal Model

In this section the signal model, based on the theory described in [2], [3] and [21], is briefly summarized while the description is restricted to one received signal. It is assumed that the receiver is in the far field of the transmitter, the narrow band assumption holds and that the antenna has a flat frequency response. Then the vector $\mathbf{u}$, which is used to describe signal and interferer, can be defined. Figure 5 shows an arbitrary antenna array with N randomly distributed antenna elements and the vector $\mathbf{u}$.

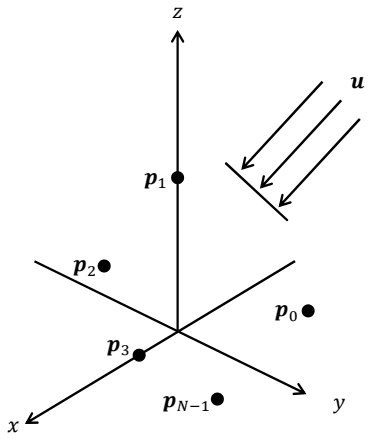


Fig. 5. Multiantenna Model.

Then, $\mathbf{u}$ can be written, depending on ϕ and θ as

$$\mathbf{u}(\phi,\theta) = \begin{pmatrix} -\sin\theta\,\cos\phi \\ -\sin\theta\,\sin\phi \\ -\cos\theta \end{pmatrix} = -\mathbf{a}(\phi,\theta) \tag{2}$$

and the wave number $\mathbf{k}$, relative to the origin of the given coordinate system can be calculated as

$$\mathbf{k}(\phi,\theta) = \frac{2\pi}{\lambda}\cdot\mathbf{u}(\phi,\theta). \tag{3}$$

In the following, the angles ϕ and θ are omitted. Furthermore, it is assumed that an N-element antenna (cf. Figure 5) receives this signal from a defined DoA. Then, the time-dependent output vector is described by the following equation:

$$\mathbf{x}(t) = \exp(-\mathrm{j}\mathbf{p}\mathbf{k})s(t) + \mathbf{n}(t) = \mathbf{a}s(t) + \mathbf{n}(t) \tag{4}$$

Next, the so called spatial covariance matrix $\mathbf{R}$ can be estimated using the estimation operator $E\{\cdot\}$ as

$$\begin{aligned}\mathbf{R} &= E\{\mathbf{x}(t)\mathbf{x}^H(t)\} \\ &= \mathbf{a}E\{\mathbf{s}(t)\mathbf{s}^H(t)\}\mathbf{a}^H + E\{\mathbf{n}(t)\mathbf{n}^H(t)\} \\ &= \mathbf{a}\mathbf{P}\mathbf{a}^H + \sigma^2\mathbf{I}\end{aligned} \tag{5}$$

Equation (5) can be re-written using a unitary matrix $\mathbf{U}$ and a matrix of the N Eigenvalues $\Lambda = \mathrm{diag}\{\Lambda_0, ..., \Lambda_{N-1}\}$ [22].

$$\begin{aligned}\mathbf{R} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H \\ &= \mathbf{U}_s\mathbf{\Lambda}_s\mathbf{U}_s^H + \mathbf{U}_n\mathbf{\Lambda}_n\mathbf{U}_n^H\end{aligned} \tag{6}$$

The Eigenvalues of noise (index n) and signal (index s) have been separated. For a real-word implementation only a limited number of samples can be recorded and used to estimate the spatial covariance matrix. Following the conventions of [21] this matrix is called $\hat{\mathbf{R}}$.

The above discussion assumes that the DoA of a useful signal should be estimated. While this use case is described in more detailed in the following section, subsection IV-B discusses the case of interferer suppression, i.e., removing interfering signals.

IV. Spatial Algorithms

A. Direction of Arrival Estimation (DoA)

In this work two DoA-estimation algorithms are considered. The Capon beamformer generates a so called spatial spectrum by using a beamsteering approach. This simple approach is limited, especially if two signal sources impinge from closely separated elevation and azimuth angles. A great benefit is the very low computational complexity and a smooth spatial spectrum. The MUSIC algorithm [21] uses a subspace based approach, which yields very high DoA-estimation accuracy at the cost of increased computational complexity. Both algorithms generate a spatial spectrum, where the maximum gives an estimate of the DoA of the incoming signal.

1) Capon Beamfomer: For the Capon beamformer, the spatial spectrum is defined as:

$$P_{\mathrm{CAP}} = \frac{1}{\mathbf{a}^H(\phi,\theta)\hat{\mathbf{R}}^{-1}\mathbf{a}(\phi,\theta)} \tag{7}$$

It can be seen that the Capon beamformer algorithm is based on the inverse spatial covariance matrix $\hat{\mathbf{R}}^{-1}$. During the search phase the steering vector $\mathbf{a}(\phi,\theta)$ is generated for all ϕ and θ, and the values for the spatial spectrum $P_{\mathrm{CAP}}(\phi,\theta)$ are calculated and stored. The final step of the algorithm is to find the maximum in the two-dimensional search space.

2) MUSIC Algorithm: The MUSIC spectrum is defined as:

$$P_{\mathrm{M}} = \frac{\mathbf{a}^H(\phi,\theta)\mathbf{a}(\phi,\theta)}{\mathbf{a}^H(\phi,\theta)\hat{\mathbf{U}}\hat{\mathbf{U}}^H\mathbf{a}(\phi,\theta)} \tag{8}$$

After the spatial spectrum has been estimated a search algorithm estimates the absolute maximum.

B. Interferer Suppression

For interferer suppression a simplified version of the Applebaum array [2] is used. Again, the estimated spatial covariance matrix $\hat{\mathbf{R}}$ is used to calculated the weights $\mathbf{W}$.

$$\mathbf{W} = \mu\hat{\mathbf{R}}^{-1}\mathbf{U}_d^* \tag{9}$$

In equation (9), μ is an arbitrary scalar constant which can be used to scale the weights $\mathbf{W}$. Again the inverse of the covariance matrix $\hat{\mathbf{R}}$ is used. Finally, $\mathbf{U}_d$ in equation (9) allows to steer the beam into a desired direction. For the GNSS case it is assumed that the received signal is composed of the useful signal, noise and the dominant interfering signal. Applying the weights $\mathbf{W}$ to the input signal, received at the N antennas, the dominant signals are removed and, in the GNSS case, the output signal has a noise like characteristic. Since the GNSS signal is about 20 dB below the noise floor the interferer suppression algorithm does not affect the useful signal. Additionally, if the satellite position is already known based on another information source (e.g., GNSS assistance data) $\mathbf{U}_d$ can be used to perform beamsteering and increase signal receive power. Again, the estimated spatial covariance matrix $\hat{\mathbf{R}}$ and the decomposition in equation (6) can be used to efficiently calculate the inverse of the spatial covariance matrix.

$$\hat{\mathbf{R}}^{-1} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}^{-1}\hat{\mathbf{U}}^H \tag{10}$$

As can be seen in equation (10), the inverse of $\hat{\mathbf{R}}$ can be calculated based on the unitary matrix $\hat{\mathbf{U}}$ and the matrix $\hat{\mathbf{\Lambda}}^{-1}$ which has the reciprocal Eigenvalues of $\hat{\mathbf{R}}$ on the main diagonal.

C. Real-time Receiver Tests

The whole receiver signal processing chain has been developed and simulated in MATLAB. This Golden Reference model has been used during the receiver design process (see Section VI) to validate the correctness of the real-time C++-based receiver results. Figure 6 shows the signal processing path for the case of DoA estimation.

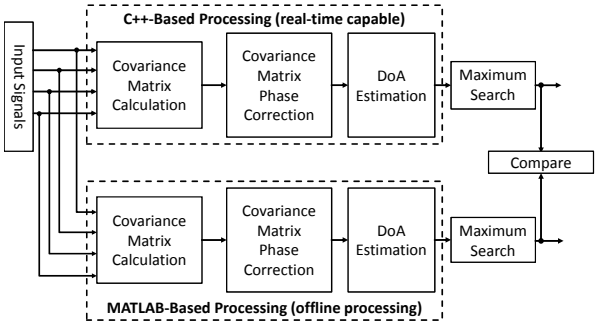


Fig. 6. MATLAB and Real-Time C++ Comparison.

The modulated carrier signals with random elevation and azimuth angles were generated in MATLAB for each sensor element and for various array geometries (i.e., circular, rectangular and uniform linear). Additionally, additive white Gaussian noise has been added to the signals (cf. equation (4)). These signals were used as input signals during C++-based and MATLAB based offline processing.

Both processing paths in Figure 6 estimate the covariance matrix, spatial spectrum, as well as the azimuth and elevation angles using floating-point precision (see Section VI for implementation details). Since input signal and data precision are

identical, the results could be directly compared, which eases the debugging of the real-time capable C++-receiver.

Additionally, the effect of a reduced precision (e.g., single precision calculations) can easily be investigated. In case of a DoA application the results show for example that the spatial spectrum of the Capon Beamformer is slightly degraded while the MUSIC spectrum is identical based on a resolution of 1° (see Section VI).

V. SDR-Based Receiver Overview

The following subsections give an overview of the hardware (i.e., SDR, host computer and low cost multiantenna) used to realize the DoA estimation, while the software is described in detail in Section VI. Mainly Ettus SDRs, equipped with daughterboards and connected to a host computer using 10 GBit/s connections are used for analog preprocessing, analog-to-digital conversion and realization of the signal processing algorithms (cf. Figure 7). On the host computer the Ettus API is used to establish the connection, control data transfer and configure the Ettus daughterboards. Moreover, the DoA-estimation, interferer suppression and calibration algorithms, as well as the GUI are implemented on the host computer. For maximal flexibility (i.e., center frequency, antenna dimensions and geometry, as well as number of antenna elements) and minimal costs, the receiver antenna array is manufactured in-house based on simple dipole antennas.

A. Receiver Hardware-Setup

A general approach of low cost multiantenna receivers for GNSS receivers has been described in [7]. Since the hardware should be used to evaluate various DoA and interferer suppression algorithms, the concept presented in this work replaces the FPGA development board used in [7] with a commercially available SDR [23]. This architecture features substantially more flexibility, which comes at significantly higher costs. A reasonable trade-off, which is acceptable during this early phase of the receiver design. The proposed receiver hardware is based on an Ettus SDR USRP X300 equipped with two SBX daughterboards [23]. Each daughterboard has a frequency range from 400 MHz to 4.4 GHz, allows duplex operation, 40 MHz bandwidth and 16-bit ADC resolution. Each X300 device can be equipped with two daughterboards, therefore a 4 channel SDR-receiver requires four SBX daughterboards and two X300.

Figure 7 shows the setup based on multiple, independent receiver units, each generating their own LO (Local Oscillator) signal. As the phase relation of the received signals is a key factor for most DoA and interferer suppression algorithms, and the LO-phase will be added to the input signal phase, independent LOs will generate useless input signals. If the phase offset between the individual LOs is known, they can be canceled out by correcting the unwanted phase shift in software. To address this issue, the SDR-receivers, used in the presented setup, have two separate inputs, one connected to the antenna and one connected to a synchronization signal that is distributed to all receivers from a central signal source.

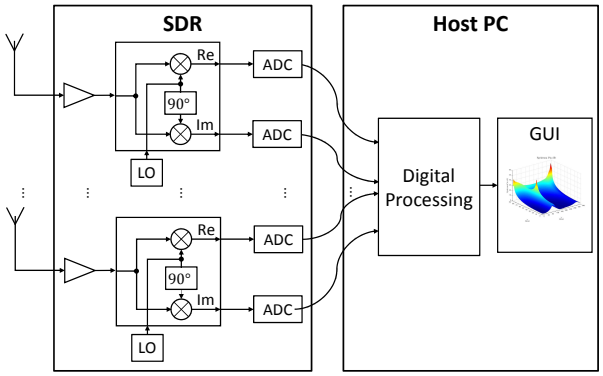


Fig. 7. General Schematic of a Multiantenna-Receiver.

Measuring results show that switching over to the synchronization signal every five seconds to re-calibrate the LO-phase error correction values is sufficient to get an overall stable measurement situation. Additionally, a time-invariant phase error is introduced by slightly different cable lengths (i.e., connections between antenna array and receiver). This error was measured once and is added as a time-invariant complex correction factor to the dynamically measured correction factors.

B. Software Overview and GUI

A high level schematic of the demonstrator software is shown in Figure 8. On the left hand side the four 16-bit digital input streams enter the signal processing stage and the spatial covariance matrix is calculated. The subsequent block performs the calibration of the spatial covariance matrix by applying time-varying and time-invariant complex correction factors.

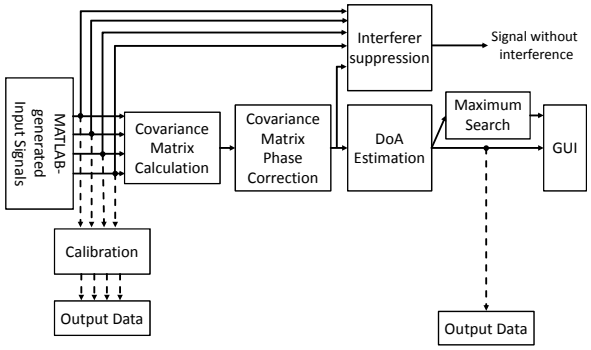


Fig. 8. Real-Time Demonstrator Schematic.

Based on the corrected covariance matrix, the estimation algorithm (e.g., MUSIC, Capon) generates the spatial spectrum, which is displayed in the GUI. A parallel task searches for the maximum in the spatial spectrum. Its numerical results (i.e., elevation and azimuth) are also displayed in the GUI (cf. Figure 17). For debugging purposes the software allows reading out the four channel input data, as well the output of the estimation algorithm. The data files can be used to compare the results of the C++-based processing of the real-time demonstrator and the MATLAB-based Golden Reference model (see Section III).

C. Antenna Setup

For the design of the 2x2 multiantenna array a two step approach has been taken [1]. First, a single ground plane antenna with four radials has been designed and manufactured (cf. Figure 9).

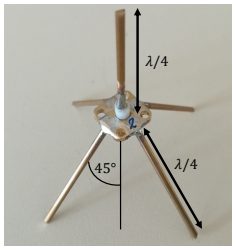


Fig. 9. Single dipole realization.

This type of antenna is low cost, easy to build and allows simplified antenna tuning [24] as well as adaptation to other frequency bands. The driven element and the four radials do have a mechanical length of approximately $\lambda/4$ for the selected center frequency f_c (cf. Figure 10). The antenna impedance is tuned to 50 Ω and the Voltage Standing Wave Ratio (VSWR) has been measured as a quality metric. Based on the approach described in [1] two more antennas (GPS and DECT) have been designed and manufactured (cf. Figure 10). For GPS this type of antenna is not the optimal choice and will be replaced by patch antennas in future designs.

System	Center Frequency	Wavelength λ	Antenna Spacing	Radial Length
ISM	2.45 GHz	12.2 cm	6.1 cm	3.1 cm
GPS	1.575 GHz	19.0 cm	9.5 cm	4.8 cm
DECT	1.89 GHz	15.9 cm	7.9 cm	4.0 cm

Fig. 10. Center Frequencies.

Figure 11 shows the resulting VSWR-plot of a single antenna for ISM-, GPS-, and DECT-center frequencies. It can be seen that all antennas achieve an VSWR ≈ 1 for the required center frequency (cf. Figure 10).

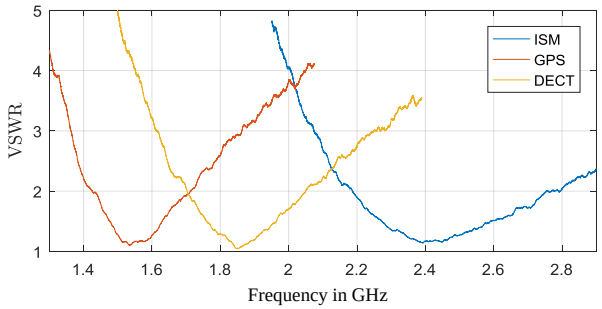


Fig. 11. VSWR-Measurement used for Antenna Tuning.

Additionally, it can be observed that the bandwidth increases for higher center frequencies. This behavior is attributed to relationship of conductor diameter and center frequency. Since at this time bandwidth is not a critical issue this effect will be ignored. As expected, things change when the single elements are combined in 2x2 array as shown in Figure 12 for the case of the ISM-, DECT- and GNSS-frequency band 2x2 multiantenna. In the construction shown, the electronic beam pattern is omni-directional for the azimuth angle, while there is no radiated energy at an elevation angle of $\phi = 0°$. This is a prefect setup for ground based signals and interferer detection systems. It will lead to problems if the desired incoming signals have larger elevation angles.

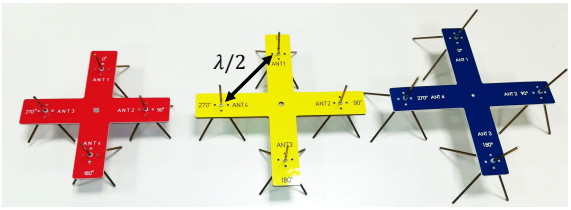


Fig. 12. Low-Cost Antennas for ISM (red), DECT (yellow) and GNSS (blue).

Again the VSWR is used to assess the quality of the manufacturing and tuning process. For the ISM antenna the results are shown in Figure 13. The figure includes all four antennas of the array.

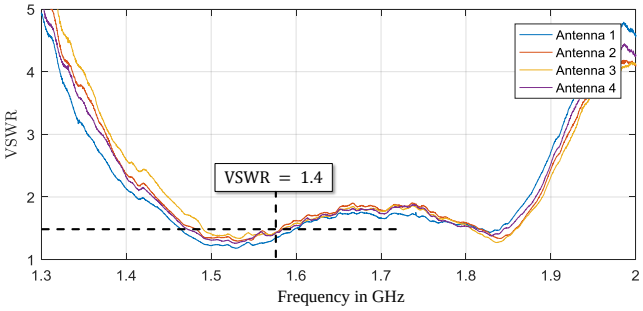


Fig. 13. 2x2 ISM Antenna.

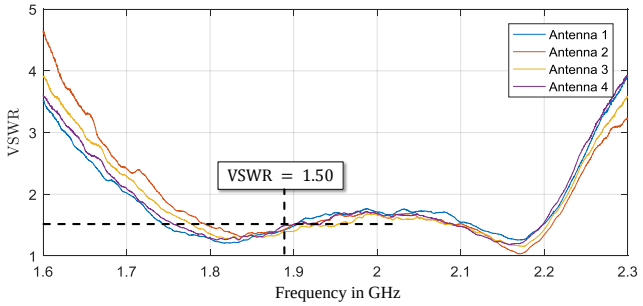


Fig. 14. 2x2 DECT Antenna.

Figures 13-15 show the results of the multiantenna realization for ISM, DECT and GPS receivers. Firstly, the figures highlight the consistency of the achieved antenna VSWR despite of the rather simple manufacturing process. Secondly, as expected, the VSWRplot has a significantly different characteristic compared to the single antenna. As shown, the array arrangement features a wider bandwidth.

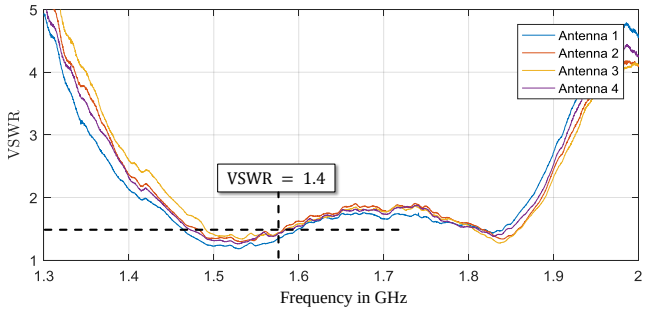


Fig. 15. 2x2 GPS Antenna.

VI. Receiver Software Realization

This section discusses details of the signal processing block realization shown in Figure 8. As mentioned in Section II the software should meet the following key constraints:

- Modular software architecture, e.g., implementing a new estimator or interferer suppression algorithm should be as easy as programming the algorithm itself.
- Modular hardware architecture, e.g., changing antenna array dimensions should be as easy as changing the description of the antenna positions, changing the center frequency should just be a change of a single variable.
- Real-time DSP performance without any sample-drop combined with an optimal GUI-operation.

Furthermore, the phase-synchronization described in Section V has to be implemented. To achieve these goals, a state-of-the-art DSP-software design flow is employed. The software is solely written in C++, using a cross-platform capable framework, originally developed for professional audio DSP-applications [25]. Besides the ability of displaying live measurement snapshots of the input signals in the time domain, the resulting spatial spectrum can be captured at any moment in time and stored to data files. This allows to analyze all parameters for various signal situations in post-processing using software like MATLAB (see Section III).

A. Concurrent Data Processing

To make use of modern multicore-CPUs and meet the throughput requirements, the computations are spread over multiple threads running in parallel, arranged in a software-pipeline structure, where each thread is a consumer of the previous thread's data and a producer for the following thread. Passing data from one thread to another is done by simply swapping buffers.

Figure 16 shows the data flow. All data exchange buffers are allocated twice at start-up. As memory allocation is a system call with unpredictable execution time on general purpose operating systems, avoiding memory allocation on the high and medium priority threads turns the operations invoked on these threads to function calls with fully predictable execution time. This guarantees that the thread's job will be predictably finished before the next data buffer is available for processing.

Samples are received by blocking calls to the Ettus UHD API [26], which invokes the 10 GBit Ethernet interface and returns as soon as a whole block of samples has been received from the hardware units and filled into the buffer passed to the API call. This buffer is forwarded to the sample processing thread afterwards, which returns the buffer it processed in the previous run to the receive thread to be filled again. This enables the new sample block to be processed, while another thread handles in parallel the acquisition of the following sample block.

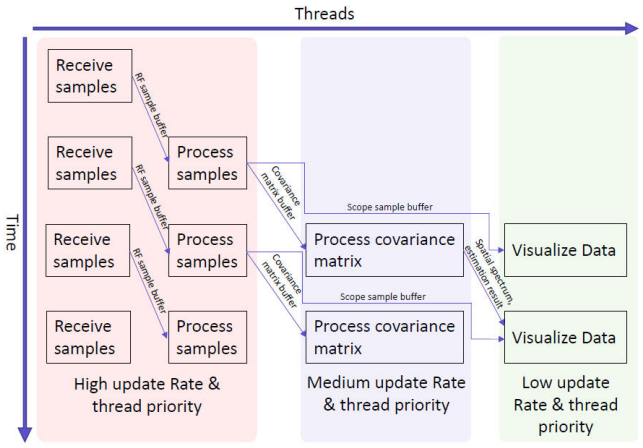


Fig. 16. Multithreaded Software Pipeline.

The sample processing thread fills a buffer if needed and then accumulates samples into the covariance matrix. Computation of this matrix is done by extensive use of SIMD-instructions on sub-vectors that exactly fit one cache-line of the CPU and uses an additional thread, not shown in the figure, to parallelize the matrix computation even further.

After a covariance matrix calculation is finished, the phase correction factors are applied to the matrix, which leads to much smaller computational load, compared to correction on a sample-basis. Depending on the covariance matrix accumulation length, which can be modified using the GUI at runtime, the accumulation process is done over several sample blocks. Thus, in general it takes several runs of the sample processing thread until a covariance matrix is handed over to the covariance matrix processing thread, which realizes the described estimator algorithm. This is why the update rate of the covariance matrix thread is slightly lower. However, the DoA-algorithms invoked on this thread, usually do some computational heavy tasks like Eigenvalue-decomposition and matrix inversion, so the broader time-slot for this thread gives it the ability to finalize computations, before the next covariance matrix is ready.

The estimation algorithms in general are expected to generate a spatial spectrum in the form of a 90x360 matrix (in case of a usual angular resolution of 1° - other values are possible) and two vectors with azimuth and elevation angles of the estimated source positions. Those buffers are again handed over to the GUI-thread that visualizes the spatial spectrum and prints out the positions of sources detected in a given interval. As updating the GUI is scheduled by the operating system, frame drops are theoretically possible at this point. However, those drops will not interrupt the processing activity.

Practically, a GUI framedrop almost never happens, which leads to a smooth presentation of the spatial spectrum.

A special case is handled when the receiver switches over to the synchronization signal. In this case, the covariance matrix computation will be paused and the phase correction value table will be updated, depending on the measured input signal phase offsets.

B. Object-Oriented Signal Processing

Object-oriented signal processing increases flexibility, as it allows a modular structure that directly models the signal flow block-diagram. Classes are used to encapsulate domains, e.g.,

- SDR-hardware
- Sample buffers
- Covariance matrix calculation
- Phase correction measurement and application
- DoA-algorithm
- Interferer suppression
- Spatial spectrum visualization

An important feature of C++ is the ability to describe (fully virtual) interface classes. This feature has been used to describe a generic DoA-algorithm class, consuming a covariance matrix and generating a spatial spectrum, as well as a pair of estimation vectors that can be overridden by an actual implementation. A Capon Beamformer, as well as a MUSIC estimator algorithm have been implemented, which can be chosen at runtime. As mentioned in the earlier sections, further algorithm development is one of the main goals. Thus, implementing new algorithms and switching from the one the other at runtime, while remaining within the same real-world signal situation, is a highly effective feature of the demonstrator. Another powerful options comes from the SDR-hardware abstraction layer, which is currently under development for its next iteration. This next generation will allow to use a completely different receiver hardware, abstracted by the same IO-interface class thus requiring minimal or no changes to the algorithm and visualization part of the software.

C. Cross-Platform Implementation

The abstraction approach described in the previous subsection allows for portability of the code to various processing platforms. In a first version, this allows to build software from the same codebase that runs on all three major operating systems (Microsoft Windows, Linux and Mac OS) without code changes. Therefore, various parts of the software can be implemented on different operating systems and could be seamlessly integrated. This approach significantly speeds up development time as team members could exactly use their development tools of choice. For the final application this results in the key benefit that the whole application or parts of it can be easily ported to an IoT-device. By design, an embedded Linux platform, as used for most IoT-devices, is a fully compatible target for the application, which radically enhances the code re-use factor for upcoming development. Furthermore, deployment to mobile platforms, like Android or iOS, are suitable options.

VII. Experimental Results

This section presents the real-time GUI developed. Additionally, compared to [1], quantitative results of DoA-estimation use cases are shown and prove the validity of the presented approach.

A. Measurement Setup

The described SDR-based demonstrator combined with the low-cost multiantenna array has been used to perform indoor measurements in the ISM-band at 2.45 GHz. Since multipath and interfering signals are expected in the utilized frequency band, the environment is close to real-world applications.

A real-time GUI (cf. Figure 17) is urgently required to setup measurement parameters, save debug data and control correct dynamic behavior during the measurements. The GUI features some additional options (e.g., taking a data snapshot, real-time modification of receiver parameters, selection of the DoA-Algorithm), which help to improve overall measurement results, and ease software debugging. The receiver raw data storage capability allows fast development of new algorithms.

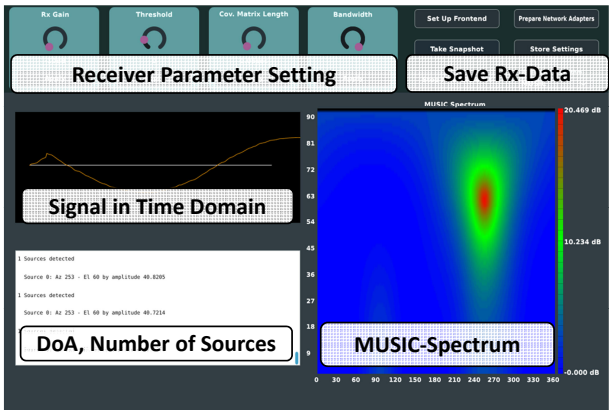


Fig. 17. Graphical User Interface of the Multiantenna-Receiver.

Besides, a first profiling has be conducted to evaluate the computational requirements of the three threads shown in Figure 16. The profiling results show that about 53% of the overall processing time is consumed by the GUI and the user interaction (i.e., the green block in Figure 16) while 45% is required for the covariance matrix calculations and the DoA algorithm (blue block in Figure 16). The high priority thread (red block in Figure 16) only consumes about 1.5% of the overall processing time. These numbers are a good starting point for optimization and for comparison of various DoA-estimation algorithms.

B. Measurement Results

This section presents first quantitative measurement results for the multiantenna DoA-estimation in the ISM-frequency band. The measurement setup is composed of a single dipole transmit antenna and a multiantenna receiver as shown in Figure 18. The setup guarantees a constant distance between transmitter and multiantenna. Additionally, the setup ensures that the the receiver operates under far field conditions.

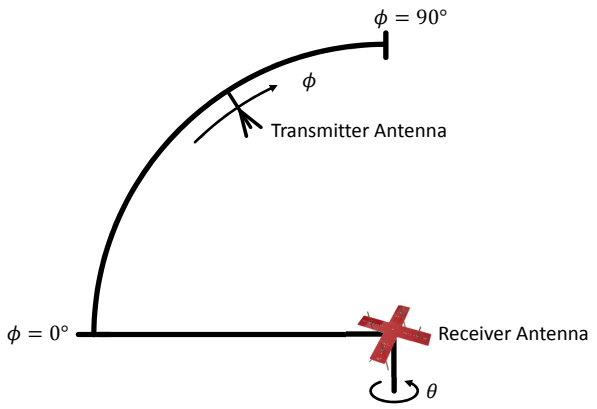


Fig. 18. Measurement setup.

Elevation angles are measured with a $10°$ spacing moving the transmit antenna from $10°$ to $90°$. Azimuth angles are modified by rotating the multiantenna. Three different azimuth angles have been selected ($90°$, $135°$, and $180°$).

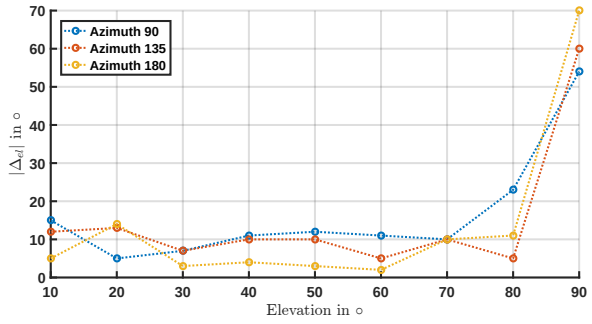


Fig. 19. Elevation error.

Figures 19 and 20 show the absolute estimation error for the elevation angle estimation $|\Delta_{el}|$ and the azimuth angle estimation $|\Delta_{az}|$.

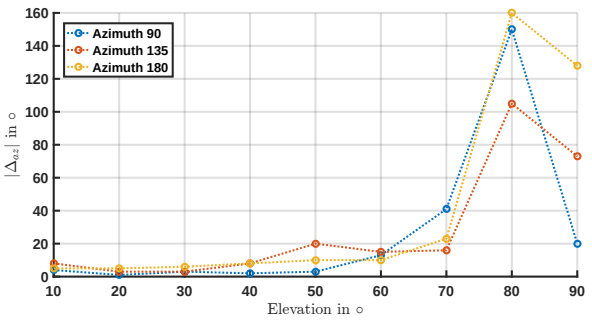


Fig. 20. Azimuth error.

For the elevation angle it can be seen in Figure 19 that the estimation error $|\Delta_{el}|$ is well below $20°$ for elevation angles smaller than $60°$. Because of the dipole beampattern and ambiguities, the estimation error increases rapidly for elevation angles larger than $70°$. A similar result is shown in Figure 20 for azimuth error $|\Delta_{az}|$. The overall results show errors about $10°$ for moderate elevation angles and again the error increases rapidly for elevation angles larger than $70°$.

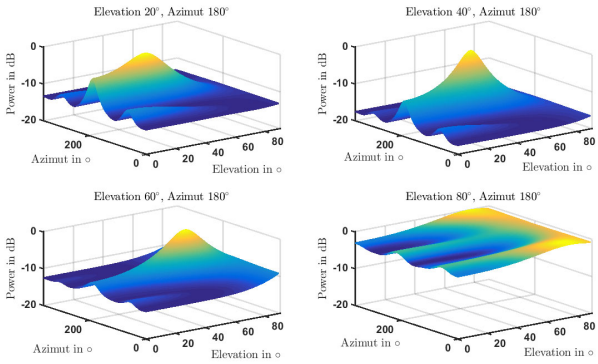


Fig. 21. 3D-DoA-estimation plot.

Figure 21 presents the DoA-estimation for four different elevation angles while a constant azimuth angle of $180°$ has been chosen. Again, it can be seen that the DoA is estimated correctly for elevation angles smaller than $60°$ and that the MUSIC spectrum degrades for larger elevation angles.

VIII. Adaptation to other Systems

As shown in [1] the proposed system is able to estimate the DoA of incoming RF-signals in the ISM-frequency band. In this section the simple adaptation of the hardware and software of the SDR-based receivers to other systems will be presented and a qualitative estimation of the required working time is given. Because of the high flexibility of the SDR-based approach and the low-cost antenna manufacturing process, switching to another system requires three main steps:

- Specification and manufacturing of the multiantenna
- Modification of constants in the SDR software code
- Measurements and calibration of the overall system

Thus, switching the system will only take a few hours, while the main time will be invested in measurements and calibration of the overall system. Modification of the SDR-code is mainly a modification of some variables. For the multiantenna manufacturing a computer numerical controlled (CNC) mill has been used to realize the antenna holding, while the single dipole antenna have been manufactured and tuned manually.

IX. Conclusion and Future Work

Spectral efficiency, robustness and security are critical design parameters of wireless IoT-sensor nodes and other wireless systems. Since costs (i.e., silicon area, power consumption) of multiantenna IoT-sensor nodes, compared to single antenna sensor nodes, are significantly higher, a detailed cost-benefit analysis has to be performed in a first step. This paper presents a modular and flexible hardware-/software-architecture, based on an SDR, which realizes the analog preprocessing and the AD-conversion. The modular C++-code realizes all digital signal processing parts, allows simple debugging and features easy extendability. The presented modular and generic approach supports porting the existing software to embedded platforms to reduce size and power consumption in a next step. Finally, a simple technique to realize low-cost antenna arrays

supports the overall approach. Measurements and simulations validate functional correctness and the demonstrator shows real-time capability of the overall receiver. The presented concept and design framework has been used to realize multiantenna receivers operating in the ISM-, DECT- and GNSS-frequency band.

ACKNOWLEDGMENT

The authors would like to thank the team of the Central Area of Electrical Engineering and Computer Science (ZBE) for their support during the antenna manufacturing process. Also, we like to thank Kai Eßmann for the 3D-design of the multiantenna mounting.

REFERENCES

[1] J. Buttgereit, E. Volpert, H. Hartmann, D. Fischer, G. C. Kappen, and T. Gemmeke, "Real-Time SDR-Based ISM-Multiantenna Receiver for DoA-Applications," in *Proceedings of the Eleventh International Conference on Advances in Circuits, Electronics and Micro-electronics, CENICS*, 2018, pp. 5–10.

[2] R. T. Compton, *Adaptive Antennas, Concepts and Performance*. Prentice Hall, 1988.

[3] H. van Trees, *Optimum Array Processing*. John Wiley and Sons, Inc., 2002, ISBN: 9780471093909.

[4] Cisco Internet Business Solutions Group (IBSG), "The internet of things," 2011.

[5] "Number of IoT Devices," 2018, URL: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/ [accessed: [2019-02-26]].

[6] "World Population," 2018, URL: https://www.populationpyramid.net/world/2025/ [accessed: 2019-02-26].

[7] G. C. Kappen, C. Haettich, and M. Meurer, "Towards a robust multi-antenna mass market GNSS receiver," in *Proceedings of the 2012 IEEE/ION Position, Location and Navigation Symposium*, April 2012, pp. 291–300.

[8] E. Kaplan, C. J. Hegarty, *Understanding GPS: Principles and Applications*. Artech House, 2006.

[9] Russian Institute of Space Device Engineering, "Global Navigation Satellite System GLONASS Interface Control Document," 2008.

[10] China Satellite Navigation Office, "Beidou Interface Control Document, Open Service Signal (Version 2.0)," 2013.

[11] European Union, "European GNSS (GALILEO) Open Service, Signal-in-Space Interface Control Document," 2016.

[12] Global Positioning Systems Directorate System Engineering Integration, "Interface Specification IS-GPS-200," 2018.

[13] European GNSS Agency, "GNSS User Technology Report," 2018.

[14] C. Durkovich, "GPS and Critical Infrastructure," 2015.

[15] European Commission, "The European Programme for Cricital Infrastructure Protection (EPCIP)," 2016.

[16] R. H. Mitch, R. C. Dougherty, , M. L. Psiaki, S. P. Powell, B. W. O'Hanlon, J. A. Bhatti, T. E. Humphreys, "Signal Characteristics of Civil GPS Jammers," in *Proceedings of the 24th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2011)*, 2011, pp. 1907–1919.

[17] M. S. Braasch and A. J. van Dierendonck, "GPS receiver architectures and measurements," *Proceedings of the IEEE*, vol. 87, no. 1, pp. 48–64, Jan 1999.

[18] M. L. Psiaki and T. E. Humphreys, "GNSS spoofing and detection," *Proceedings of the IEEE*, vol. 104, no. 6, pp. 1258–1270, June 2016.

[19] B. Walke, *Mobilfunknetze und ihre Protokolle, Mobile Communication Systems and their Protocols*. Teubner, 1988.

[20] J. Xue, S. Biswas, A. C. Cirik, H. Du, Y. Yang, T. Ratnarajah, and M. Sellathurai, "Transceiver Design of Optimum Wirelessly Powered Full-Duplex MIMO IoT Devices," *IEEE Transactions on Communications*, pp. 1955 – 1969, 2018.

[21] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, Jul 1996.

[22] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed. Baltimore: Johns Hopkins University Press, 1989.

[23] "Ettus Homepage," 2014, URL: http://www.ettus.com/ [accessed: 2019-02-26].

[24] T. A. Milligan, *Modern antenna design*. Macmillan, 1985. [Online]. Available: https://books.google.de/books?id=sxUoAQAAMAAJ

[25] "Juce Homepage," 2018, URL: https://juce.com/ [2019-02-26].

[26] "Ettus API," 2014, URL: http://files.ettus.com/manual/page_coding.html/ [accessed: 2019-02-26].

A Low-Voltage Folded-Cascode OP Amplifier with a Dynamic Switching Bias Circuit and Application to Switched Capacitor Filters

Hiroo Wakaumi

Tokyo Metropolitan College of Industrial Technology
Tokyo, Japan
email: waka781420j@ab.auone-net.jp

***Abstract*—Wideband filters employing Operational Amplifiers (OP Amps) are required in sensing devices such as video cameras for environment sensing. A high-speed low-voltage Folded-Cascode (FC) OP Amp with a Dynamic Switching Bias (DSB) circuit capable of processing video signals, which enables low power consumption, high gain with wide bandwidths, and a wide dynamic range, was proposed. Through simulations, it was shown that the OP Amp with the reduced 3-V power supply is able to operate at a 14.3 MHz dynamic switching rate, allowing processing video signals, and a dissipated power of 57 % compared to that in the conventional 5-V power DSBFC OP Amp while keeping a 0.6 V wide output dynamic range. The 2nd-order switched capacitor Low-Pass Filter (LPF) and the 4th-order switched capacitor LPF were tested as its applications. The response of the former was near the theoretical frequency response within frequencies below 5 MHz. The sample-hold circuit in the latter was optimized considering the feed-through phenomenon. The latter showed practical level gains in frequencies over 5 MHz within a stop-band while showing a sharp roll-off near the theoretical frequency response within frequencies below 4 MHz. The power dissipation of either of these switched capacitor LPFs was also reduced to nearly 57 % of that in each switched capacitor LPF with the conventional 5-V power DSBFC OP Amp.**

Keywords-CMOS; operational amplifier; dynamic switching; switched capacitor circuit; filter.

I. INTRODUCTION

This work is an extension of work originally presented in SENSORDEVICES 2018 [1]. Wideband filters are essential for signal processing in video electronic appliances. Specifically, a wideband Low-Pass Filter (LPF) is needed in sensing devices such as a CCD (Charge-Coupled Device) camera with a monitor handling a wide bandwidth video signal of over 2 MHz. The CMOS (Complementary MOS) Switched Capacitor (SC) techniques suitable for realizing analog signal processing ICs (Integrated Circuits), have promising use in video signal bandwidth circuits. It has been demonstrated that SC techniques using CMOS Operational Amplifiers (OP Amps) are useful for implementing analog functions such as filters [2]-[5]. Although CMOS OP Amps are suitable for such filter ICs, the use of several OP Amps results in large power consumption. Especially, the power consumption of OP Amps in high speed operation becomes large because they have wideband properties. There is a possibility that this causes unstable operation.

Until now, several approaches have been considered to decrease the power consumption of OP Amps, including the development of ICs that work at low power supply voltages [6][7]. In the two-stage Folded-Cascode (FC) OP Amp operating at the low power of 1 V [7], resistive biasing and capacitive level shifter are required to increase the output voltage swing. The requirement of four resistors for the resistive biasing makes it difficult to realize as an IC. A clocked current bias scheme for FC OP Amps suitable for achieving a wide dynamic range has typically been proposed to decrease the power consumption of the OP Amp [8][9]. Since the circuit requires complicated four-phase bias-current control pulses and biasing circuits, it is not suitable for the high speed operation and results in a large layout area.

Recently, the author proposed an FC CMOS OP Amp with a Dynamic Switching Bias circuit (DSBFC OP Amp), of simple configuration, to provide low power consumption while maintaining high speed switching operation suitable for processing video signals [10]. This OP Amp operating at the 5-V power supply voltage is not necessarily enough for use in low-voltage signal processing applications under the progress of miniaturization of equipment. That is, the development of OP Amps with a still lower power supply voltage is expected to decrease their power dissipation.

In this paper, a Low-Voltage (LV) DSBFC OP Amp with the 3-V power supply voltage [1] is proposed, which enables low power consumption and is suitable for achieving wide bandwidths and realizing as an IC. The point of view in design for architecture and operation of the LV DSBFC OP Amp is discussed in Section Ⅱ. Simulation results for performance of the LV DSBFC OP Amp are shown in Section Ⅲ. As application examples of this OP Amp, its practicability for a 2nd-order SC Butterworth LPF and a 4th-order SC Butterworth LPF is also evaluated in Section Ⅳ. Finally, conclusions of this work are summarized in Section Ⅴ.

II. LOW-VOLTAGE DSB OP AMP CONFIGURATION

Figure 1 shows a configuration of an LV DSBFC OP Amp, in which the power supply voltages (±1.5 V) were reduced to 60 % of the previous ones (±2.5 V). The DSB method is also used for implementing low power

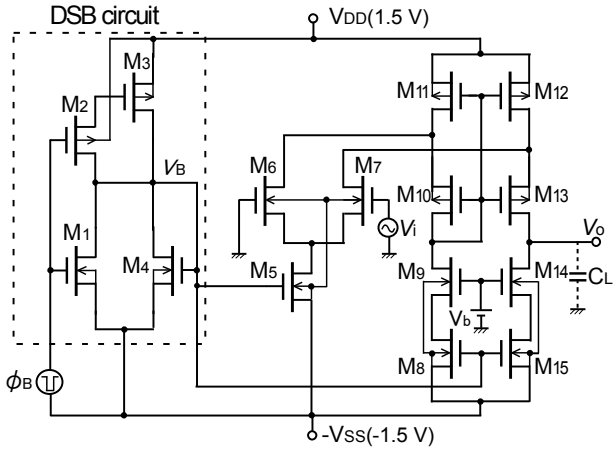


Figure 1. Configuration of the low-voltage DSBFC OP Amp.

TABLE Ⅰ. LOW-VOLTAGE DSBFC OP AMP DESIGNED VALUES.

FET	W/L (μm/μm)	FET	W/L (μm/μm)
M1	15/2.5	M6, M7	2000/2.5
M2	30/2.5	M8, M15	92/2.5
M3	50/4	M9, M14	1055/2.5
M4	44/6	M10, M13	2000/2.5
M5	187/2.5	M11, M12	390/2.5

consumption. When the power supply voltage is simply reduced, the gain of OP Amps is restrained and their bandwidths become low. So, in the newly developed circuit, each FET (Field Effect Transistor) size of the LV DSBFC OP Amp was optimized to achieve high-speed switching operation of 14.3 MHz. This OP Amp has a DSB circuit suitable for low power dissipation and an FC OP Amp to achieve a wide dynamic range even in low power supply voltages. The DSB circuit consists of a bias circuit of M1-M4. The FC OP Amp consists of a current mirror of M10-M13 and current sources of M8, M9, M14, and M15. The current sources M5, M8, and M15 of the FC OP Amp are controlled by switching a bias voltage V_B of the DSB circuit dynamically from V_B* to -1.5 V to reduce the power consumption still more. Table Ⅰ shows its designed values. A minimum channel length of p-MOS FETs and n-MOS FETs is 2.5 μm. In order to achieve almost the same transconductance g_m as that in the conventional 5-V power DSBFC OP Amp, a channel width W of M11 and M12 and that of M10 and M13, in p-MOS current mirror circuits, were increased by nearly fourfold and twofold, respectively. Each W of n-MOS current sources M9, M14, M8, and M15 was increased by nearly one and a half. The bias voltage of V_B* at the on state of the FC OP Amp was adjusted to nearly 0 V (larger than the conventional one) to decrease W of the current source M5 maintaining a high switching speed of the DSB circuit. W of M1-M3 was increased by over twofold. W of M4 was adjusted to an optimum value. The bias voltage Vb of the current source consisting of M9 and M14 was set at 0.15 V.

In the DSB circuit, when an external control pulse ϕ_B driving an inverting switching circuit of M1-M4 is -1.5 V, the OP Amp turns on by setting V_B at an appropriate level of nearly 0 V by enabling M3 and M4 to operate in the saturation region, and operates normally as an operational amplifier. Conversely, when ϕ_B becomes 1.5 V, the OP Amp turns off by setting V_B at nearly -1.5 V, enabling M1 to operate in a low impedance and M3 in a high impedance. This high impedance status of M3 occurs because the gate of M3 is set at a potential determined by the capacitive coupling between gate and source of M2 and between gate and drain of M3 at the transition of ϕ_B to 1.5 V. Therefore, the OP Amp does not dissipate power at all during this off period, which brings about low power consumption.

III. SIMULATION RESULTS

The LV DSBFC OP Amp performance was tested through SPICE simulations. The power supply voltages V_{DD} and Vss are 1.5 V. Typical performances compared with those of the conventional DSBFC OP Amp with a power supply of 5 V are shown in Table Ⅱ. The values of parameters of open loop gain, phase margin, unity gain frequency, slew rate, and settling time are almost the same as those in the conventional 5-V power DSBFC OP Amp. As the inherent static nonlinearity of the LV DSBFC OP Amp, the total harmonic distortion THD for the 10 kHz input signal, enabling 0.6 Vp-p to output, was 0.73 %, which is a little large compared to the conventional one. However, this is less than 1 %.

The LV DSBFC OP Amp operated in a dynamic switching mode with a Duty Ratio (DR) of 50 % and a switching frequency, *fs*, of 14.3 MHz as shown in Figure 2. The output sine wave voltage for the input signal of 1 mV was nearly equal to that in the static operation mode of this OP Amp. Like this, the distortion by the dynamic operation seems to be hardly seen. In the dynamic switching operation mode of 50 % duty ratio, the power dissipation was 9.3 mW, which is 66 % of that (14.0 mW) observed in the static operation mode as shown in Figure 3. This is also 57 % of that (16.3 mW) of the conventional 5-V power DSBFC OP Amp. This shows this OP Amp's extremely low power consumption characteristics due to the reduced effect of power supply voltages (60 % of that in the conventional 5-V power

TABLE Ⅱ. TYPICAL PERFORMANCES FOR THE LOW-VOLTAGE 3-V POWER AND CONVENTIONAL 5-V POWER DSBFC OP AMPS. C_L=1 PF.

Performance parameters	3V power DSBFC OP Amp - this work	5V power DSBFC OP Amp
Power supply voltages	±1.5 V	±2.5 V
Switching frequency *fs*	14.3 MHz	14.3 MHz
Open loop gain Ao	47.1 dB	51 dB
Phase margin θ	32.8 degrees	34.2 degrees
Unity gain frequency *fu*	603.7 MHz	709 MHz
Slew rate SR (C_L=10 pF)	131 V/μs	140 V/μs
Settling time ts	10 ns	12 ns
Distortion THD (*fin*=10 kHz, *Vo*=0.6 Vp-p)	0.73%	0.40%
Power dissipation (C_L=5 pF) in 50 % switching duty ratio	9.3 mW	16.3 mW

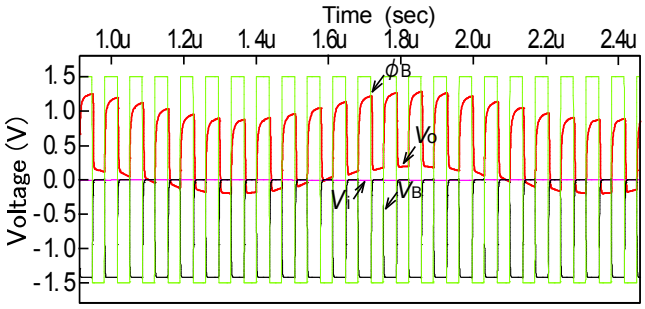


Figure 2. Simulation waveforms of the low-voltage DSBFC OP Amp. Input signal frequency f_{in}=1 MHz, V_{in}=1 mV, f_S=14.3 MHz, C_L=2 pF.

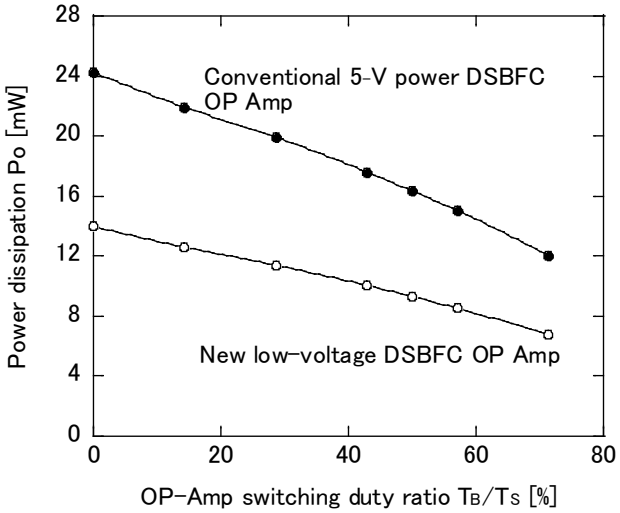


Figure 3. Power dissipation vs. OP-Amp switching duty ratio in the DSB mode. f_S=14.3 MHz, C_L=5 pF.

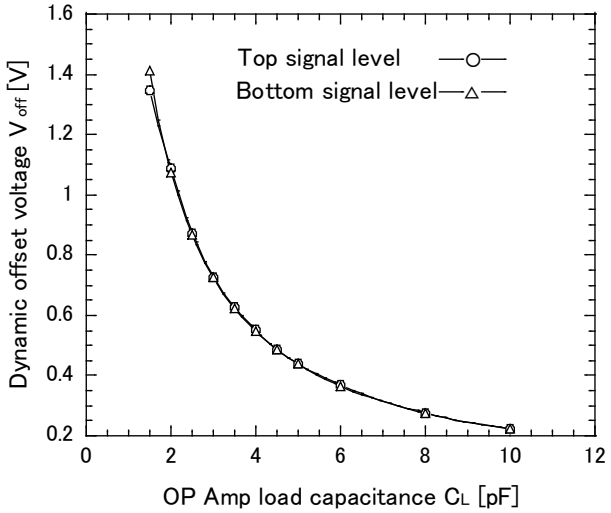


Figure 4. Dynamic offset voltage vs. OP Amp load capacitance in the LV DSBFC OP Amp. f_{in}=0.5 MHz, V_{in}=1 mV.

DSBFC OP Amp) and dynamic switching operation.

The LV DSBFC OP Amp switches dynamically to the off state. At this time, though p-MOSFETs M11 and M12 remain the on-state, MOSFETs M6, M7, M9, M10, M13, and M14 change to the on-state weakly. The output terminal of *Vo* of the OP Amp is set at a voltage depending on the load capacitance through the capacitive coupling between the drain and the gate of the MOSFET M13. So, a large output swing in *Vo* occurs at the off-state transition. A dynamic offset voltage Voff (defined as the difference of the on-state and the off-state output voltages) at the off-state transition of the OP Amp vs. load capacitance C_L was tested (Figure 4). In small load capacitances less than 1.5 pF, top and bottom signal levels of the dynamic off swing do not match. This causes distortion at an output signal of the OP Amp. Therefore, we can see that the load capacitance C_L over 2 pF is desirable for its operation.

IV. APPLICATION TO SC LPFS

To demonstrate the practicability of the above LV DSBFC OP Amp, the feasibility of its application to two kinds of SC LPFs was investigated.

A. 2nd-Order SC LPF

At first, a 2nd-order SC IIR (Infinite Impulse Response) LPF with the Butterworth frequency characteristic was tested. When a sampling frequency f_S=14.3 MHz (equal to four times of NTSC (National Television System Committee) color sub-carrier frequency 3.58 MHz) and a cutoff frequency f_C=2 MHz, respectively, were chosen for this LPF, the discrete-time transfer function using the z-transform is given by (1) [11].

$$H_1(z) = -\frac{0.11735(1+z^{-1})^2}{1-0.82524z^{-1}+0.29464z^{-2}} \quad (1)$$

The circuit configuration and operation waveforms realizing this transfer function are shown in Figures 5 and 6, respectively [11]. This SC LPF was designed referencing an SC biquadratic circuit with integrators in the reference [12]. It consists of a sample-hold circuit with a holding capacitor Cs1 and a CMOS sampling switch controlled by ϕ_{SH}, CMOS switches ϕ_1, ϕ_2 for charge transfer, capacitors A-E, G, and I, and two LV DSBFC OP Amps (OP Amps 1 and 2). In this SC LPF, in order to enable easily to determine the capacitance value of each capacitor, the coefficient of A is set to be equal to that of B. Capacitors in this SC LPF can be basically divided into two groups. In one group (C, D, E, and

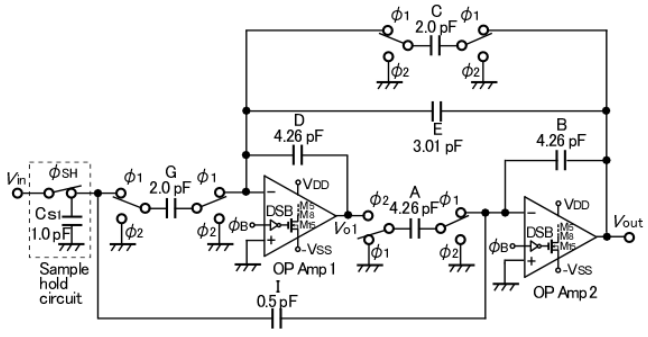


Figure 5. Configuration of the 2nd-order SC LPF.

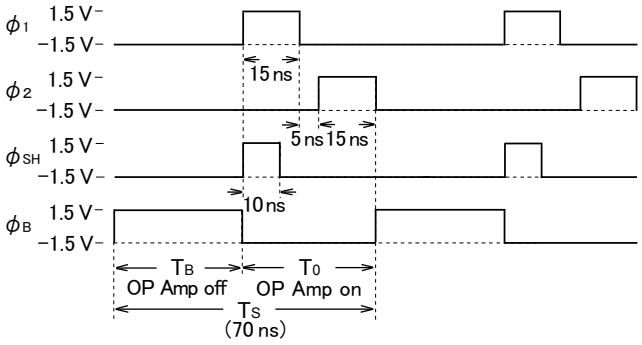


Figure 6. Operation waveforms of the 2nd-order SC LPF.

G), charges are supplied to OP Amp 1. In another group (A, B, and I), charges are supplied to OP Amp 2. Even if a capacitance of each capacitor group is multiplied by constant times, the transfer function of the SC LPF does not change because the multiplication of two capacitor coefficients is performed in the numerator and the denominator of its transfer function, respectively. Therefore, capacitances of integral capacitors B and D are here chosen as a reference capacitance in each group and each coefficient of B and D is normalized to 1. At this time, 1 for every normalized coefficient of A, B, and D is obtained because A and B were set to the same coefficient as described above. In Figure 5, when the coefficients of A, B, and D are normalized to 1, other coefficients are determined in the following.

I=K=0.11735

G=2K+2I=0.4694

C=1+b_1+b_2=0.4694

E=1-b_2=0.70536

Here, b_1=-0.82524 and b_2=0.29464. When the smallest coefficient of I (=0.11735) is replaced as a reference capacitance of 0.5 pF, each capacitance in the SC LPF IC is set in proportion to the above coefficient as shown in Figure 5.

Because an SC LPF's input signal is desirable to be maintained by a sample-hold circuit for stabilizing, this sample-hold circuit is applied to the SC LPF. In this case, the transfer function is multiplied by the following zero-order hold function due to a sample-hold effect.

$$H_s(j\omega)=\frac{\sin(\omega T_s/2)}{\omega T_s/2} \quad (2)$$

Here, Ts represents the one cycle period of sampling pulses. Therefore, when the transfer function (1) is replaced using $z=e^{j\omega T_s}$, the magnitude of the transfer function of the 2nd-order SC LPF considering the sample-hold effect is given by (3).

$$|H(j\omega)|=\frac{\sin(\omega T_s/2)}{\omega T_s/2}\ |H_1(j\omega)| \quad (3)$$

Here, $|H_1(j\omega)|$ in (3) is the following function.

$$|H_1(j\omega)|=\frac{0.2347(1+\cos(\omega T_S))}{\sqrt{1.76783-2.13678\cos(\omega T_S)+1.65128\cos(2\omega T_S)}} \quad (4)$$

The sampling switch was designed to a channel width / channel length W/L=105/2.5 (μm/μm) for each of p-MOSFET and n-MOSFET. Other CMOS switches were designed to 75/2.5 (μm/μm). CMOS switches are turned on and off by non-overlapping two-phase clock pulses ϕ_1, ϕ_2, swinging from -1.5 V to 1.5 V. These sampling and CMOS switches were designed to have a balanced structure with each equal L and W of p-MOS and n-MOS FETs to suppress a feed-through phenomenon as much as possible. This phenomenon is easy to be caused by a capacitive coupling between gate and output terminals.

Major CMOS process parameters are given as a gate insulating film thickness t_{ox}=50 nm, a p-MOSFET threshold voltage V_{TP}=-0.6 V, and an n-MOSFET threshold voltage V_{Tn}=0.6 V.

The operation principle of this SC LPF is as follows. The output signal V_{o1} of OP Amp 1 is obtained as an additional output of an integrated signal of V_{in} using a negative integrator (capacitor D, G SC circuit, and OP Amp 1), an integrated signal of V_{out} using a negative integrator (capacitor D, C SC circuit, and OP Amp1), and a signal multiplied V_{out} by E/D. The output signal V_{out} is an additional output of an integrated signal of V_{o1} using a positive integrator (A SC circuit, capacitor B, and OP Amp 2), and a signal multiplied V_{in} by I/B. V_{out} is basically fed back to an input of OP Amp 1 like this. V_{in} is also integrated twice and added after being decreased by an appropriate capacitance ratio. Due to these integration using positive and negative integrators, addition and feedback operations, the function of LPF is achieved.

The actual operation of the SC LPF is described in the following. In this SC LPF, charge transfer operations through clock pulses ϕ_1, ϕ_2, are performed during the on-state period T_o of the LV DSBFC OP Amps (when the control pulse ϕ_B is set at -1.5 V). The off-state period T_B (the remaining period of the one cycle period Ts) is separately provided to realize low power consumption for this SC LPF. An input signal V_{in} is sampled during the sampling phase of ϕ_{SH} (10 ns) in the on-state period T_o. After sampling operation, its corresponding charge is stored on the holding capacitor C_{s1}. The voltage at the off-state transition of ϕ_{SH} is transferred to an output terminal V_{out}, charging all capacitors, during the clock phase of ϕ_1. During the subsequent clock phase of ϕ_2, each charge of two capacitors C and G is discharged and each charge of remaining capacitors is redistributed. During such on-state period of T_o, the LV DSBFC OP Amps turn on by setting a bias voltage of V_B at an appropriate level of nearly 0 V and operate normally as operational amplifiers.

Subsequently, ϕ_B becomes 1.5 V at the off-state transition of the OP Amps, while ϕ_2 is switched off. During this off period T_B, the OP Amps turn off and so these do not dissipate power at all. Therefore, when T_B is set relatively long as compared to the one-cycle period Ts, the power

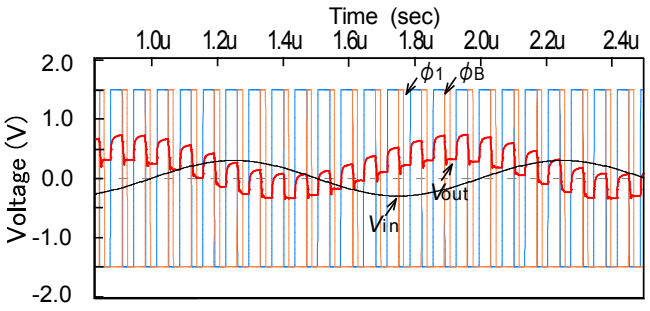


Figure 7. Simulation waveforms for the 2nd-order SC LPF. V_{in}=0.3 V_0-p, f_{in}=1 MHz, C_L=4 pF.

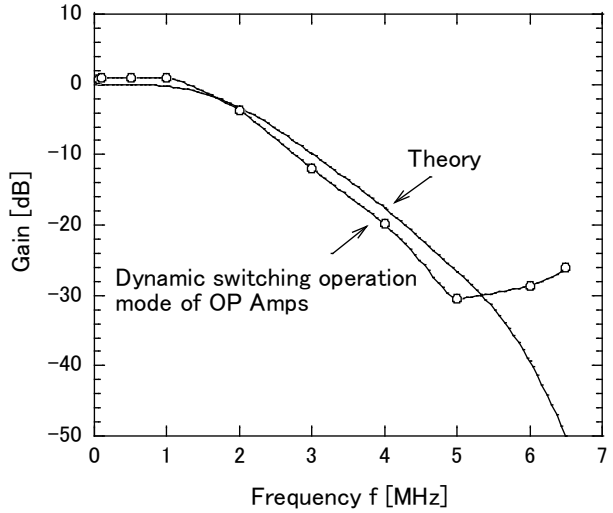


Figure 8. Frequency response of the 2nd-order SC LPF in the DSB mode of the OP Amp. T_B=35 ns.

consumption of the SC LPF is expected to become lower than that observed in ordinary static operation for an SC LPF using conventional OP Amps.

Operation waveforms for an input signal of 1 MHz with an amplitude of 0.3 V and an output load of 4 pF are shown in Figure 7. In the dynamic switching operation, an output load of the LV DSBFC OP Amp increases to nearly 5 pF including internal capacitances of the SC LPF. For the pass-band frequency signal (≦2 MHz), almost the same signal as the input one was obtained. The frequency response of the SC LPF in the dynamic switching operation of the LV DSBFC OP Amp is shown in Figure 8. Figure 9 shows the frequency response of the SC LPF in the static operation mode of OP Amps. Both frequency responses are almost the same. This means that there is almost no gain deterioration by employing the DSB operation of the LV DSBFC OP Amps. The frequency response was near the theoretical one from 100 kHz to near 5 MHz. In the high frequency range over 6 MHz, it deteriorated due to a sampling phase effect. The gain below -26 dB was obtained at over 6 MHz within the stop-band. Though this stop-band gain is not low enough, it is expected to be improved by making the roll-off steeper through the increase of filter order.

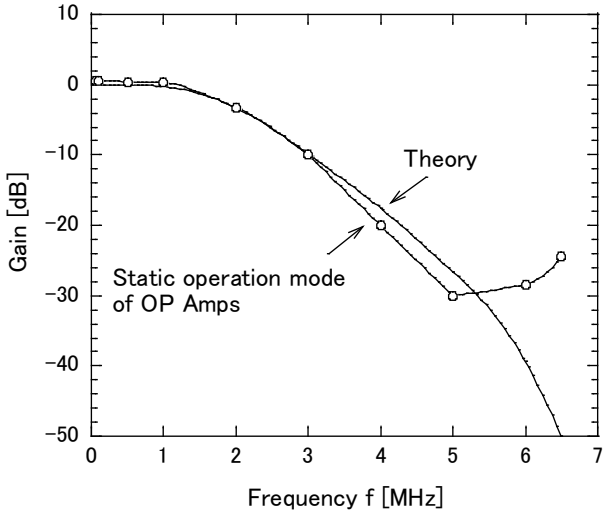


Figure 9. Frequency response of the 2nd-order SC LPF in the static operation mode of the OP Amp. ϕ_B = -1.5 V.

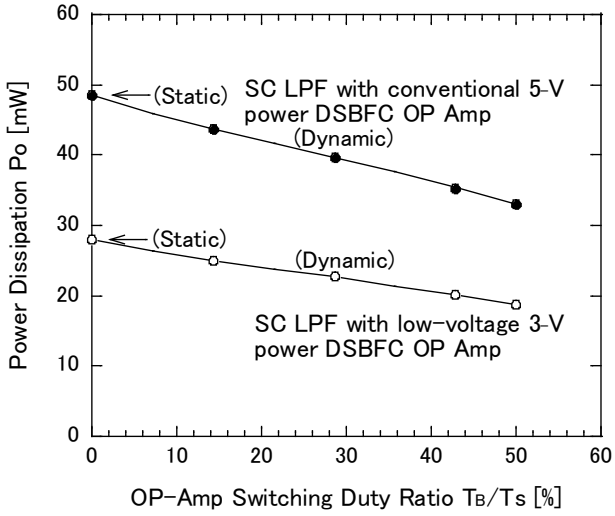


Figure 10. Power dissipation vs. OP Amp switching duty ratio in the 2nd-order SC LPFs. f_{in}=1 MHz.

The power dissipation vs. the OP Amp switching duty ratio in the 2nd-order SC LPF is shown in Figure 10. The power dissipation of the SC LPF itself decreased in proportion to the off period T_B of the OP Amp. In the dynamic switching operation mode of T_B=35 ns (DR=50 %), the power dissipation of the SC LPF (18.7 mW) was reduced to 66.8 % as compared to that in the static operation of the OP Amps (28.0 mW). Thus, the dynamic switching operation of the LV DSBFC OP Amp is useful for reducing the power dissipation of the SC LPF. This power dissipation was 56.8 % compared to that in the SC LPF using the conventional 5-V power DSBFC OP Amp (32.9 mW). This low power characteristic was realized by the low power supply voltages and dynamic switching operation.

B. 4th-Order SC LPF

As another application example of the above LV DSBFC OP Amp, a 4th-order SC IIR LPF with the Butterworth frequency characteristic was tested. The high filter order of the fourth was selected because it is expected to achieve a sharp roll-off gain characteristic. The 4th-order SC LPF was designed to achieve a roll-off characteristic with a gain of below -30 dB at over 4 MHz within the stop-band. The other design condition was set as follows. That is, the sampling frequency *f*s=14.3 MHz and the cutoff frequency *f*c=2 MHz were chosen as the same values as those in the 2nd-order SC LPF, that enables it to process video signals. Under this condition, the discrete-time transfer function is given using the z-transform by (5). That is, (5) can be rewritten like (6) considering the independence of two transfer functions.

$$H_2(z) = \frac{-0.10573(1+z^{-1})^2}{1-0.74578z^{-1}+0.16869z^{-2}} \cdot \frac{-0.13976(1+z^{-1})^2}{1-0.98582z^{-1}+0.54484z^{-2}} \quad (5)$$

$$H_2(z) = H_{21}(z) \cdot H_{22}(z) \quad (6)$$

The circuit configuration realizing this transfer function is shown in Figure 11 [13]. Its operation waveforms are the same as those in the 2nd-order SC LPF shown in Figure 6. The 4th-order SC LPF was designed referencing a SC biquadratic circuit with integrators [12] in the same way as the 2nd-order SC LPF. This SC LPF consists of two-stage biquadratic circuits cascading two 2nd-order SC LPFs of LPF1 and LPF2, provided with a sample-hold circuit with a holding capacitor C_{S1} and a sampling switch controlled by ϕ_{SH}, CMOS switches ϕ_1 and ϕ_2, capacitors A_1, B_1, C_1, D_1, E_1, G_1, I_1, A_2, B_2, C_2, D_2, E_2, G_2, and I_2, and four LV DSBFC OP Amps. The transfer function of this SC LPF circuit is shown in (7).

$$H_2(z) = (-)\frac{D_1I_1+(A_1G_1-2D_1I_1)z^{-1}+D_1I_1z^{-2}}{D_1B_1+(A_1C_1+A_1E_1-2D_1B_1)z^{-1}+(D_1B_1-A_1E_1)z^{-2}} \cdot (-)\frac{D_2I_2+(A_2G_2-2D_2I_2)z^{-1}+D_2I_2z^{-2}}{D_2B_2+(A_2C_2+A_2E_2-2D_2B_2)z^{-1}+(D_2B_2-A_2E_2)z^{-2}} \quad (7)$$

The same way of thinking as that in the 2nd-order SC LPF is applicable in determining each capacitance of LPF1 (A_1, B_1, C_1, D_1, E_1, G_1, and I_1) and LPF2 (A_2, B_2, C_2, D_2, E_2, G_2, and I_2) as well. As shown in (5) and (6), each transfer function for the LPF1 and LPF2 is independent of each other. Therefore, the normalization of the coefficients in the LPF1 and LPF2 can be made independently. In Figure 11, when the coefficients of A_1, B_1, D_1, A_2, B_2, and D_2 are normalized to 1, the transfer function (7) is changed to (8).

$$H_2(z) = (-)\frac{I_1+(G_1-2I_1)z^{-1}+I_1z^{-2}}{1+(C_1+E_1-2)z^{-1}+(1-E_1)z^{-2}} \cdot (-)\frac{I_2+(G_2-2I_2)z^{-1}+I_2z^{-2}}{1+(C_2+E_2-2)z^{-1}+(1-E_2)z^{-2}} \quad (8)$$

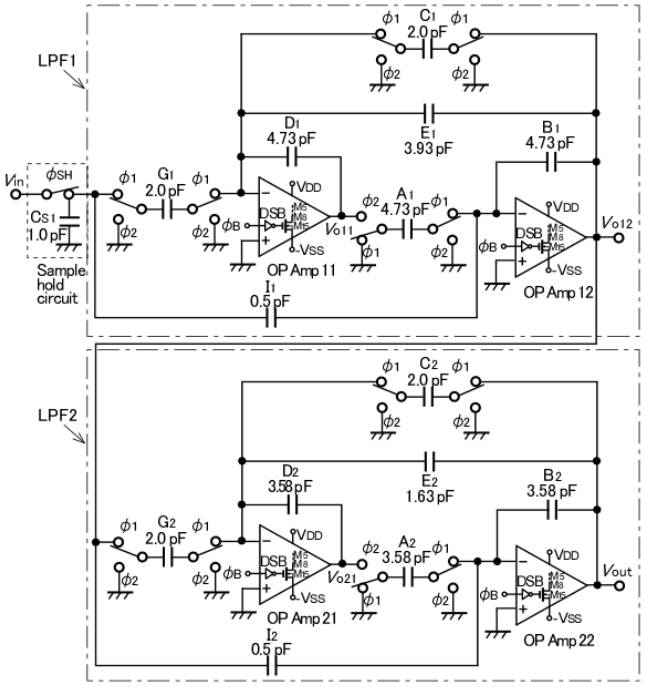


Figure 11. Configuration of the 4th-order SC LPF.

Because the coefficients in (8) are equal to those in (5), other coefficients are determined as follows.

$I_1=K_1=0.10573$

$G_1=4K_1=0.42292$

$C_1=1+b_{11}+b_{12}=0.42291$

$E_1=1-b_{12}=0.83131$

$I_2=K_2=0.13976$

$G_2=4K_2=0.55904$

$C_2=1+b_{21}+b_{22}=0.55902$

$E_2=1-b_{22}=0.45516$

Here, b_{11}=-0.74578, b_{12}=0.16869, b_{21}=-0.98582, and b_{22}=0.54484. When the smallest coefficients of I_1=0.10573 in the LPF1 and I_2=0.13976 in the LPF2 are replaced as a reference capacitance of 0.5 pF, each capacitance in the 4th-order SC LPF IC is set in proportion to the above coefficients as shown in Figure 11. A sample-hold circuit is also applied in this SC LPF to maintain its input signal for stabilizing. At this time, the transfer function (5) is multiplied by the zero-order hold function (2) in the same way as that in the previous 2nd-order SC LPF. Therefore, when the transfer function (5) or (6) is replaced using $z=e^{j\omega T_s}$, the magnitude of the transfer function of the 4th-order SC LPF considering the sample-hold effect is given by

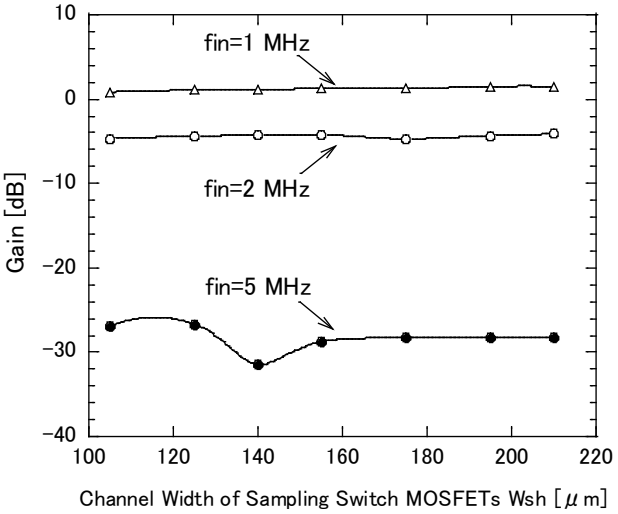


Figure 12. Gain vs. channel width of sampling switch MOSFETs. Cs1=1 pF.

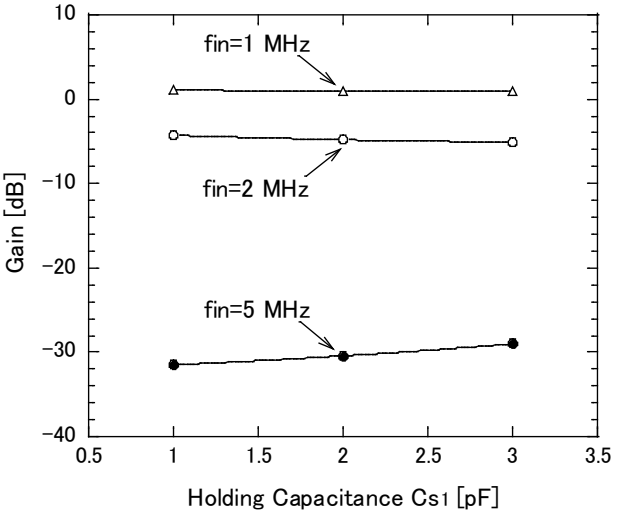


Figure 13. Gain vs. holding capacitance. W/L of sampling switch MOSFETs=140/2.5 (μm/μm).

(9).

$$|H(j\omega)|=\frac{\sin(\omega T_S/2)}{\omega T_S/2}\ |H_{21}(j\omega)|\cdot|H_{22}(j\omega)| \qquad (9)$$

Here, $|H_{21}(j\omega)|$ and $|H_{22}(j\omega)|$ in (9) are the following functions, respectively.

$$|H_{21}(j\omega)|=\frac{0.21146(1+\cos(\omega T_S))}{\sqrt{1.58464-1.74317\cos(\omega T_S)+0.33738\cos(2\omega T_S)}} \qquad (10)$$

$$|H_{22}(j\omega)|=\frac{0.27953(1+\cos(\omega T_S))}{\sqrt{2.26869-3.04587\cos(\omega T_S)+1.08968\cos(2\omega T_S)}} \qquad (11)$$

In this 4th-order SC LPF, the load capacitances of LV DSBFC OP Amps were set as a 2 pF because twofold capacitances as compared to that in the 2nd-order SC LPF are loaded. The sample-hold circuit in this SC LPF was designed

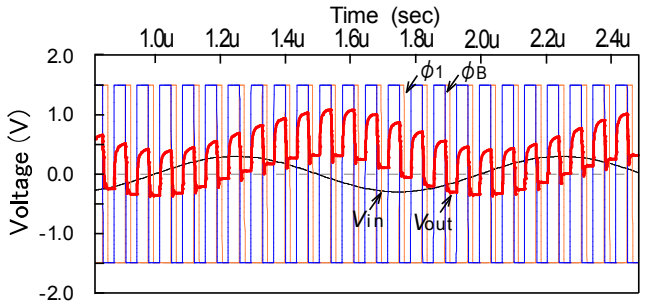


Figure 14. Simulation waveforms for the 4th-order SC LPF. V_{in}=0.3 V_0-p, f_{in}=1 MHz, C_L=2 pF.

as follows considering the feed-through phenomenon.

Figure 12 shows the gain of the 4th-order SC LPF in the DSB mode of the LV DSBFC OP Amp vs. the channel width Wsh of each of p-MOSFET and n-MOSFET in the sampling switch. The gain of the 4th-order SC LPF became minimum at a Wsh of nearly 140 μm when the input signal frequency *f*in is equal to 5 MHz within the stop-band, while its gain remains almost unchanged for input signals of 1 and 2 MHz within the passband. This is thought to be due to the following phenomenon. When Wsh is larger than 140 μm, the feed-through via the difference of capacitive coupling between gate and output terminals of the above MOSFETs does not become negligible at the off-state transition of the sampling switch and so the gain corresponding to *f*in=5 MHz increases. When Wsh is smaller than this value, a driving ability of the sampling switch becomes insufficient, which brings about an increase of the gain. Like this, Wsh of the sampling switch is optimized to 140 μm. The feed-through in the sample-hold circuit is also dependent on a holding capacitance. Figure 13 shows the gain of the 4th-order SC LPF in the DSB mode of the LV DSBFC OP Amp vs. the holding capacitance. As the holding capacitance Cs1 increases, the gain corresponding to *f*in=5 MHz in the stop-band deteriorates little by little. That is, we can see that a smaller capacitance is desirable as Cs1. So, the Cs1 of 1 pF in this 4th-order SC LPF was also chosen.

Other CMOS switches were designed to 75/2.5 (μm/μm), which is the same one as that in the 2nd-order SC LPF. In this 4th-order SC LPF consisting of a cascade connection of two different 2nd-order SC LPFs of LPF1 and LPF2 and a sample-hold circuit for common use, the operation principle is similar to the previous 2nd-order SC LPF

Operation waveforms for an input signal of 1 MHz with an amplitude of 0.3 V and an output load of 2 pF are shown in Figure 14. In the 4th-order SC LPF, the output signal amplitude nearly close to the input signal was also obtained for passband frequency signals. The frequency response of the 4th-order SC LPF in the dynamic switching operation of the LV DSBFC OP Amps is shown in Figure 15. The roll-off characteristic in near 3-4 MHz was greatly improved compared to that in the 2nd-order SC LPF. The response was near the theoretical one from 100 kHz up to near 4 MHz. At 4 MHz within the stop-band, the gain below -28 dB was obtained. In the high frequency range over 5 MHz within the stop-band, although it deteriorated due to a sampling phase

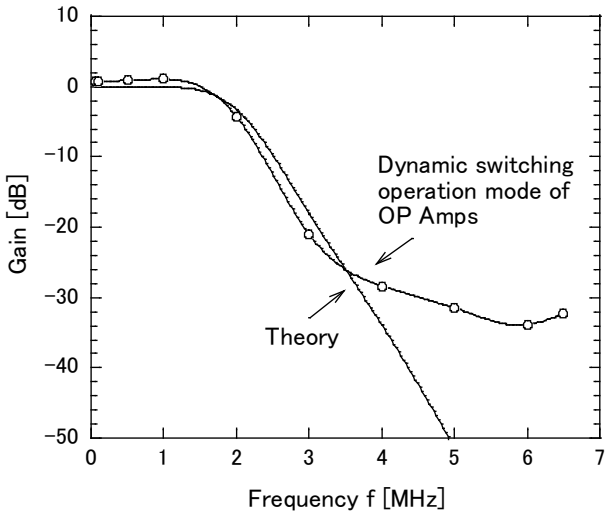


Figure 15. Frequency response of the 4th-order SC LPF in the DSB mode of the LV DSBFC OP Amp. T_B=35 ns.

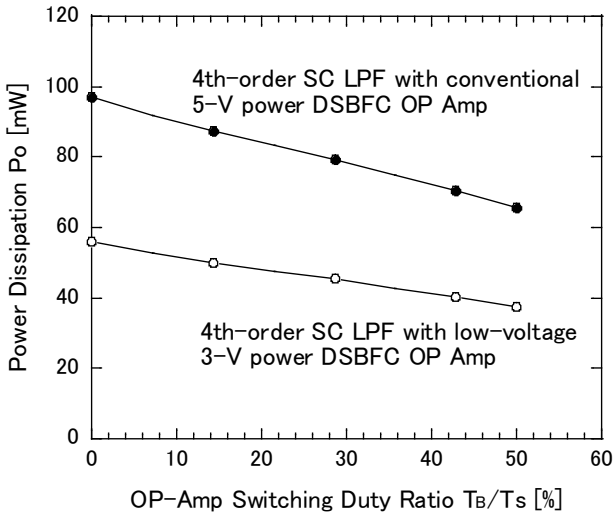


Figure 16. Power dissipation vs. OP Amp switching duty ratio in the 4th-order SC LPFs. f_{in}=1 MHz.

effect, the gain below -31.5 dB (a practical level) was achieved. In this way, a wide stop-band with a high attenuation (a sharp roll-off characteristic) in the high frequency response became possible due to the two-stage biquadratic SC LPF configuration with the increased filter order of the fourth. Like this, it is clear that the LV DSBFC OP Amp is also applicable to the high-order SC LPF.

The power dissipation vs. the OP-Amp switching duty ratio in the 4th-order SC LPF with the 3-V power LV DSBFC OP Amps compared to that in the 4th-order one with conventional 5-V power DSBFC OP Amps is shown in Figure 16. The power dissipation of the 4th-order SC LPF with the LV DSBFC OP Amps itself decreased in proportion to the off-state period T_B of the OP Amps. In the dynamic switching operation mode of T_B=35 ns (=50 % switching duty ratio) and $\phi_1 = \phi_2$ =15 ns, the power consumption of this 4th-order SC LPF (37.4mW) decreased to 66.8 % as

TABLE Ⅲ. TYPICAL PERFORMANCES FOR THE 2ND-ORDER AND 4TH-ORDER SC LPFS EMPLOYING THE LV DSBFC OP AMPS.

Performance parameters	Simulation results	
	4th-order SC LPF	2nd-order SC LPF
Sampling and switching frequency fs	14.3 MHz	14.3 MHz
Input signal amplitude	0.3 $V_{0\text{-}P}$	0.3 $V_{0\text{-}P}$
Cutoff frequency fc	2 MHz	2 MHz
Gain at a stop-band over 5 MHz	≦-31.5 dB (Dynamic mode)	≦-26.0 dB (Dynamic mode)
Power consumption	56.0 mW (Static)	28.0 mW (Static)
	37.4 mW (Dynamic:T_B/T_S=50 %)	18.7 mW (Dynamic:T_B/T_S=50 %)

compared to that in the static operation of the LV DSBFC OP Amps (56 mW). This value is twice as large as that in the 2nd-order SC LPF with the LV DSBFC OP Amps (18.7 mW) because the 4th-order SC LPF consists of a cascade-connection of two 2nd-order SC LPFs. However, the power consumption of this 4th-order SC LPF with the LV DSBFC OP Amps at 50 % switching duty ratio was reduced to 56.9 % compared to that in the 4th-order SC LPF with conventional 5-V power DSBFC OP Amps (65.7 mW). Thus, even when the DSBFC OP Amps are applied to the two-stage biquadratic circuits of SC LPF, the dynamic operation of these LV DSBFC OP Amps enabling low power consumption as compared to their static operation is also useful for reducing the power consumption of SC LPF. Typical characteristics of the 4th-order SC LPF compared with those of the 2nd-order SC LPF employing the LV DSBFC OP Amps are listed in Table Ⅲ.

V. CONCLUSIONS

A high-speed Low-Voltage (LV) Folded-Cascode (FC) Operational Amplifier (OP Amp) with a Dynamic Switching Bias (DSB) circuit capable of processing video signals, which enables low power consumption, high gain with wide bandwidths, and a wide dynamic range, was proposed. Through simulations, it was shown that the OP Amp with the reduced 3-V power supply is able to operate at a 14.3 MHz dynamic switching rate, allowing processing video signals, and a dissipated power of 57 % compared to that in the conventional 5-V power DSBFC OP Amp while keeping a 0.6 V wide output dynamic range. The response of the 2nd-order Switched Capacitor Low-Pass Filter (SC LPF) tested as its application was near the theoretical frequency response within frequencies below 5 MHz. The power dissipation of this LPF was also reduced to 56.8 % of that in the 2nd-order SC LPF with conventional 5-V power DSBFC OP Amps. As another application of the LV DSBFC OP Amps, the 4th-order SC LPF was tested to achieve a practical sharp roll-off gain characteristic. The roll-off response of this 4th-order SC LPF was greatly improved compared to that in the 2nd-order SC LPF. At over 5 MHz within the stop-band, a practical level gain below -31.5 dB was achieved. The power consumption of this 4th-order SC LPF decreased to 56.9 % of that in the 4th-order SC LPF with conventional 5-V power DSBFC OP Amps.

Thus, it was confirmed that the 3-V power DSBFC OP

Amp is useful for high speed operation, low power consumption, and greatly reducing the power dissipation of the SC LPFs. This circuit should be useful for the realization of low-power wide-band signal processing ICs. For example, it has a possibility of changing current A/D (Analog-to-Digital) converters, D/A (Digital-to-Analog) converters, and active-RC LPFs in CCD cameras into low-power SC circuits employing this LV DSBFC OP Amp, and of realizing low-power SC versions instead of current SC Amps for prevention of the deterioration of CCD camera's output signal. It has also applicability to Amps and/or LPFs required in various kinds of sensing devices and video equipment. Furthermore, it is noteworthy that the performance is expected to be improved still more by employing MOSFETs with a minimum shorter channel length than 2.5 μm used in this work.

As shown above, although the frequency response was improved to a practical level by increasing the filter order, there might be a limitation in the filter order, because one OP Amp per one order LPF must be used and so power dissipation will increase in proportion to the filter order.

REFERENCES

[1] H. Wakaumi, "A Low-Voltage Folded-Cascode OP Amplifier with a Dynamic Switching Bias Circuit," The Ninth International Conference on Sensor Device Technologies and Applications (SENSORDEVICES 2018), Sept. 2018, pp. 60-64.

[2] R. Gregorian and W. E. Nicholson, "CMOS Switched-Capacitor Filters for a PCM Voice CODEC," IEEE J. Solid-State Circuits, vol. SC-14, no. 6, pp. 970-980, Dec. 1979.

[3] R. Dessoulavy, A. Knob, F. Krummenacher, and E. A. Vittoz, "A Synchronous Switched Capacitor Filter," IEEE J. Solid-State Circuits, vol. SC-15, no. 3, pp. 301-305, June 1980.

[4] A. Iwata, H. Kikuchi, K. Uchimura, A. Morino, and M. Nakajima, "A Single-Chip Codec with Switched-Capacitor Filters," IEEE J. Solid-State Circuits, vol. SC-16, no. 4, pp. 315-321, Aug. 1981.

[5] J.-T. Wu, Y.-H. Chang, and K.-L. Chang, "1.2V CMOS Switched-Capacitor Circuits," 1996 IEEE International Solid-State Circuits Conference Digest of Technical Papers (42nd ISSCC), pp. 388-389, 479.

[6] Z. Kun, W. Di, and L. Zhangfa, "A High-Performance Folded Cascode Amplifier," 2011 International Conference on Computer and Automation Engineering (ICCAE 2011), IPCSIT, vol. 44, 2012, pp. 41-44.

[7] S. H. Mirhosseini and A. Ayatollahi, "A Low-Voltage, Low-Power, Two-Stage Amplifier for Switched-Capacitor Applications in 90 nm CMOS Process," Iranian J. Electrical & Electronic Engineering, vol. 6, no. 4, pp. 199-204, Dec. 2010.

[8] D. B. Kasha, W. L. Lee, and A. Thomsen, "A 16-mW, 120-dB Linear Switched-Capacitor Delta-Sigma Modulator with Dynamic Biasing," IEEE J. Solid-State Circuits, vol. 34, no. 7, pp. 921-926, July 1999.

[9] H.-L. Chen, Y.-S. Lee, and J.-S. Chiang, "Low Power Sigma Delta Modulator with Dynamic Biasing for Audio Applications," 50th Midwest Symp. Circuits and Systems 2007 (MWSCAS 2007), pp. 159-162.

[10] H. Wakaumi, "A Folded-Cascode OP Amplifier with a Dynamic Switching Bias Circuit," Engineering Letters, vol. 23, issue 2, pp. 92-97, June 2015.

[11] H. Wakaumi, "A Switched-Capacitor Low-Pass Filter with Dynamic Switching Bias OP Amplifiers," Advances in Science, Technology and Engineering Systems J., vol. 2, no. 6, pp. 100-106, Nov. 2017.

[12] T. Takebe, A. Iwata, N. Takahashi, and H. Kunieda, Switched Capacitor Circuit, Tokyo: Gendai Kohgaku-Sha, Apr. 2005.

[13] H. Wakaumi, "A Fourth-Order Switched-Capacitor Low-Pass Filter with a Dynamic Switching Bias OP Amplifiers," 6th International Conference on Advanced Technology & Sciences (ICAT'Riga), Sept. 2017, pp. 147-151.

Seismic Observation and Structural Health Monitoring of Buildings by Improved Sensor Device Capable of Autonomously Keeping Accurate Time Information

Narito Kurata
Faculty of Industrial Technology
Tsukuba University of Technology
Tsukuba City, Ibaraki, Japan
e-mail: kurata@home.email.ne.jp

Abstract - **In this research, sensor devices were developed for application to seismic observation to understand damage conditions after earthquakes and to structural health monitoring for the maintenance of buildings and civil infrastructures. To apply the sensor devices, they must be densely installed in a broad area and measurement data with synchronized time must be obtained. It is desirable that the sensor devices themselves keep accurate time information even in environments with no available network or Global Positioning System (GPS) signals. Therefore, a sensor device was developed that keeps accurate time information autonomously using a Chip Scale Atomic Clock (CSAC), which consumes ultra-low power, can be mounted on a small board, and is an ultra-high precision clock. This paper explains the CSAC and a mechanism to add highly accurate time information to the measured data using the CSAC. Next, the paper discusses the process of development from prototype to practical device as well as improvement results to solve challenges identified in the actual use at a bridge. In this paper, a new procedure for time synchronization between devices is described. In addition, the communication system of measurement data newly constructed using "fluentd" which is open source software for data collection is detailed. Finally, the paper demonstrates the usability of the developed sensor device using a case study where seismic observation and structural health monitoring was implemented by installing the improved devices in an actual building. In particular, structural health monitoring of the building based on the evaluation by the inter-story deformation was made possible by the practical device developed in this research securing autonomous time synchronization.**

Keywords-Time Synchronization; Chip Scale Atomic Clock; Earthquake Observation; Structural Health Monitoring; Acceleration Sensor.

I. INTRODUCTION

Due to degradation of buildings and civil infrastructures, such as bridges, and highways over time, automation of inspection for their maintenance and management is an urgent social issue. Also, since there are many earthquakes in Japan, it is required to detect the damage of the structure immediately after the earthquake and to grasp the situation of the urban damage. In order to detect those abnormal situations, data collection and analyses by a group of sensors are necessary [1]. Sensors were developed for seismic observation and structural health monitoring applying wireless sensor network technology, and their performance in a skyscraper was demonstrated [2]-[4]. One important challenge in this research was time synchronization among sensors. To analyze a group of data measured by multiple sensors and assess structural safety, time synchronization among the sensors must be kept. In the wireless sensor network system, the time synchronization was materialized through transmission of wireless packets among the sensors [4]. However, the wireless sensor network technology is not practically applied to multiple buildings, long-span structures, such as bridges, or broad urban spaces. If sensors installed in various locations are capable of keeping accurate time information autonomously, this problem can be solved. Using Global Positioning System (GPS) signals is effective for outdoor situations, but it is not available inside buildings, underground, under bridges, or in tunnels. Therefore, a prototype sensor device capable of maintaining accurate time information autonomously was developed using a Chip Scale Atomic Clock (CSAC) [5]-[7], which is a high-precision clock and very accurate compared with crystal resonators [8][9]. Then, the prototype device was upgraded for higher functionality and practical application to develop a practical device [10].

In addition, in order to apply the developed sensor device to earthquake observation, logic to detect the occurrence of an earthquake and store data of only earthquake events was implemented and confirmed its function in shaking table experiment [11][12]. These tests confirmed the performance of the sensor device that can maintain accurate time information autonomously and showed that the device is applicable to seismic observation and structural health monitoring of buildings and civil infrastructures.

In this article, Section II shows the existing time synchronization methods and describes their problems and achievement of the development of sensor device proposed in this study. Section III describes CSAC and explains the mechanism for providing ultra-high accurate time information to sensor data by the CSAC. Section IV describes the development of a practical module from its prototype. Section V lists problems that were extracted when the practical modules were installed on an actual bridge, and describes details of improvements made to cope with these problems. Further, Section VI shows a new procedure for time synchronization between devices. The communication system of measurement data newly constructed using "fluentd" which

is open source software for data collection is detailed in Section VII. Finally, Section VIII shows an example of seismic observation and structural health monitoring by applying the developed sensor device to a real building. In order to obtain the inter-story deformation of the building for the purpose of structural health monitoring shown in this paper, time synchronization of sensors is required. The sensor device developed in this research can hold accurate absolute time information autonomously, so it is easy to secure time synchronization of many sensor devices without wiring or network.

II. State of the Art

A time synchronizing function is indispensable for sensor devices that are used for seismic observation and structural health monitoring. Unless a data group where time synchronization is ensured is obtained, a time history analysis employing phase information cannot be made. For example, it is difficult to clarify a phenomenon where seismic waves propagate through the ground. Moreover, it is not possible to make a modal analysis or an analysis for damage evaluation of a structure. Many studies have been carried out so far in relation to the time synchronized sensing, including the GPS that makes use of a radio clock or a satellite, and the Network Time Protocol (NTP) [13] designed for time synchronization on the internet. There are also studies where time synchronization is realized by making use of the characteristics of a radio sensor network where a propagation delay is small. For example, time-synchronizing protocols have been studied, which include Reference Broadcast Synchronization (RBS), Timing-sync Protocol for Sensor Networks (TPSN), and Flooding Time synchronization Protocol (FTSP) [14]-[18]. However, although these time synchronizing technologies are widely used even now, they cannot constitute an optimum means for sensor devices for use in seismic observation and structural health monitoring. Specifically, the GPS cannot be used inside a building, and the time synchronizing accuracy of the NTP is not sufficient. The time synchronizing method employing the radio technology is highly useful, but it is not ensured that the radio communication is always available. In particular, if the wireless communication is interrupted at the time of an earthquake, time synchronization cannot be performed.

In this study, a prototype of a sensor module for autonomously keeping accurate time information is developed by making use of a CSAC that is an ultra-high accurate clock, and an improvement is carried out on the prototype for a practical application. Even though a tremendous number of sensors are installed in the buildings and civil infrastructures, in case accurate time information can autonomously be given to the data measured by those sensors, time synchronization can be ensured between the sensors only by collecting the data using an arbitrary means and by realigning the data utilizing the time information. The data group where the time synchronization is ensured by using the sensor device proposed in this paper is available for an analysis intended to grasp a seismic phenomenon or evaluate the damage of a structure.

III. Time Stamping Mechanism Using Chip Scale Atomic Clock

A CSAC has time accuracy equivalent to that of a rubidium atomic clock and is very accurate compared with crystal resonators [5][7]. The CSAC can achieve ultra-precision time measurement at a level of some ten picoseconds, consumes low power and is small enough to be mounted on a circuit board (Table I). The development of the CSAC started with the support of Defense Advanced Research Projects Agency, and the commercial product was released by an American company in 2011 and is still available for purchase. Recently, ultra-small atomic clock systems, which can be mounted on general communication terminals, such as smart phones, have been proposed, and further downsizing and price reduction are expected. If the sensor device is equipped with a CSAC and a mechanism that adds time stamping for every sample of measured data, the sensor device can create data having high-accuracy time information. Each sensor device autonomously keeps highly accurate time information even if the GPS signals and network communication are unavailable. Therefore, by collecting the measured data by means, such as 3G, Wi-Fi, Ethernet, etc., a data group ensuring time synchronization can be obtained.

To configure a sensing system composed of multiple sensor devices equipped with a CSAC, one device is set as a master device and other devices as slave devices must be synchronized by defining absolute time information. The main controller of each sensor device is equipped with an input/output connector for 1 Pulse Per Second (PPS) of the CSAC. Using this connector, the master device outputs 1 PPS signal, and each slave device inputs it to synchronize and match the phase of the CSAC in each slave device. The CSAC keeps accurate time, but it does not have absolute time information. Therefore, it must be defined separately. At initial settings, the GPS module installed in the main controller is used. Absolute time information is transmitted from the master device to the slave device by the IEEE 1588 standard. Once all the sensor devices are synchronized at the beginning, they continue keeping highly accurate time information autonomously. It is only necessary to install the sensor device in an arbitrary place and collect data. As mentioned above, any means of data collection, such as Ethernet, Wi-Fi, or 3G, are available as the measured data records accurate time stamping.

TABLE I. Specifications of CSAC

Model	SA.45s
RF output	10 MHz
1 PPS output	Rise/fall time: < 10 ns Pulse width: 100 µs
Power consumption	< 120 mW
Outside dimensions (mm)	40 × 35 × 12
Frequency accuracy	$\pm 5 \times 10^{-11}$
Aging	$< 9 \times 10^{-10}$/month

The sensors are also suitable for use as mobile measurement and a mobile sensing system because the sensors can measure and collect data even if GPS signals are not available, and a wireless or wired network cannot be used.

IV. Development of Sensor Device Equipped with CSAC

The general sensor device is composed of a sensor chip, CPU, filter, A/D converter, memory, and network interface, and a crystal oscillator is used as the clock of the CPU. If CSAC is installed in the sensor device and measurement is performed while correcting the clock of the CPU, a delay occurs because CSAC's clocking accuracy is too high. Therefore, a mechanism having a special Field Programmable Gate Array (FPGA) was developed to add time information from the CSAC to the data measured by the sensor directly. The FPGA not only adds the time information of the CSAC to the measured data but can also incorporate logic, such as seismic detection. A prototype device was developed first to identify challenges, and then a practical device that has solved the challenges was developed.

A. Prototype Device

Fig. 1 shows the developed prototype device [8][9]. The prototype device is composed of a mainboard, sensor board, and wireless communication board. The mainboard incorporates a CSAC, FPGA, CPU, memory, network interface, etc. The sensor board is detachable. Two types of sensor board were developed, one of which is an acceleration sensor board equipped with a microelectromechanical systems (MEMS) accelerometer, temperature sensor, anti-aliasing filter, A/D converter, etc., and the other is an external sensor board that can connect an analog sensor externally via the Bayonet Neill-Concelman (BNC) connector (Fig. 2). The communication board is also detachable and can collect data using wireless network once mounted on the mainboard. Two types of dedicated board equipped with Wi-Fi or 3G were developed.

B. Practical Device

The following improvements were made for high functionality and practical use of the prototype device [10]-[12].

1) *3-channeled external analog sensor input interface*
2) *24-bit A/D converter*
3) *Enhanced FPGA*
4) *Separate wireless communication to use commercially available Raspberry Pi*
5) *Time synchronization by IEEE 1588*

The practical device is composed of a board and Raspberry Pi having a wireless communication function (Figs. 3 and 4) and is enclosed in a dedicated case (Fig. 5). As shown in Fig. 3, the board of the practical device is composed of a main control unit and a sensor unit. The main control unit is equipped with a CSAC, FPGA, GPS, CPU, memory, and network interface (Table II). The main control unit controls measurement of the sensor while producing time stamping, based on highly accurate time information from the CSAC. The device sends the measurement data via Ethernet or wireless communication to the network after saving data in an SD card. The device saves two types of data, one of which is regularly measured data and the other is extracted seismic event data. The device sends data containing seismic events solely to the network immediately after an earthquake using the FPGA to detect start and end of the earthquake. A GPS is installed in the device to initialize and adjust the time information. The sensor takes measurements following a command from the main control unit. The sensor is equipped with a 3-axis MEMS accelerometer, 3-channel external analog sensor input interface, temperature sensor, anti-aliasing filter, and A/D converter. The wireless communication uses commercially available Raspberry Pi to enable data collection with wireless communication. Either Wi-Fi or 3G is selectively used for the wireless communication.

Compared with the prototype device, the practical device incorporates a 3-channel external analog sensor input interface and 24-bit A/D converter, so it can connect with a sensor requiring a wide dynamic range, such as a servo accelerometer. In addition, the sensor device can be used as a data logger by connecting with three strain sensors, displacement centers, etc. Using Raspberry Pi for the wireless communication, the device can quickly respond to new wireless communication formats. Furthermore, an interface of IEEE 1588 standard for time synchronization of network was incorporated to initialize measurement timing among sensor devices and synchronize them.

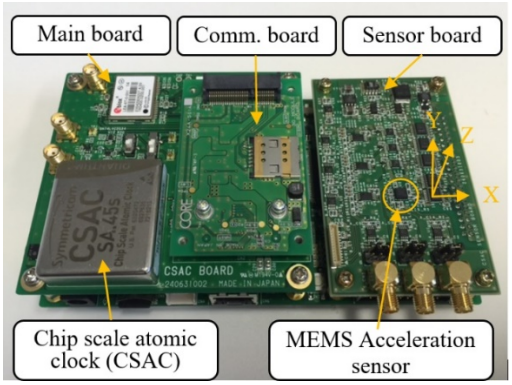


Figure 1. Prototype Device.

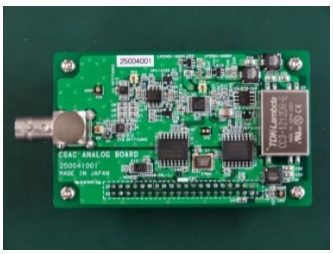


Figure 2. External Sensor Board.

V. IMPROVEMENT OF PRACTICAL DEVICE

The developed practical device was installed onto an actual bridge, and measurement was performed for three months to identify challenges in practical use. Based on those challenges, the practical device was improved for higher stability and operability, and the specified performance was confirmed [1]. Improvement items and contents are as follows.

A. Reduction in Built-in SD Card Access

To avoid failures caused by total service life consumed by the number of rewrite cycles resulting from the quality of the SD card (micro SD card) or compatibility issues, the timing to write in the file system was reviewed. By setting the interval of the write timing to the file system from 5 seconds to 240 seconds, access is suppressed to a maximum of 1/48, as shown in Table III.

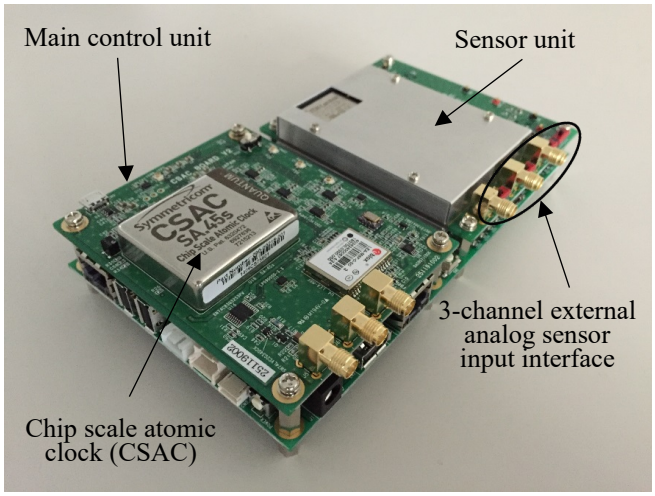


Figure 3. Board Configuration of the Practical Device.

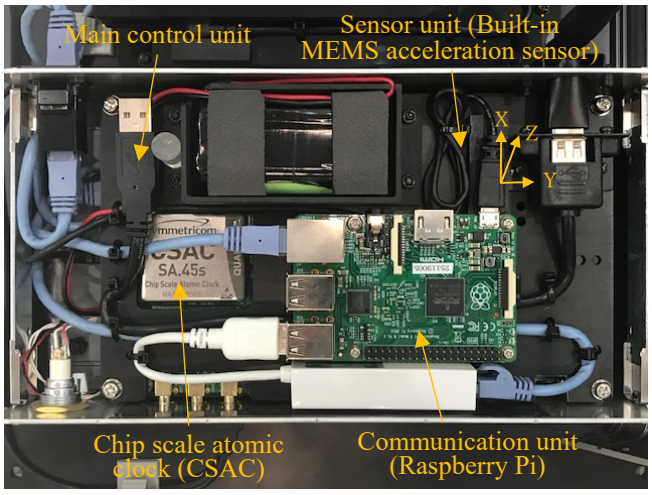


Figure 4. Board Configuration of the Practical Device in Dedicated Case.

However, risk of data loss during emergency black out increases, which means it is a trade-off. Table III shows the effects of reduction in access to the built-in SD card.

Figure 5. Dedicated Case for Practical Device.

TABLE II. SPECIFICATION OF CONTROL UNIT

CPU	RZ/A1L CPUcore:ARM Cortex A9 with Neon 384MHz (Max) Instruction / data cache: 32KB/32KB L2 cache: 128KB 128MHz
RAM (SDRAM)	128Mbyte CL=2 @64MHz (16bit Bus) x2 Micron:MT48LC32M16A2P-75IT
RAM (Internal RAM)	3Mbyte (RZ/A1L)
ROM (serial-FLASH)	64Mbyte Serial Multi I/O x1 Spansion: S25FL512SDPMFIG11 (64Mbyte)
Ethernet	10/100BASE-PHY
FPGA	XC6SLX45
CSAC	SA.45s CSAC Power consumption: <120mW Volume: <17 cm^3 Frequency Accuracy: Max. $\pm 5x10^{-11}$ Short term Stability(@1000s): < $1x10^{-11}$ Aging Rate (Typical): < $9x10^{-10}$ /month Operating Temperature: -10ºC to +70ºC
GPS	LEA-M8F
Battery	Lithium-ion rechargeable battery (3.7V) 2 cell
Outside dimension	180 × 100 mm

TABLE III. EFFECTS OF REDUCTION IN ACCESS TO BUILT-IN SD CARD

	Current	After improvement
Frequency of writing	Once per 5 second	Once per 240 second
Data loss during emergency blackout with sampling rate of 100 Hz	Approx. 500 sampling	Approx. 24000 sampling

B. *File Compression of Measured Data*

In the actual work installed onto the bridge, transmission time for the measured data must be reduced. To reduce the data transmission time and achieve storing long-term measurement data, a file compression function was added. From the viewpoint of durability, it is desirable to use a relatively small capacity SD card, which is considered to be manufactured in a stable production line rather than a large capacity one which might be manufactured in a premature production line. The file compression function is important as a means to materialize long-term data recording in a small capacity SD. The software was improved to compress the measured data to be recorded in the SD card after 75 minutes.

C. *Change of Local PC Collecting Display*

A tool was developed to collect the measured data saved in the built-in SD card of the practical device placed in the LAN, which was temporarily installed on the site into a local PC using a browser. However, there was a possibility that time information not related to measurement data might be displayed depending on the startup environment in the measurement data download screen (Fig. 6). Specifically, by adjusting the time of Raspberry Pi with the FPGA on the baseboard, improvements were made so that the correct time information of the measurement data is displayed regardless of the activation environment. However, even without this improvement, there is no effect on the high precision time information recorded in the measurement data.

D. *Correction of PTP Control Software*

The Precision Time Protocol (PTP) defined in the IEEE-1588 standard is a means to synchronize time of computers on a LAN highly accurately. The CSAC is a high precision clock, but it does not keep absolute time information. Therefore, clock adjustment was made with the CSAC and then the absolute time was set to the timer counter of the FPGA by the PTP in the practical device as described in detail in Section VI.

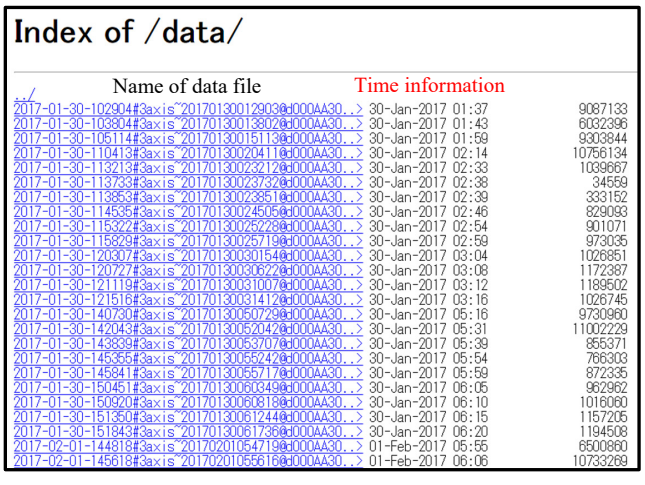


Figure 6. Local PC collecting tool display screen.

In the actual work at the bridge, there was a risk that the absolute time set by the PTP may be initialized due to plugging and unplugging of the LAN cable. Such phenomenon is not desirable for the practical device although it follows the standard implementation and specifications of Linux as the PTP must be operated with a LAN cable connected. Therefore, the settings were corrected so that the absolute time would not be initialized and would be kept between two practical devices (master device and slave device) whose absolute time had been set by the PTP even if the Ethernet cable was plugged and unplugged.

E. *Minimization of Wiring Delay in the FPGA*

Depending on the circuit design of the FPGA, the time stamping may not be correctly recorded due to fluctuation of timing to read and write the memory from the FPGA. The wiring delay of the SDRAM control signal in the FPGA was minimized so that the timing to read and write the memory from the FPGA would not fluctuate. The FPGA circuit was redesigned and implemented in the new practical sensor device. It was confirmed that the continuous operation test for 2 months was conducted for the four devices and the time information was correctly displayed.

VI. TIME SYNCHRONIZATION PROCEDURE AMONG PRACTICAL DEVICES

There are two methods for absolute time synchronization to the practical devices. The first is to synchronize each practical device directly with GPS by absolute time. The second is to synchronize one device (master device) with absolute time by GPS. Thereafter, other devices (slave devices) are synchronized by the master device. For practical reasons, the second method is usually used. The procedure is shown in Fig. 7 and the details are shown below.

In this system, the signal source for clock adjustment and for absolute time definition are called “clock source” and “time source”, respectively. Also, there are “time constant” for determining the follow-up speed to the signal source as parameters related to clock adjustment, and “absolute time synchronization interval” for indicating absolute time synchronization execution as parameters related to the absolute time synchronization.

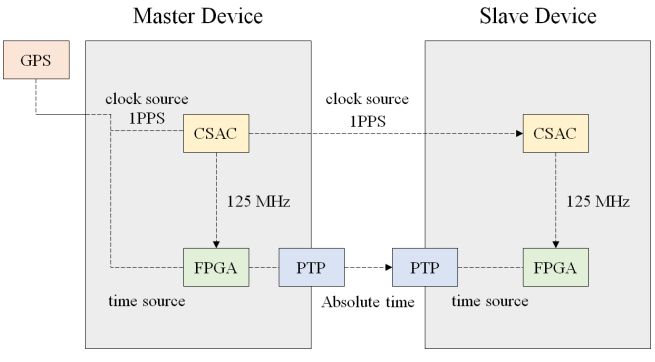


Figure 7. Time Synchronization Procedure among Practical Devices.

A. *Time Synchronization of Master Device*

When synchronizing the master device to the absolute time, GPS is used as clock source and time source. Clock adjustment and absolute time definition are performed according to preset time constant and absolute time synchronization interval. By receiving the GPS signal, the master device maintains high clock accuracy and absolute time accuracy. The time constant and absolute time synchronization interval are set to 1,000 seconds and 1,440 seconds, respectively. The absolute time is synchronized with the FPGA timer counter at the time interval set on the basis of 0:00:00 on the day on which this setting was made.

B. *Time Synchronization of Slave Device by Master Device*

Clock adjustment of the CSAC of the slave device is performed by utilizing the 1 PPS output of the CSAC of the master device. When the clock adjustment of the CSAC of the slave device is completed, the absolute time is set to the timer counter of the FPGA by the PTP. When absolute time setting by the PTP is completed, more precise time synchronization (less than a second) is performed by using the 1 PPS input of CSAC of the master device again.

VII. Communication of Sensor Data of Practical Device

The practical device sends the measurement data via Ethernet or wireless communication to the server on the cloud by "fluentd" after saving data in an SD card as shown in Fig. 8. Fluentd is an open source software called a data collector or data log collection tool, and it provides a function to collect log data and convert it to JavaScript Object Notation (JSON) and output it. JSON is one of lightweight data description languages and is designed to be used for passing data between various software and programming languages. "Input function" and "output function" are modularized, and by adding a plug-in module, fluentd can correspond to various data sources and output destinations. Data in the JSON structure in a format conforming to fluentd is transmitted in a binary MessagePack. The continuity and loss of measurement data can be confirmed by sampling period and time stamp. In addition, tag of the data is updated at the start of measurement or at the time of event detection at the date, hour, minute, and second as an identifier for one measurement, and the same tag is given after the end of measurement or the event.

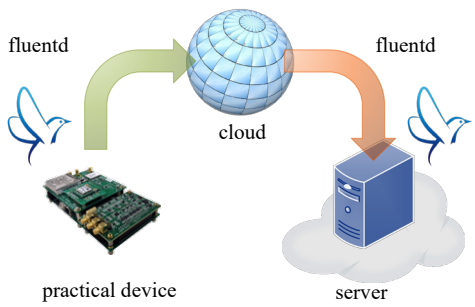


Figure 8. Data transmission from device to server by fluentd.

VIII. Application to Actural Building

The developed practical devices were installed in an actual building and seismic observation started in October 2017. The building is a three-story reinforced concrete building built in Tsukuba, Ibaraki, Japan (Fig. 9). In each floor, one practical device was installed. Fig. 10 shows the plan of 2nd floor of the building and locations of the practical devices installed. Fig. 11 shows how the devices have been installed. As shown in both figures, dedicated installation space was made next to the staircase from 1st to 3rd floors. At each installation location, the device was screwed to the acrylic plate and it was fixed to the floor with an anchor. The device for the rooftop is fixed on the ceiling of the 3rd floor because the rooftop floor is outside.

The sensor device is able to use the built-in MEMS accelerometer and any analog sensor connected to the external input terminal solely. The device is set to use the built-in MEMS accelerometer. Table IV shows the specifications of the MEMS accelerometer. Because of the excellent noise performance of the built-in MEMS, it is possible to measure building vibrations from small earthquakes as well as large ones. This new practical device can detect earthquake occurrence and save seismic event data. The least square calculation for values measured by the accelerometer in each direction is derived by the following equation.

$$l_{\mathrm{r}} = \sqrt{\frac{1}{N}\sum_{i=N-1}^{0}\{(x_i - r_{xi})^2 + (y_i - r_{yi})^2 + (z_i - r_{zi})^2\}}$$

where, r_x, r_y, r_z are correction values of the following zero point.

$$r_x = \frac{1}{N}\sum_{i=N-1}^{0} x_i,\ r_y = \frac{1}{N}\sum_{i=N-1}^{0} y_i,\ r_z = \frac{1}{N}\sum_{i=N-1}^{0} z_i$$

Figure 9. External appearance of building.

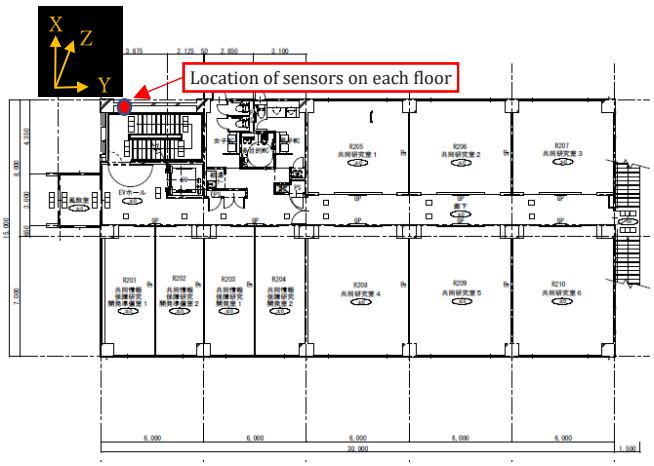


Figure 10. Plan view of 2nd floor of the building and locations of practical device installed.

The device was set to judge that an earthquake has occurred when the least square value l_r for 2 seconds exceeds 1 cm/sec^2 and to save the measurement data until the earthquake ends under the condition of a sampling frequency of 100 Hz and a number of correction values of zero point N = 200 in the above equation. The device also saves data 10 seconds before the point that the device judged that the earthquake has occurred and 10 seconds after the earthquake has ended. As shown in Section III, each practical device is equipped with Ethernet, Wi-Fi, and 3G as communication means and can use any of them. For this time, it was decided to use 3G for data transfer to the data server in order to store only the measurement data at the event of the earthquake.

Table V shows a list of seismic records observed from October 2017 when the system was installed to November 2017. When an earthquake with Japanese seismic intensity scale of 1 or more occurred, measurements were made reliably. Figs. 12 and 13 show measured acceleration of an earthquake in the building, which occurred on November 3, 2017 (No. 8 in Table V).

TABLE IV. SPECIFICATIONS OF MEMS ACCELERATION SENSOR

Model	LIS344ALH
Measurement direction	3
Maximum acceleration (± G)	2
Outside dimensions (mm)	4 × 4 × 1.5
Consumption current (mA)	0.68
Stand-by power consumption (µA)	1
Detection sensitivity	660 mV/G ± 5%
Noise characteristics	50 µG/√Hz
Operating temperature (°C)	-40 - +85

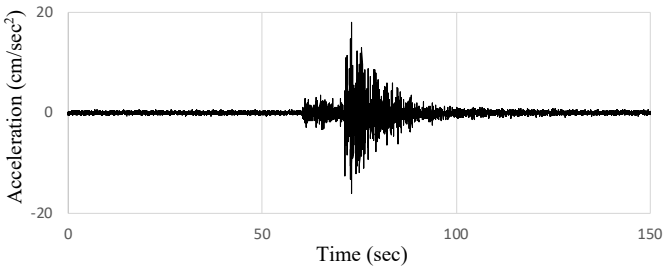
(a) Ceiling of 3rd floor (Floor of rooftop)

(b) Floor of 1st floor

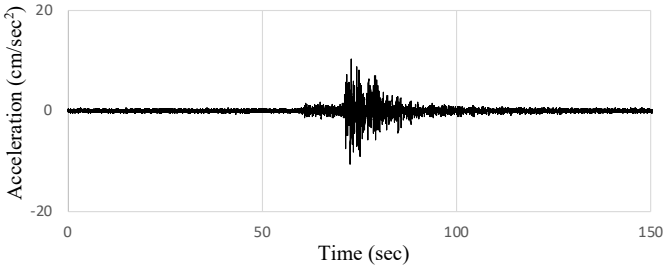
Figure 11. Photo of how practical devices are installed.

TABLE V. LIST OF EARTHQUAKE RECORD MEASURED BY PRACTICAL DEVICES

No	Date	Time	Name of Epicenter	Magnitude/ Depth(km)	Local/Max . Intensity
1	06/10/2017	16:59	Fukushimaken-oki	6.3/57	1/2
2	06/10/2017	23:56	Fukushimaken-oki	5.9/53	2/5 lower
3	07/10/2017	16:20	Ibarakiken-nanbu	3.4/43	1/1
4	12/10/2017	15:12	Fukushimaken-oki	5.2/26	1/2
5	15/10/2017	19:05	Ibarakiken-hokubu	3.0/7	1/1
6	18/10/2017	07:40	Ibarakiken-nanbu	3.7/45	1/2
7	02/11/2017	22:31	Ibarakiken-oki	4.3/74	1/3
8	**03/11/2017**	**21:38**	**Ibarakiken-hokubu**	**4.8/8**	**2/3**
9	05/11/2017	17:40	Ibarakiken-nanbu	2.9/43	1/1
10	15/11/2017	01:21	Ibarakiken-nanbu	3.8/20	1/2
11	26/11/2017	15:55	Ibarakiken-hokubu	3.9/4	1/2
12	30/11/2017	22:02	Ibarakiken-naubu	3.9/42	1/3

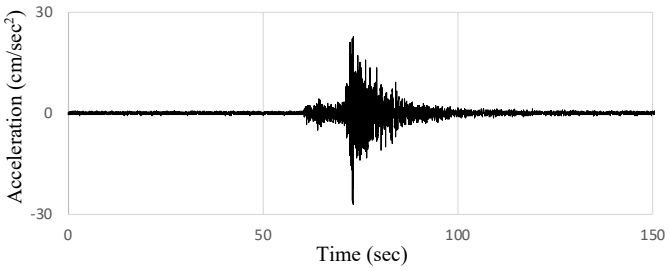


(a) Roof top floor (3rd ceiling)

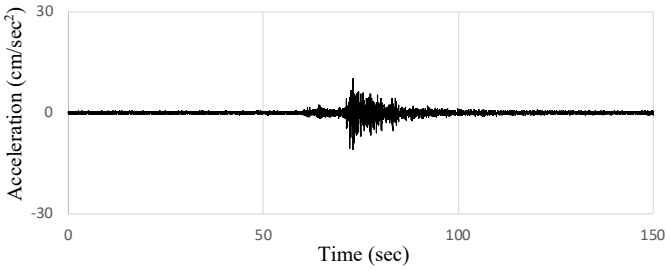


(b) 1st floor

Figure 12. Measured acceleration data (X direction).



(a) Roof top floor (3rd ceiling)



(b) 1st floor

Figure 13. Measured acceleration data (Y direction).

Figs. 12 and 13 show measured results of the 1st floor and the 3rd ceiling (floor of the rooftop) in a horizontal direction X and Y, respectively. For the 1st floor, an acceleration of 10.2 cm/sec^2 in X direction and 10.0 cm/sec^2 in Y direction at the maximum are observed. For the rooftop, the vibration of the earthquake is amplified and the maximum value of the acceleration is greater. The acceleration amplification factor of rooftop floor for the first floor is about 1.8 times in the X direction and about 2.2 in the Y direction.

Fig. 14 shows a Fourier spectrum of acceleration on the first floor. That is the acceleration of the seismic motion itself which is the input to the building. From this figure, dominant frequencies are observed at 2.4 Hz and 2.9 Hz in the X direction and 3.9 Hz in the Y direction. Figs. 15 and 16 show a transfer function (Fourier spectrum ratio) of the acceleration of each floor with respect to the first floor. Figs. 15 and 16 show the X direction and the Y direction components, respectively. By calculating the transfer function of each floor, it is possible to eliminate the influence of the frequency component of the seismic wave and observe only the dynamic characteristics of the building. It can be confirmed from the figure that the primary natural frequencies in the X direction and the Y direction are around 4.5 Hz and 5.5 Hz, respectively. Observation of the primary natural frequency is important for the structural health monitoring of buildings. If it moves to a lower frequency after the earthquake, it means that the stiffness of the building's structural member has decreased and the building has been damaged.

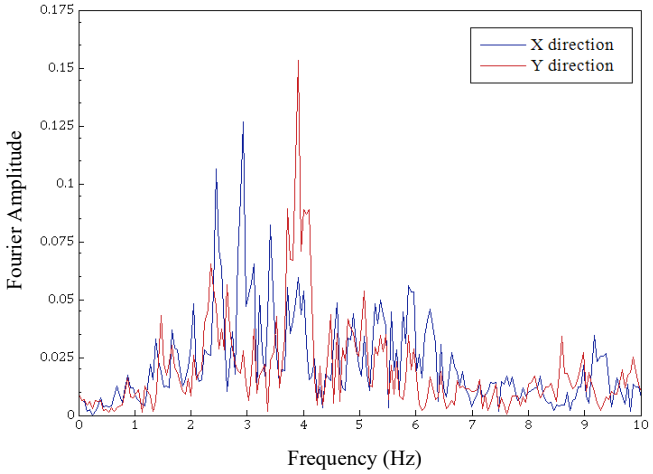


Figure 14. Fourier spectrum of acceleration on 1st floor.

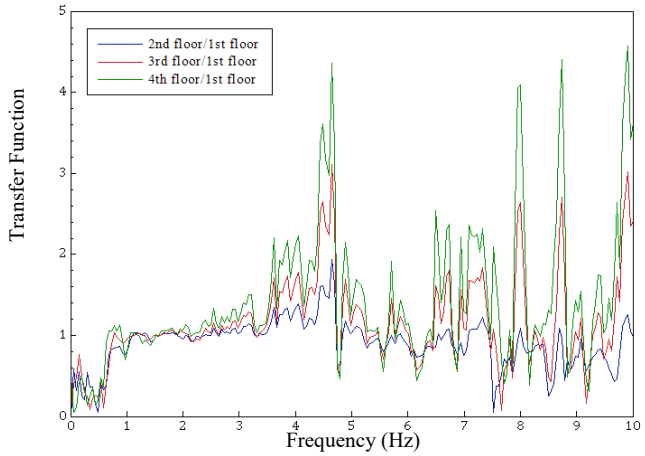


Figure 15. Acceleration of each floor relative to the first floor in X dir.

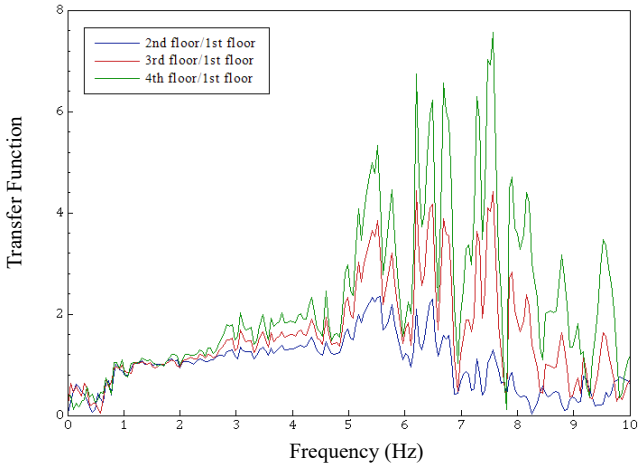


Figure 16. Acceleration of each floor relative to the first floor in Y dir.

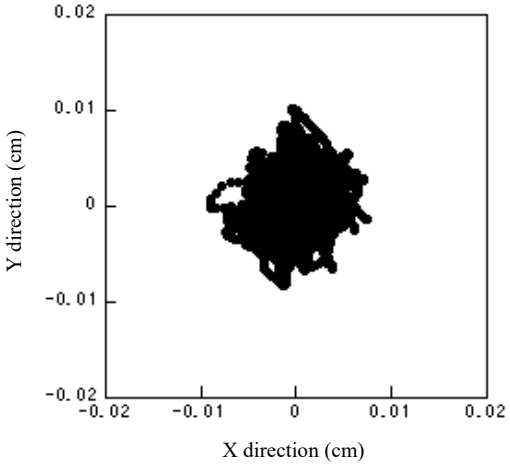


Figure 19. Inter-story deformation (1st story).

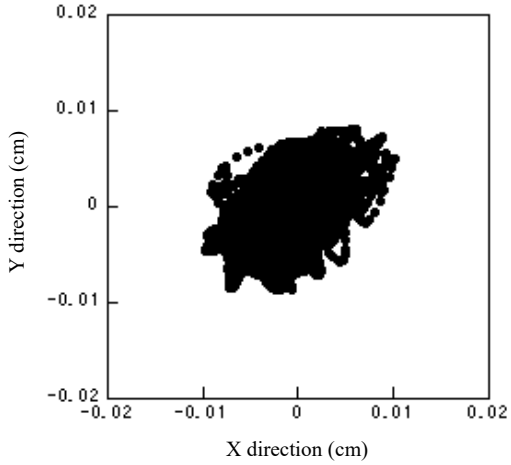


Figure 17. Inter-story displacement (3rd story).

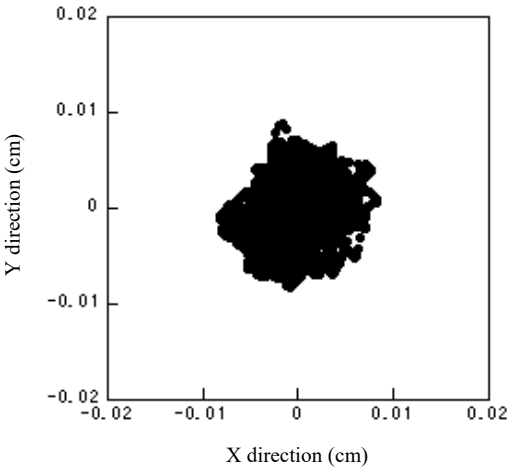


Figure 18. Inter-story displacement (2nd story).

Even if many sensors are installed in a building, analysis to evaluate the structural health of the building cannot be performed unless time synchronization between the sensors is secured [19]. Structural health of the building after the earthquake can be evaluated by the amount of deformation of each inter-story by the earthquake. Therefore, in order to evaluate the damage of the building in more detail, the inter-story deformation at each floor must be calculated. The displacement of each floor is calculated by double integration of the measured acceleration. Since time synchronization is ensured for the measurement data of this practical devices, the deformation of each inter-story can be calculated by subtracting the displacement of the upper floor and the lower floor. The value obtained by dividing the inter-story deformation of the building by the floor height is defined as the inter-story deformation angle. The maximum inter-story deformation angle during the earthquake and the damage of the building have the following relation. If the maximum inter-story deformation angle of the building during the earthquake is within 1/200, the building is assumed to be undamaged. The possibility of damage is high if it exceeds 1/100. Figs. 17, 18 and 19 show calculated inter-story deformation of the third, the second, and the first story of the building, respectively. Since the floor height of the targeted building is 380 cm to 390 cm, the inter-story deformation angle of 1/200 means an inter-story deformation of 1.9 cm to 1.95 cm. Since the inter-story deformation shown in Figs. 17, 18 and 19 is much smaller in both the X and Y directions, it can be evaluated that the building is undamaged against this earthquake.

Function and usability of the improved practical device, which was installed onto an actual building, were verified by the case study of seismic observation and structural health monitoring. In particular, such structural health monitoring of the building was made possible by the practical device developed in this research securing autonomous time synchronization.

IX. CONCLUSION

This paper describes research and development on the sensor devices that can keep highly accurate time information autonomously using the built-in Chip Scale Atomic Clock (CSAC) for the purpose of application to seismic observation and structural health monitoring of buildings and civil infrastructure. First, a process of development was explained from the prototype, which uses the mechanism that adds highly accurate time information to the measured data with the CSAC to the practical device. Then, challenges in practical use were identified by the long-term measurement implemented by installing the developed practical device onto a bridge. Based on the challenges, further improvements were made for stability and operability of the practical device, and its performance was confirmed. A new procedure for time synchronization between devices and the communication system of measurement data newly constructed using "fluentd" which is open source software for data collection were detailed. In addition, the improved practical device was installed onto an actual building, and its function and usability were verified by the case study of seismic observation and structural health monitoring. Finally, structural health monitoring of the building based on the evaluation by the inter-story deformation was made possible by the improved practical device developed in this research securing autonomous time synchronization.

As one of challenges, the operating method of the sensing system must be considered depending on the objective of measurement because the CSAC is aged over time although it is a highly accurate clock. Another challenge is the high cost of manufacture. The CSAC is expected to be mounted on all computers and smart phones in the near future, but only one American company manufactures and sells it at this time. It is expected that many companies will also participate in the business, and CSACs are actively used in various fields. Further verification will be performed by using the sensing system in actual buildings and civil infrastructures.

ACKNOWLEDGMENT

This research was partially supported by the New Energy and Industrial Technology Development Organization (NEDO) through the Project of Technology for Maintenance, Replacement and Management of Civil Infrastructure, Cross-ministerial Strategic Innovation Promotion Program (SIP). This research was also partially supported by JSPS KAKENHI Grant Number JP16H01717 and JP16K01283.

REFERENCES

[1] N. Kurata, "Improvement and Application of Sensor Device Capable of Autonomously Keeping Accurate Time Information for Buildings and Civil Infrastructures," The Ninth International Conference on Sensor Device Technologies and Applications (SENSORDEVICES 2018) IARIA, Sep. 2018, pp. 114-120, ISBN: 978-1-61208-660-6.

[2] N. Kurata, B. F. Spencer, and M. Ruiz-Sandoval, "Risk Monitoring of Buildings Using Wireless Sensor Network," Journal of Structural Control and Monitoring, vol. 12, Issue 3-4, pp. 315-327, July-Dec. 2005, doi: 10.1002/stc.73.

[3] N. Kurata, M. Suzuki, S. Saruwatari, and H. Morikawa, "Actual Application of Ubiquitous Structural Monitoring System using Wireless Sensor Networks," Proc. the 14th World Conference on Earthquake Engineering (14WCEE) IAEE, Oct. 2008, pp. 1-8, Paper ID:11-0037.

[4] N. Kurata, M. Suzuki, S. Saruwatari, and H. Morikawa, "Application of Ubiquitous Structural Monitoring System by Wireless Sensor Networks to Actual High-rise Building," Proc. the 5th World Conference on Structural Control and Monitoring (5WCSCM) IASCM, July 2010, pp. 1-9, Paper No. 013.

[5] S. Knappe, V. Shah, P. D. D. Schwindt, and J. Kitching, "A microfabricated atomic clock," Applied Physics Letters, vol. 85, Issue 9, pp. 1460-1462, Aug. 2004, doi:10.1063/1.1787942.

[6] Q. Li and D. Rus, "Global Clock Synchronization in Sensor Networks," IEEE Transactions on Computers, vol. 55, Issue 2, pp. 214-226, Jan. 2006, ISSN: 0018-9340.

[7] R. Lutwak et al., "The Chip-Scale Atomic Clock - Prototype Evaluation," Proc. the 39th Annual Precise Time and Time Interval (PTTI) Meeting, ION, Nov. 2007, pp. 269-290.

[8] N. Kurata, "Disaster Big Data Infrastructure using Sensing Technology with a Chip Scale Atomic Clock," World Engineering Conference and Convention (WECC2015) WFEO/UNESCO/SCJ/JFES, Dec. 2015, pp. 1-5.

[9] N. Kurata, "Basic Study of Autonomous Time Synchronization Sensing Technology Using Chip Scale Atomic Clock," Proc. the 16th International Conference on Computing in Civil and Building Engineering (ICCCBE2016) ISCCBE, July 2016, pp. 67-74.

[10] N. Kurata, "Development of Sensor Module for Seismic and Structural Monitoring with a Chip-Scale Atomic Clock," Proceedings of the 16th. World Conference on Earthquake Engineering (16WCEE) IAEE, Jan. 2017, pp. 1-8, Paper No.583.

[11] N. Kurata, "An Autonomous Time Synchronization Sensor Device Using a Chip Scale Atomic Clock for Earthquake Observation and Structural Health Monitoring," The Eighth International Conference on Sensor Device Technologies and Applications (SENSORDEVICES 2017) IARIA, Sep. 2017, pp. 31-36, ISSN: 2308-3514, ISBN: 978-1-61208-581-4.

[12] N. Kurata, "Development and Application of an Autonomous Time Synchronization Sensor Device Using a Chip Scale Atomic Clock," Sensors & Transducers Journal, Vol. 219, Issue 1, pp.17-25, January 2018, ISSN: 2306-8515, e-ISSN 1726-5479.

[13] D. L. Mills, "Internet time synchronization: the network time protocol," IEEE Transactions on Communications, vol. 39, Issue 10, pp. 1482-1493, Oct. 1991, doi:10.1109/26.103043.

[14] M. Maroti, B. Kusy, G. Simon, and A. Ledeczi, "The Flooding Time Synchronization Protocol," Proc. the 2nd International Conference on Embedded Networked Sensor Systems (SenSys '04), Nov. 2004, pp. 39-49, ISBN:1-58113-879-2.

[15] Q. Li and D. Rus, "Global Clock Synchronization in Sensor Networks," IEEE Transactions on Computers, vol. 55, Issue 2, pp. 214-226, Jan. 2006, ISSN: 0018-9340.

[16] J. Elson, L. Girod, and D. Estrin, "Fine-Grained Network Time Synchronization using Reference Broadcasts," Proc. 5th Symposium on Operating Systems Design and Implementation (OSDI'02), Dec. 2002, pp. 147-163.

[17] S. Ganeriwal, R. Kumar and M. B. Srivastava, "Timing-sync Protocol for Sensor Networks," Proc. the 1st international conference on Embedded networked sensor systems (SenSys '03), Nov. 2003, pp. 138-149.

[18] K. Romer, "Time Synchronization in Ad Hoc Networks," Proc. the 2nd ACM International Symp. on Mobile Ad Hoc Networking & Computing (MobiHoc'01), Oct. 2001, pp. 173-182.

[19] C. Boller, F. K. Chang, and Y. Fujino. eds. "Encyclopedia of Structural Health Monitoring," John Wiley & Sons, 2009.

Generation of Geodetic Lines and Duplication of Triangulated Convex Surfaces

Anna von Pestalozza
Chair of Short-Time Dynamics
University of the Federal Armed Forces
Hamburg, Germany
pestalozza@hsu-hh.de

Arash Ramezani
Chair of Short-Time Dynamics
University of the Federal Armed Forces
Hamburg, Germany
ramezani@hsu-hh.de

***Abstract*—In the following paper, research about geodetic lines as well as about surface duplication is presented. The calculation of geodetic lines plays an important role in many applications, such as the minimisation of material in manufacturing processes. Many manufacturing steps, such as cutting or attaching layers on curved surfaces, suffer from loss of material. In order to minimise wastage of material, geodetic lines can be employed to find a cutting pattern for the given material with minimal distortion. This paper presents an automatable algorithm that numerically calculates geodetic lines on any given surface. The result is evaluated with a practical example by comparing the numerical result and the analytical solution. The creation of multiple layers serves the purpose to reinforce a given structure to increase its stability, which is commonly done in manufacturing processes. This paper presents an algorithm, which calculates the coordinates of multiple attached layers with any given thickness of layers. Furthermore, the point of maximum curvature is determined.**

Keywords-Geodetic Line; Surface Analysis; Surface Orientation; Surface Curvature.

I. INTRODUCTION

This paper is based on the assumptions that the material of the given surface is of finite thickness and of low elasticity, which leads to the necessity of minimising loss of material as well as to the necessity to calculate size and coordinates of further layers. The project in which the presented research is embedded aims particularly at protection gear such as helmets and vests. It is therefore presumed that the surfaces which have to be investigated are mainly convex. Moreover, it is assumed that the given triangulation is so fine that on a convex surface the scalar products of the normal vectors of adjacent triangles are always positive. For the implementation of the presented algorithms the software tool Matlab was used for implementation as well as for evaluation.

A general overview over the research about geodetic lines is given in [1], which will be described more in detail in the following paper. The starting point of the research on the geodetic lines is the approach given in [2] for finding geodetic lines between two points. The main idea is to successively calculate distances from the starting point which is improved by the fast marching method. In [1], an algorithm for extracting the geodetic line as well as for further improving it is derived, which will be repeated in the following as well as deepen.

For cutting a curved surface either sectional planes or geodetic lines can be used. The graphical approximation of flattened material stripes of an originally curved surface having been cut by the procedures of applying sectional planes and calculating geodetic lines clearly show that the cutting with the geodetic lines provides straight edges when flattened whereas the sectional planes result in curved edges which leads to a higher amount of material loss. However, sectional planes are much easier to apply and less time consuming than the analytical calculation of geodetic lines which is not even possible in many cases. Thus, this paper aims to provide an algorithm which approximates analytical geodetic lines on any given surface [1].

Using geodetic lines, a cutting pattern can be derived which is not explicitly explained in this paper. However, it is not only one layer which is treated when producing objects so that several layers have to be taken into account. This paper presents an approach how to duplicate layers. For this purpose, finite thinness of the single layers is assumed which is in practice always true so that the overall area of a single layers slightly increases on a convex structure proportional to the number of layers (assuming that the layers are applied on the outer/convex face of the given surface). This shift of coordinates is addressed in this paper as well as the analysis of the orientation of the surface. As a result, an algorithm is presented which uniformly applies layers of any given thickness and size to the *outside* of any given convex surface which will be derived, e.g., which thickens a surface without changing its shape. The algorithm can be easily adapted to shift the coordinates to the inner face if needed.

The paper is divided in nine sections. After the introduction, the calculation of the shortest distance on a triangulated mesh is shown in Section II followed by Section III about the extraction of the geodetic line. In Section IV a straightening algorithm for improvement of the geodetic line is presented. In Section V exemplary results of the applied algorithm which is derived in Section II-IV are evaluated. Section VI explains the basic considerations and concepts of the creation of multiple layers and challenges which have to be faced when duplicating layers. The duplication of layers is then described in Section VII. In

Section VIII the point of maximum curvature. Its relevance for research is shortly paraphrased in Section IX which is a concluding paragraph and an outlook to future work on this project at the end of this paper.

II. Calculation of Shortest Distances on a Triangulated Mesh

In this paper, the procedure of the Fast Marching Method (FMM) is used [4] for calculating the shortest distances on a triangulated mesh. Basically, this method approximates the distances of all points surrounding the starting point successively by a wave front until it reaches the given ending point. For the following procedure it is assumed that starting and ending point of the geodetic line which is to be approximated are given.

A. Procedure

In the FMM, the vertices of all triangles in the mesh are divided into several groups which are sets of vertices.

1) Fixed vertex set (FVS): contains initially only the starting points; vertices which are points of the shortest distance are added in the procedure.

2) Close vertex set (CVS): contains initially no vertices; vertices which are close to the point that is investigated in the current iteration of the loop are added.

3) Unprocessed vertex set (UVS): contains all vertices of the mesh that are not contained in FVS.

Two situations can be distinguished: Only one starting point is given and more than one starting point is given. If there is only one starting point, the distances T_i of its direct neighbours have to be calculated and the neighbours are added to the CVS. If there is more than one starting point, the points a_0, a_1, and a_2 which are part of a triangle of the mesh containing exactly two points in FVS have to be determined. After computing their distances T_0, T_1, and T_2 to the starting value, the points a_0, a_1, and a_2 are added to CVS. After these initial steps, the following loop starts:

- The point a_i, i = 0, 1, 2 with the shortest distance T_i to the starting value is moved to FVS and is now the point of origin for further investigations. This point is called trial.
- The distances T_i of all points in UVS∪CVS which are adjacent to triangles containing trial and a point in FVS are computed and moved to CVS.

In each iteration, one point is added to FVS and its neighbours are added to CVS. The algorithm terminates when FVS contains every vertex which is part of a line resulting in the shortest distance from starting to ending point.

B. Calculation of Distance T

For calculating the distance T the method presented in [3] is used. It requires that one point, P_1, of known distance T_1 is the origin and that another point, P_2, of known distance T_2 is on the x-axis.

1) Procedure: The distance T_3 of the third point P_3 is calculated in terms of T_1, T_2 and the connecting vectors v_i with $v_i = P_i - P_1$, in particular $(v_3)_x$ and $(v_3)_y$, i.e., the projections of v_3 onto the new basis vectors, which are calculated as follows: To change the default, adjust the template as follows:

1. One point is set as the origin (P_1):

$$v_i = P_i - P_1$$

2. The coordinate system is transformed, where:

$$e_x = \frac{(p_2 - p_1)}{|p_2 - p_1|} = \frac{(v_2)}{|v_2|}$$

$$e_y = \frac{(v_3 - e_x \cdot (e_x \cdot v_3))}{|(v_3 - e_x \cdot (e_x \cdot v_{3)})|} = \frac{(v_3 \cdot |v_2|^2 - v_2 \cdot (v_2 \cdot v_3))}{|(v_3 \cdot |v_2|^2 - v_2 \cdot (v_2 \cdot v_3)|}$$

3. The distance of v_3 to the new coordinate system is computed where O_x is the x-coordinate at which the origin of the new coordinate is located and O_y is the relative y-coordinate:

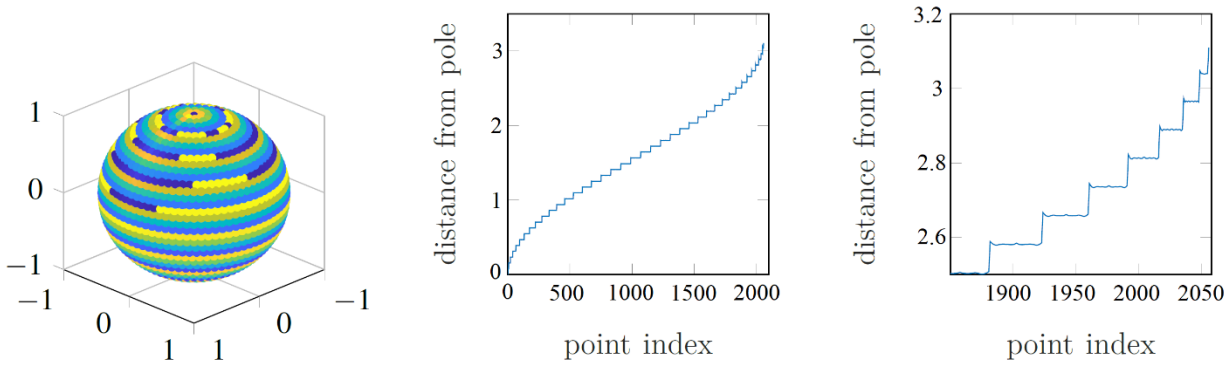
$$O_x = \frac{1}{2}\frac{(v_2)_x^2 + T_1^2 - T_2^2}{(v_2)_x}$$

$$O_y = \pm\sqrt{T_1^2 - \frac{((v_2)_x^2 + T_1^2 - T_2^2)^2}{4(v_2)_x^2}} = \pm\sqrt{T_1^2 - O_x^2}$$

$$T_3 = O_x \cdot e_x + O_y \cdot e_y - v_3$$

A challenge with this method is that there are always two possible virtual origins due to $\pm O_y$. In [4] it is stated that this is solved by calculating both distances and taking the larger value. However, there are situations where the smaller value is the correct one. This happens, presumably, mostly or only when P_3 is not in front of the wavefront but beside. Such a situation occurs when the distance of a point in the CVS is recalculated. To mitigate this issue in a simple way, the recalculated value for the distance T is only stored if it is smaller than the existing one.

2) Accuracy: The algorithm was tested on a sphere with equally spaced points as shown in Figure 1. The starting point, i.e., the point with distance $T = 0$ is chosen to be the north pole. The points are numerated such that one whole circle at constant θ is taken. Thus, plotting the distance over the index results in plateaus of constant distance as shown in Figure 1b and 1c.



(a) The colourmap was chosen to be repeating 10 times so that intra-ring fluctuations are visible.

(b) Staircase-distances emerging from the ring-after-ring-data-structure

(c) Zoom of the upper „stair“ shows oscillations (visible in intra-ring colour changes in (a))

Figure 1. Distances on a homogeneously sampled sphere. The index starts at the south pole and increases ring by ring until the north pole is reached. Distances are calculated from the south pole via the fast marching method.

III. EXTRACTING THE GEODETIC LINE

In the second section, the shortest distance from starting to ending point on the triangulated mesh is determined. In order to approximate the geodetic line, a line of shortest distance can be backtracked along the points in FVS. The real geodetic line, however, does not necessarily consist only of vertices but of points on edges of the triangles as well. In the following the first approximation of the geodetic line is denoted by Γ_0.

A. Method of Minimum Distance

The Method of Minimum Distance approximates Γ_0 with regard to the calculated distances *T*. It iterates the following procedure and can be modified through two different options:

1. The neighbor *N* of the previous point is determined which fulfills one of the following requirements:

1. Option 1: *N* has the lowest distance T_N of all provided neighbours.

2. Option 2: *N* is the point of neighbours for which the value of the distance T_N added to the distance from the previous point *p* is minimal

2. The resulting neighbor N is appended to Γ_0.

This method extracts the geodetic line very quickly but does not provide a good approximation, neither with Option 1 nor Option 2, especially when the grid is very uniform. Also, the points of the geodetic line are still only located on vertices. Therefore, the *gradient method* was implemented.

B. The Gradient Method

The gradient method provides an approach to extract the geodetic line dissociated from the vertices. To determine the direction in which the geodetic line propagates the gradient of the distance *T*, approximated with the three distances for each point in each triangle, is used.

1) Approximation of the gradient in a triangle:

The gradient in a triangle with vertices *i*, *j* and *k* is given by

$$(\vec{\nabla}T)_{(i,j,k)} = -\frac{\vec{n}}{|n|^2} \times \left(T_i \vec{e}_{jk} + T_j \vec{e}_{ki} + T_k \vec{e}_{ij}\right),$$

where

$$\vec{e}_{a,b} = \vec{x}_b - \vec{x}_a$$

are the vectors connecting the vertices *a* and *b* and $\vec{n}$ is the

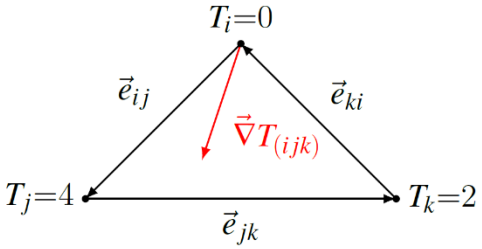


Figure 2. Three points (*i*, *j*,*k*) of a triangle with distance values *T* and the approximated gradient $(\vec{\nabla}T)_{(i,j,k)}$

surface normal of the triangle:

$$\vec{n} = \vec{e}_{ki} \times \vec{e}_{jk}$$

Note that the connecting vectors $\vec{e}_{ij}$, $\vec{e}_{jk}$ and $\vec{e}_{ki}$ are circular, i.e., that

$$\vec{e}_{ij} + \vec{e}_{jk} + \vec{e}_{ki} = 0$$

In Figure 2, a sketch of a triangle with its gradient is shown for an example set of distance values T_i, T_j, T_k.

2) Extracting the Geodetic Line with the Gradient Method:

The basic concept of the gradient method is to generate a line g for each triangle from the previous point p of the geodetic line and the gradient of T

$$g : \vec{x}(\lambda) = \vec{p} + \lambda \vec{\nabla} T_{(ijk)}$$

and to find its point of intersection with the edges of adjacent triangles. For the choice of edges to intersect g with, one has to consider whether the previous point p is on a vertex or an edge. If p is on a vertex, the following procedure is applied:

1. The negative gradients of the adjacent triangles are computed.
2. A triangle is determined whose negative gradient points into the triangle itself.
3. The line g is intersected with the edge of that triangle on the opposite side.
4. The point of intersection is added to Γ_0.

If no triangle is found whose negative gradient points into the triangle itself, the neighbour N with the smallest distance T to the previous point p is added to Γ_0.

If p lies on an edge, a different procedure is used:

1. The triangle which is adjacent to p and was not used for the prior calculation of p itself has to be identified.

2. The line g is intersected with the two remaining edges, if the negative gradient points into the triangle.

If the negative gradient does not point into the triangle, the previous p is moved to the vertex of the same edge that has the smaller distance T.

Special case: It might happen that p lies on a boundary edge. This case can be resolved by moving p to the vertex of the same triangle with a smaller distance T. If p lies on a boundary vertex, the above-mentioned procedure can be applied without further arrangements. As already mentioned, his is a special case. Therefore, this will not be considered in the further course.

3) Performance of the Gradient Method

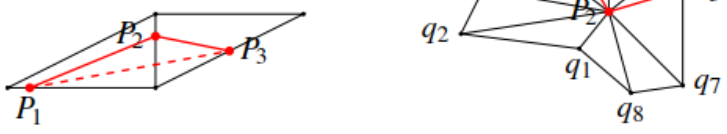
The algorithm approximates the real geodetic line in many test cases very precisely in accurate time. In case that real geodetic line runs near or along a line of edges without passing through several triangles or without changing the lane over the course of many points, the calculated geodetic line tends to stick to one lane and very late moves over to the other. This cannot be taken care of by the improvement algorithm which is described in the next section unless it is run for a lot more iterations than usual which is expensive. However, this special case is not problematic unless one wants to find the real geodetic line with even higher accuracy than already provided. For this, one could calculate the geodetic line and refine the triangulation around it to redo the whole calculation with the new triangulation until it converges.

IV. Improving the Approximation of the Geodetic Line

In the previous section we have generated an initial approximation Γ_0 for the geodetic line between two points on a triangulated mesh in three-dimensional space. As this is just a first approximation, an algorithm for improving Γ_0 is required. The improvement can be achieved by moving the points on vertices of the geodetic line along the edges of the mesh to shorten the length of Γ_0.

A. Criterion for Improvement of the Geodetic Line

According to [2] the shortest path is given by the straightest path for triangulated surfaces. 'Straight' is defined as follows: After taking all triangles that the approximation Γ_{i-1} passes through and unfolding them into a plane, the path Γ_i is the shortest when it is a straight line in the planar view. Therefore, the algorithm for improvement aims at straightening the path in the unfolded planar view.



(a) $P_{i,j}$ lies on an edge. (b) $P_{i,j}$ lies on a vertex.

Figure 3. The two cases of a point of the geodetic Γ_i needing correction. For readability $P_{i,-1,j}$, $P_{i,,j}$ and $P_{i,+1,j}$ have been replaced by P_1, P_2 and P_3, respectively. The dashed line denotes the corrected path.

B. The straightening algorithm

For this section the *i*-th version of the path is denoted as Γ_i and $P_{i,j}$ the *j*-th point of the *i*-th path. For the following let $P_{i,j}$ be the point to be corrected using the information about $P_{i,j+1}$ and $P_{i,j-1}$. The idea is to locally straighten the path by moving the central point of the three, i.e., $P_{i,j}$. To ensure that the geodetic line converges and actually becomes shorter with each iteration, the updated $P_{i+1,j}$ for the updated path Γ_{i+1} is calculated using the points which have already been updated during this iteration, i.e., $P_{i,+1,j-1}$ instead of $P_{i,j-1}$. For readability, we omit the "+1" in $P_{i,+1,j-1}$, but take care of it by only keeping one Γ stored and updating it with each step during each iteration.

There are always two cases to be considered: $P_{i,j}$ lies on an edge or on a vertex as can be seen in Figure 3.

1) $P_{i,j}$ lies on an edge:

If $P_{i,j}$ lies on an edge, the following steps are needed to improve the line:

1. The two triangles adjacent to $P_{i,j}$ are unfolded.
2. The point of intersection of the connecting line between $P_{i,j+1}$ and $P_{i,j-1}$ and the edge that $P_{i,j}$ lies on are calculated. If the point of intersection does not lie between the two vertices of the edge, the closer vertex is chosen to be the corrected point instead in this case.

As an example, let us assume the three points P_1, P_2 and P_3, of which P_2 lies on an edge and is the point which has to be corrected. Since P_2 is on an edge, there are only two triangles adjacent to P_2. For this example, let us call the points of the first triangle *A*, *B* and *C*, and of the second triangle *B*, *C* and *D*. P_2 lies subsequently on the edge connecting *B* and *C*, P_1 lies in the triangle limited by *ABC* and P_3 lies in the triangle *BCD*. Applying the procedure explained above, first, the points *D* and P_3 have to be rotated around $\overline{CD}$ to be in the same plane as *ABC*. The rotated points will be denoted *D'* and P_3'. When defining *C* as the origin, every investigated point is moved by $-\vec{c}$, which is the position vector of *C* in the original coordinate system. The position vectors of the points P_3' and *D'* are then given by:

$$\vec{p}_3{}' \times= \vec{v}(\vec{v}\cdot\vec{p}_3)\ + \cos(\phi)\,(\vec{v}\times\vec{p}_3)\times\vec{v} + \sin(\phi)(\vec{v}\times\vec{p}_3)$$
$$\vec{d}' \times= \vec{v}(\vec{v}\cdot\vec{d})\ + \cos(\phi)\,(\vec{v}\times\vec{d})\times\vec{v} + \sin(\phi)(\vec{v}\times\vec{d})$$

Where the edge connecting the vertices *B* and *C* is given by

$$\vec{v} = \vec{b} - \vec{c}$$

And the angle by which the points have to be rotated by:

$$\phi = \pm\arccos\frac{(\vec{n}_1\times\vec{n}_2)}{|\vec{n}_1|\times|\vec{n}_2|}$$

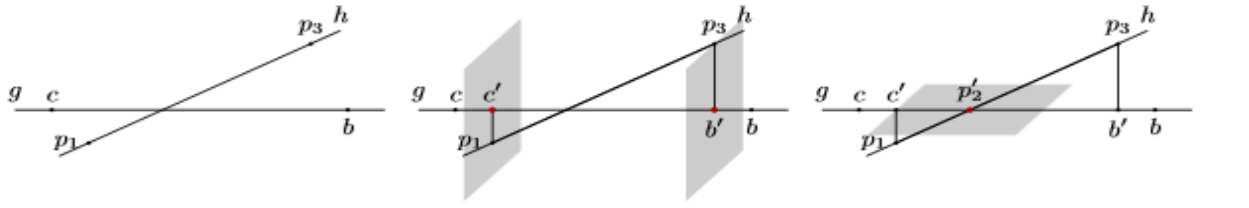
As the second step, the point of intersection of the edge $\overline{BC}$, in the following denoted as *h*, and $\overline{P_1,P_3{}'}$, in the following denoted as *g*, has to be calculated. The points *C'* and *B'* are found, which are defined as the points on *h* with the shortest distance to P_1 and P_3' which can be seen in Figure 4.

A plane at *C'* with normal vector $\vec{n}_2 = \vec{p}_1 - \vec{c}'$ can then be inserted as can be seen in Figure 4c. This plane can then be intersected with the line to find $\vec{p}_2$:

$$g: \vec{x} = \vec{p}_3 + \lambda(\vec{p}_1 - \vec{p}_3) = \vec{p}_2$$

With

$$\lambda = \frac{(\vec{c}' - \vec{p}_3)\cdot\vec{n}_2}{\vec{n}_2\cdot\vec{v}_2}$$

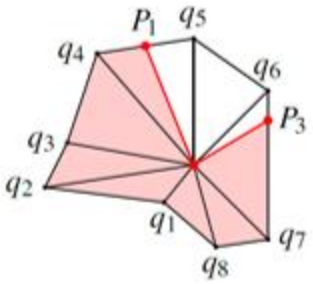


(a) Initial situation (two lines defined by four poitns)

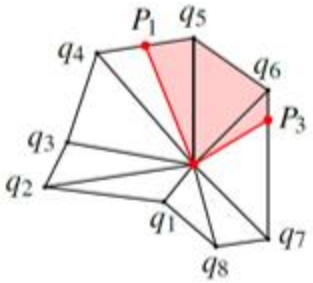
(b) Two planes (perpendicular to *g*) added for finding *c'* and *b'*

(c) Plane (perpendicular to $\vec{p}_1 - \vec{c}'$) added for finding $\vec{p}_2$

Figure 4. Intersecting two lines



(a) Angle θ_r



(b) Angle θ_l

Figure 5. Notation of angles.

2) $P_{i,j}$ lies on a vertex

If the point that is to be corrected coincides with a vertex, the procedure becomes more complicated. Let S_k be the set of triangles that have $P_{i,j}$ as the central vertex, then several cases can be distinguished. Firstly, there are two simple cases which can be easily taken care of numerically:

a) If all three points ($P_{i,-1,j}$, $P_{i,j}$ and $P_{i,+1,j}$) are part of the same triangle, $P_{i,j}$ is removed from Γ.

b) If $P_{i,\,j+1}$ *or* $P_{i,j-1}$ lies on an edge that is not part of the boundary of S_k, it is removed from Γ.

For all other cases ($P_{i,\,j+1}$ and $P_{i,j-1}$ belong to two different triangles) the vertices around $P_{i,j}$ are sorted and the left and right hand angles, θ_l and θ_r, are calculated in order to characterize the vertex as can be seen in Figure 5. These angles are given by the sum of the central angles of the triangles which are obtained by splitting the star-like structure of S_k along the path $P_{i,-1,j} \rightarrow P_{i,,j} \rightarrow P_{i,+1,j}$.

Three main cases can be distinguished:

1. $\theta = 2\pi$: euclidean
2. $\theta = \theta_l + \theta_r > 2\pi$: hyperbolic
3. $\theta < 2\pi$: spherical

These three cases are taken care of differently where θ is defined as left or right hand angle.

a) Euclidean: S_k can be unfolded isometrically. After unfolding, $P_{i,\,j+1}$ or $P_{i,j-1}$ are joined in the unfolded S_k and the intersections with the edges added to Γ.

b) Hyperbolic:

a. If θ_l and θ_r are greater than π: no correction is needed.

b. If θ_l and θ_r are smaller than π, that side of S_k is unfolded and $P_{i,\,j+1}$ as well as $P_{i,j-1}$ are joined in the same manner as in the Euclidean case.

c) Spherical: The part of S_k with smaller $\theta_{l/r}$ is unfolded and $P_{i,\,j+1}$ or $P_{i,j-1}$ are joined as in the Euclidean case.

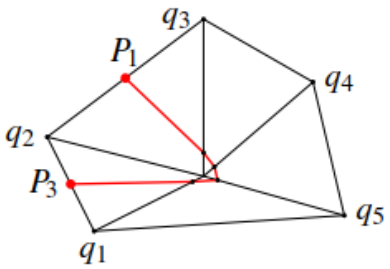
In all three cases the part of S_k with smaller $\theta_{l/r}$ has to be unfolded and the points of intersection have to be calculated.

In test runs, it was observed that points which are very close to vertices keep approaching the vertex which they are close to without coinciding and adopting its value. Therefore, every 10 iterations the path is scanned for points on Γ for which this might be the case. These points are moved to the vertex instead. All the following points that approach the same vertex are deleted from the path. This is necessary because otherwise curves in the path will never pass over a vertex. The effect of the scanning of the path and the movement of points to vertices is shown in Figure 6.

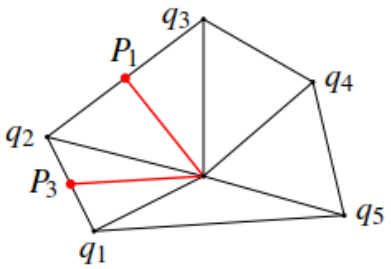
V. EXEMPLARY RESULTS

To test the capability of the gradient method and the straightening algorithm a geometry was chosen for which exact geodetic lines can be analytically computed for reference.

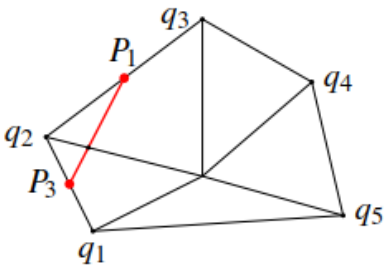
A plane with a half cylinder barrier is generated and the geodetic line between two points on either side of the half cylinder is calculated, first analytically, then using the



(a) The path is stuck at a vertex.



(b) After ten iterations all points close to a vertex are removed and replaced by one point which is directly located on the vertex.



(c) The next iteration corrects this point and the detour is removed.

Figure 6. Correction of points surrounding a vertex.

presented algorithm as shown in Figure 7.

The axis of the cylinder is set along the *y*-direction. Let us assume that the following parameters are given: the range of *x*, the step size *dx*, the range of *y*, the step size *dy*, the step size $d\phi$ for the angles on the cylinder, radius *r* of the cylinder and the displacement in *x*- direction of the cylinder axis *c* in the space. The cylinder is set in space such that there is translational invariance along the *y*-direction. Thus, the same *y*-vector can be used for each pair of *x* and *z*. The *x*-axis is separated into three parts. The first is an equally spaced vector that is limited by:

$$x = c - r - dx$$

The third is an equally spaced vector starting at

$$x = c + r + dx$$

The third part is the section in the middle of the mentioned ones and the one containing the half cylinder. This central part is given by:

$$x = c + r \cdot (-\cos(\phi))$$

For these three parts the *z*-coordinate is only unequal *zero* for the cylindrical part as can be seen in Figure 7. For the triangulation of the surface the Matlab function *delaunay* is used.

The surface can then be expressed in terms of a function *f* which is defined to be:

$$f: \mathbb{R}^2 \to \mathbb{R}$$
$$\text{i. e.}, f: (x, y) \to z(x, y)$$

This general generation of the test surface can be seen in Figure 7.

The analytical solution of the geodetic line running over the cylinder can be calculated as follows:
The starting and ending point, p_{start} and $p_{end,}$ of the desired geodetic line are set to be on different sides of the cylinder. The first step is then to find the point at which the geodetic line starts to climb the cylinder which is denoted as $\vec{\alpha}$ and the point at which the geodetic merges again to the plane which is denoted as $\vec{\beta}$ as can be seen in Figure 8a. To find these two points the half cylinder can be flattened which results in an isometric stretch of the grid in *x*-direction. The part which previously was on the right side of the half cylinder is then displaced by $r\pi - 2r$ as shown in Figure 8b, where shifted points are denoted with a prime (‘). Drawing a straight line between p_{start} and p_{end}’gives then the possibility to determine $\vec{\alpha}$ and $\vec{\beta}$ by the slope *m* of the drawn line. For this purpose, the slope is firstly calculated:

$$m = \frac{(\vec{p}_{end,y}{}' - \vec{p}_{start,y})}{(\vec{p}_{end,x}{}' - \vec{p}_{start,x})} = \frac{(\vec{p}_{end,y} - \vec{p}_{start,y})}{(\vec{p}_{end,x} + r(\pi - 2) - \vec{p}_{start,x})}$$

Let us call the *x*-component of the distance between p_{start} and the left bottom of the half cylinder $\Delta\alpha$ and the *x*-component of the distance between p_{end} and the right bottom of the half cylinder $\Delta\beta$, which are then given by:

$$\Delta\alpha = c - r - p_{start,x}$$
$$\Delta\beta = -c - r + p_{end,x}$$

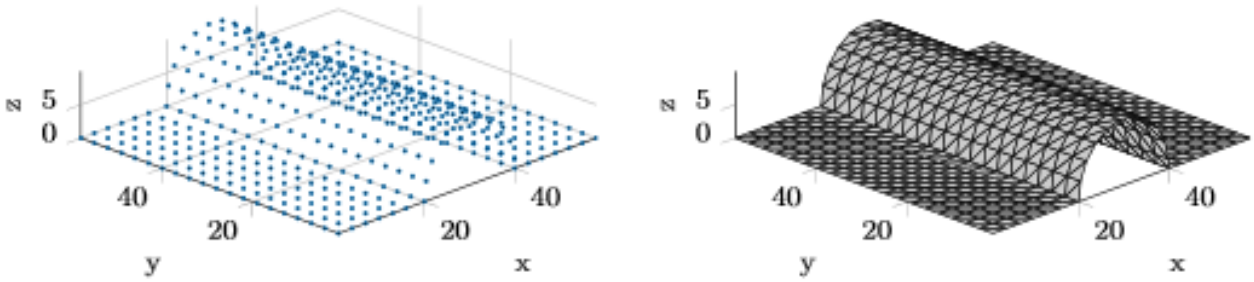
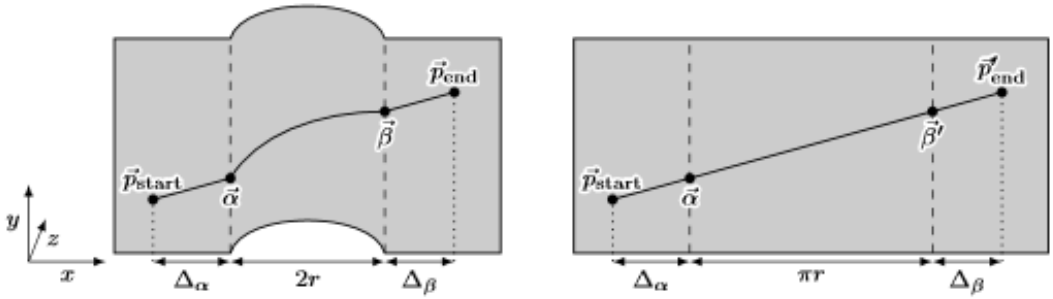


Figure 7. Scatterplot and delaunay triangulation of the test surface (Matlab).



(a) Sketch of a geodesic that goes over a cylindrical barrier. The points which it enters and leaves the barrier are highlighted ($\vec{\alpha}$ and $\vec{\beta}$)

(b) Stretching out the barrier to the right shifts $\vec{\beta}$ and p_{end} by $r\pi - 2r$

Figure 8. Flattening the half cylinder (Matlab Figure).

The coordinates of $\vec{\alpha}$ and $\vec{\beta}$ are subsequently given by:

$$\vec{\alpha} = \vec{p}_{start} + \begin{pmatrix} \Delta\alpha \\ m \cdot \Delta\alpha \\ 0 \end{pmatrix}$$

$$\vec{\beta} = \vec{p}_{end} - \begin{pmatrix} \Delta\beta \\ m \cdot \Delta\beta \\ 0 \end{pmatrix}$$

The distance in *x*,*y* and *z*-direction of the sections on the left-hand and right-hand side on the half cylinder can be easily calculated because the geodetic line is given by a straight line in a plane. The geodetic line in the central section can be analytically calculated as well:

$$y(\phi) = \alpha_y + \phi \cdot \frac{\beta_y - \alpha_y}{\pi}$$

In Figure 9, an exemplary result of the geodetic line analytically calculated is shown as a black line. The numerical result is indicated by red dots. Several aspects can be seen: Beginning at the ending point (on the right-hand side of the half cylinder) the distances of the other points are calculated by the FMM. These increase up to the starting point (on the left-hand side of the half cylinder). For clarity, the colour palette was chosen such that it is repeated five times.

In the left image the calculation is stopped after the extraction using the gradient method. For the right image the extracted geodetic line was improved by using the technique described in Subsection IV-B.

As can be seen, the straightening algorithm removed a few deviations visible close to the upper right end of the geodetic line. The straightening algorithm ran 50 times but most of the improvement was already achieved after five iterations.

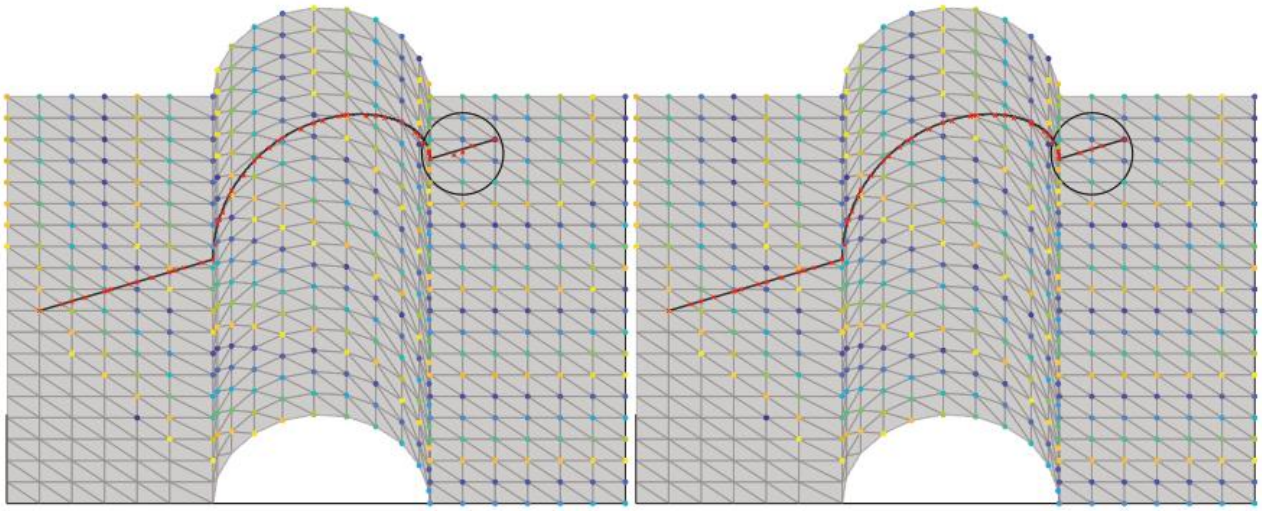


Figure 9. Result of fast marching method (both), geodetic extraction (left) and improvement algorithm (right)

VI. CREATING MULTIPLE LAYERS

When developing material structures, a given layer, which determines the structure of the desired object, has to be duplicated to ensure stability of the resulting object. For this, multiple considerations have to be taken into account which will be discussed in the following sections:

A) The layers, which extend the first one by being attached on it, do not have exactly the same size as the given one due to the thickness of the individual layers, which cannot be assumed to be infinitesimally small in practice. Thus, there are two general movements, which have to be considered. The first movement is the one, which the single investigated vertex is subjected to, i.e., the relative change of coordinates. The second one is the direction in which the surface moves as a whole. Here, the layers should shift to the *outside* of the surface, which has to be defined first.

B) In order to determine important specifications for the cutting pattern such as the thickness of the material layers and the width of the applied layers/material stripes, the point on the surface has to be found, which is subjected to the highest stress and thus experiences the highest strain in the whole layer. This point is, when regarding a surface without external loads being applied on it, the point of maximum curvature. At this point, the internal stress of the material is at its maximum. Material properties as well as manufacturing processes have to be adjusted in a way that the material at this point is still able to withstand the expected strains it might experience in its life cycle.

VII. DUPLICATING A TRIANGULATED LAYER

This section deals with the direction of the vertex shift, the direction of the surface shift and the resulting coordinates of additionally applied layers.

A. Direction of vertex shift

The direction $\vec{d}_i$, in which a single vertex moves, can be determined by the normal vectors $\vec{n}_i$ of the surrounding triangles T_i. Given are the edges of the triangles $\vec{e}_i$ and the corresponding vertices v_i. The normal vector $\vec{n}_j$ of a triangle is obtained by the cross product of two edges $\vec{e}_i$ depicted in vector form of the triangle T_j:

$$\vec{n}_j = \vec{e}_1 \times \vec{e}_2$$

If a point p_j is adjacent to i triangles T_i with normal vectors $\vec{n}_i$, then the direction $\vec{d}_i$ of the shift in p_j is given by the average of the normal vectors of the adjacent triangles as shown in 2-D in Figure 10.

$$\vec{d}_i = \frac{\sum(\vec{n}_i)}{i}$$

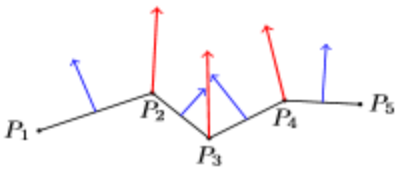


Figure 10. Calculation of normal vectors in points (red) from normal vectors in triangles (blue).

B. Direction of surface shift

1) The 'outside' of the surface

The outer face of protection gear, which is particularly aimed at in this project, is predominantly defined by convex structures such as shield, helmets and vests. The outside is subsequently generally defined as the space above the convex curvature of the surface. The problem, which arises, is that the orientation of the surface, or more specifically, of the order and direction of edges is not given by the *delaunay* triangulation in Matlab. Therefore, the calculation of the normal vectors is not necessarily right-handed which results in normal vectors that can point inside the surface when calculated with a left-handed system and outside the surface when calculated with a right-handed system. Hence, the normal vectors have to be oriented in a way that they all point in one direction which can then be easily adjusted to the outside of the surface.

Owing to the assumption that there are no large kinks in the triangulated surface, one can say that the projection of the normal vector of any triangle on the normal vectors of each of the adjacent triangles, i.e., the scalar product c_i, has to be positive when the normal vectors of the triangles point in the same direction. For this concept to work, the triangulation has to be fine enough, so that on a *convex* surface normal vectors of adjacent triangles, which point on the same side of the surface, have a positive scalar product without exception.

$$c_i = \vec{n}_i \cdot \vec{n}_{i+1} >! 0$$

2) Direction of the surface

To shift all vertices in the same direction, all normal vectors have to point in the same direction of the surface. At this point of the procedure, it does not matter, whether they point inside or outside the surface. The specified assumptions and calculations result in the following procedure to set the direction:

1) One arbitrary reference triangle T_r is determined, whose direction of the relative normal vector gives the reference normal vector for all other triangles (points to inside or outside)

2) It is iterated through all adjacent triangles of T_r and the projections of the normal vectors are calculated, i.e., the scalar product, of the normal vectors of the adjacent triangles on the reference normal vector.
3) If the scalar product is smaller than zero, the relative normal vector of the adjacent triangle has to be flipped, i.e., multiplied with "–1".
4) It is iterated for all triangles (without step a)) such that there is always one triangle in the neighborhood, which has already been checked on its direction. This triangle then sets the new reference normal vector.

At the end of this procedure, all normal vectors point on the same side of the surface, either the inside or the outside of the surface.

For the following principle, it is additionally assumed that the coordinate system is set such that the convex part of the structure points in positive *z*-direction.

To check, whether the normal vectors point inside or outside the structure, it is iterated through all normal vectors and searched for the one with the largest z-component, which is denoted as $\vec{z}_{max}$. The vector $\vec{z}_{max}$ should be the one of the triangle on top of the convex curvature of the surface, e.g., on top of the helmet. This works even with a rotated surface as long as the convex curvature points upwards. Owing to this, the scalar product of the unit vector in *z*-direction $\vec{e}_z$ and $\vec{z}_{max}$ can be used to determine the orientation on now all normal vectors because they all point on the same side of the helmet.

1) If the scalar product is positive, the normal vectors of all triangles point outside the surface. No arrangement has to be made in this case.

$$c_i > 0 \rightarrow \vec{n}_{i,final} = \vec{n}_i$$

2) If the scalar product is negative, the direction of all normal vectors has to be reversed, i.e., multiplied with "–1" so that they point outwards.

$$c_i < 0 \rightarrow \vec{n}_{i,final} = \vec{n}_i \cdot (-1)$$

The result at this point of the algorithm for multiple layers are the directions in which each point has to be shifted. Due to numerical round-off errors, the normal vectors might not all be of length 1, which is addressed in the next section.

3) Shifting the vertices

The vertices v_i of the triangulation can now be shifted *n*-times by the distance *s* in direction $\vec{d}_i$, where *n* is the number of required additional layers and *s* is the thickness of the applied layer. The coordinates for the new layers are denoted as X_{n+1} which is given in matrix form. At this stage, distance *s* is manually given. As mentioned above, the normal vectors may not be of length "1", which is why the distance *s* is divided by the length of the normal vector l in the following equation. X_n and the summand are of the same dimension as X_n stores all coordinates of *m* points in *m* rows and three columns for the corresponding *x*-, *y*- and *z*-components and the summand denotes the shift in *x*-, *y*- and *z*- direction stored in three columns for *m* points stored in *m* rows.

$$X_{n+1} = X_n + [\, n \cdot \vec{d}_i \cdot \frac{s}{l_i} \,]$$

By this, the given layer can be shifted arbitrarily often while maintaining its original shape as can be seen in Figure 11.

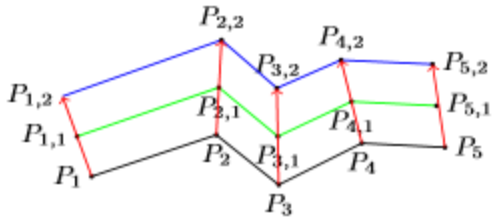


Figure 11. First layer (black) shifted two times (green and blue lines) in the direction of the normal vectors in the points (red).

VIII. Point of Maximum Curvature

The curvature of points can be calculated with an approach given in [2]. In this approach, every single triangle is investigated independently of the surrounding triangles. The curvature directly in the vertices is then set to zero. However, in this project, it is assumed that the number of elements over the whole surface, i.e., the number of triangles, is so large that the triangles can be approximated with planes such that there is no curvature within one individual triangle but the convex shape of the whole structure is obtained by putting the triangles together with different angles of inclination. The triangulated mesh is assumed to be so fine that the differences in angles of inclination of two adjacent triangles never exceeds 90 degrees. Consequently, scalar products of the normal vectors of adjacent triangles are always positive. Therefore, these scalar products of the normal vectors of adjacent triangles can be used to approximate the curvature by the difference in inclination of two adjacent triangles. The 'curvature' of the surface in a given vertex can then be determined by averaging over the scalar products of the adjacent triangles. The smallest scalar product then results in the point of maximum curvature, which can be used in future research.

IX. Conclusion and Future Work

An algorithm for calculating a geodetic line on a given surface, a technique for its further improvement and the numerical derivation of multiple layers on a surface are described. The goal was to derive an accurate numerically determined geodetic line as well as duplication of a given layer. Further steps could feature an extension of the algorithm for geodetic lines, such that several geodetic lines on one surface can be found by iterating over the algorithm. To further improve, analyse and straighten the geodetic line, the unfolding of surfaces with the least distortion could be investigated and automatized [1]. For this, the point of maximum curvature, which was derived in the last section, can be used as the point of investigation because it faces the largest distortion when being flattened. As a reference for the distortion the calculation of metric and angular change of the flattened and convex surface [5] or the elastic potential [6] could be used. At this stage the aim is that the geodetic lines, i.e., the cutting pattern, should be set in a way that the distortion when flattening is minimised because the material is cut in the 2-D plane and is then formed into the desired state or attached on a convex 3-D object. For this process of flattening, the approach given in [7] can be referred to which also presents an approach on how to measure the amount of distortion.

Furthermore, thresholds values for the thickness of layers and width of material stripes could be set with reference to the point of maximum curvature depending on material properties such as Young's Modulus and Poisson's ratio. Both values should never exceed a threshold where too much stress is applied on the material when bending over the maximum point of curvature. This threshold value for the maximum applicable stress could be set to be the ultimate tensile stress when working with prestressing. In [8] basis properties of elasto-mechanical properties of materials with and without prestressing as well as distortion are discussed which could be used as a starting point for future research in this topic.

Additionally, the change in accuracy dependent of the number of elements in the triangulated mesh could be investigated in order to define and optimize the relation between these two quantities.

Summarized, future work could contain research about appropriate thresholds, materials and manufacturing specifications with the aim to reinforce a given structure.

References

[1] A. Pestalozza, S. Weichert, A. Ramezani, H. Rothe, Generation of a Geodetic Line on any Given Surface, The 10th international Conference on Advances in System Simulation: Simulation Based Development of Defense Systems, 2018

[2] J. M. F. Linhard, Numeric-Mechanical Investigation of the Planning Process of Membrane Structures, PhD thesis, Technische Universität München, 2009.

[3] D. Martínez, L. Velho, and P. C. Carvalho, Computing geodesics on triangular meshes, Computers & Graphics 29(5):667 – 675, 2005.

[4] M. Novotni und R. Klein, Computing geodesic distances on triangular meshes, Journal of WDCG, 10(1-2):341-347, 2002.

[5] L.Ju, J. Stern, K. Rehm, K. Schaper, M. Hurdal and D. Rottenberg, Cortical Surface Flattening Using Least Square Conformal Mapping With Minimal Metric Distortion, 2nd IEEE International Symposiu on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821), 2004.

[6] K.-U. Bletzinger, A. Widhammer, Variation of Reference Strategy-A novel approach for generating optimized cutting patterns of membrane structures, Procedia Engineering, 2016.

[7] B. Lévy, S. Petitjean, N. Ray and J. Maillot, Least Squares Conformal Maps for Automatic Texture Atlas Generation, ISA (Inria Lorraine and CNRS), 2002.

[8] D. Ströbel, The Usage of the Adjustment Calculation of Elasto-Mechanical Systems, 1997.

Development of Flexible and Lightweight Ballistic Body Armor

Comparative Ballistic Studies on Different Ultra-High-Molecular-Weight Polyethylene Materials

Henrik Seeber, Arash Ramezani
Chair of High-Speed Dynamics
Helmut Schmidt University - University of the Federal Armed Forces
Holstenhofweg 85, 22043, Hamburg, Germany
Email: henrik.seeber@hsu-hh.de, ramezani@hsu-hh.de

***Abstract*—In the last years, the risk to become a victim of a gun fight or to be involved into an amok situation increased. This leads to a higher demand for ballistic protection of civilians. Therefore, another capability profile for ballistic armor is necessary, compared to the military or police sector. The focal point of civil ballistic armor is wearing comfort, weight and invisibility. This paper provides information about the development of flexible and lightweight ballistic body armor. The manufacturing process is showed, heat- and sag test are presented. Different ultra-high molecular weight polyethylene (UHMWPE) composites, in this case made by Dyneema®, are compared by using ballistic tests and will be discussed below. In this test series Dyneema® hard ballistic (HB) 26, HB50 and soft ballistic (SB) 115 is used. Aim of the project is to create a ballistic body armor, which is flexible, and body fit enough to be worn under a sweater or suit. Furthermore, the body armor should meet the fourth level of ballistic protection of the "Association of test laboratories for bullet resistant materials and constructions" (VPAM 4) as well as the requirements of the VPAM ballistic protection vests (BSW). The idea can be realized by using an UHMWPE composite. A material with well-balanced properties to fulfill the ballistic and mechanical requirements, simultaneously providing the necessary wearing comfort.**

Keywords - defense engineering; ballistic body armor; fiber-reinforced plastics; ballistic trials; material processing.

I. INTRODUCTION

Based on the findings in [1] this article provides further ballistic trials. These trails are evaluated and first series relevant findings are achieved.

Personal integrity is a basic need. Nowadays, this need is endangered by increased incidents relating to gun violence. Especially in the USA, the number of incidents rises from about 50,000 to over 60,000 between 2014 and 2017 [2]. This leads to a higher need for personal ballistic protection at the civilian markets. Included are products like soft-ballistic sweater inlays or discrete ballistic vest, which are suitable for everyday life. In this area of use weight, wearing comfort and invisibility are focal points. The ballistic protection up to the fourth level of ballistic protection of the "Association of test laboratories for bullet resistant materials and constructions" (VPAM 4) has to be ensured. These requirements make it necessary to use a material, which combines high tensile strength for a high ballistic performance and low density for a suitable weight balance. These attributes are combined in many fiber reinforced composite materials, especially ultra-high molecular weight polyethylene (UHMWPE) is well suited for the mentioned application. Already existing analyses of the ballistic behavior of UHMWPE, like in [3] or [4] described, are often on a theoretical level. This paper aims for an analysis on an applied level, with a concrete connection to a product development. Existing theoretical results are applied to the development of an actual ballistic vest. This leads to the overall aim of the project to create a flexible and lightweight ballistic vest, which meet the VPAM 4 regulations and can be worn under everyday clothes. Therefore, the project is divided into five sections:

1. Material processing (Section IV);
2. Material pre-testing (Section V);
3. Pre-ballistic testing (Section VI);
4. Ballistic testing (Section VII);
5. Finalization (future work).

This paper aims to provide data to derive a decision about the material composition, which is used for the finalization. The decision should be especially based on the results of the ballistic testing, mentioned in Section VII.

The paper is structured as follows. In Section II, preliminary considerations are introduced regarding material, processing and shape of the test objects. Section III gives a brief information about the different used UHMWPE prepreg materials and the second matrix material. Section IV is about the processing of ballistic plates namely cutting and lamination process. In Section V, heat resistance test and sag test are presented and discussed. In Section VI, the pre-ballistic test, which includes VPAM 3 (third level of ballistic protection of the "Association of test laboratories for bullet resistant materials and constructions") and VPAM 4 testing, is described and evaluated. Section VII is about the ballistic test. It includes the basic testing of the newly used materials and further, a comparative ballistic test to show the differences (in contrast) to the pre-ballistic test. The final section, Section VIII, merges all results and, these results are discussed, leading to the constructive design of the plates.

II. PRELIMINARY CONSIDERATIONS

In the beginning of the project two questions occur:

- Which material is suitable for the project (ballistic performance and weight)?
- How can it be processed to become a flexible ballistic plate?

First thoughts about the material leads to the UHMWPE in detail Dyneema® HB26. This material, in shape of hard-

ballistic-plates, was successfully tested in previous ballistic trails. The plates were made of multiple layer of pre-impregnated fiber (prepreg) material, which are fused under a certain pressure and temperature to become a solid ballistic plate. These ballistic trails have already been reported in [4].

This type of HB26 plates have an inflexible structure, making them unsuitable for this project of soft-ballistic-plates. On the other hand, the material shows a good ratio between weight and ballistic performance. In detail, areal density of the material is between 257 – 271 g/m² [5] and has an energy absorption per areal density of ~35 J·m²/kg [5].

Nowadays, newer materials are available, like the improved Dyneema® HB50 or the Dyneema® SB115. The HB50 is an enhanced HB26. The basic design is consistent, but the areal density is lower (226 – 240 g/m²) and the ballistic properties of the material have to be checked and compared in Section VII [6]. SB115 is a prepreg material, consisting of UHMWPE fibers. This material is especially produced for flexible and lightweight body armor applications, due to the very low areal density (75 – 84 g/m²) and notably flexible matrix material [7]. In Section III the mechanical properties of these materials are explained in detail to enhance the awareness for the material.

Relating to the second initial question, Dyneema® HB26 provides a promising starting point, because this basic prepreg already offers a flexible structure. Therefore, the prepreg material only needs a second flexible matrix material to hold the prepreg layers together. Due to this, a lamination process by hand is selected, which makes it possible to use a flexible cast resin. This leads to the question, which kind of cast resin has to be used as a second matrix material. It has to be considered that, firstly, the cast resin does not destroy the chemical basic structure of the prepreg material and secondly, the cast resin remains flexible. To ensure this chemical compatibility of the second matrix material the same matrix material, like it is used for the prepreg, is chosen. For that reason, a polyurethane (PUR) determines the second matrix material. This is selected out of the group of thermoplastic elastomers (TPE) [5]. Under the aspect of flexibility, a cast resin out of PUR with special flexible properties is chosen. A further description of the second matrix material is given in Section IIID.

Beside the material, also the shape of the plates is important for the ballistic performance. This is, because of the anisotropic properties, which are described in Section IIIA. In this case, the shape was predetermined, because the plates have to fit into already existing structures. Relating to the standard size of a plate carrier inlay, plates were produced in a size of 300 mm x 250 mm. Furthermore, patterns for a sweatshirt inlay were used to evaluate the behavior of the different shapes in a ballistic trail. The main body pattern has a width of 435 mm and a total height of 415 mm (Figure 1). The side body pattern is narrower and has a width of 150 mm and a height of 205 mm (Figure 2) [1].

III. USED MATERIALS

The basic material for the development of the flexible plates is UHMWPE prepreg material. A prepreg material is fiber composite material where the fibers are already combined with the matrix material. Various numbers of sublayers are implemented into a prepreg.

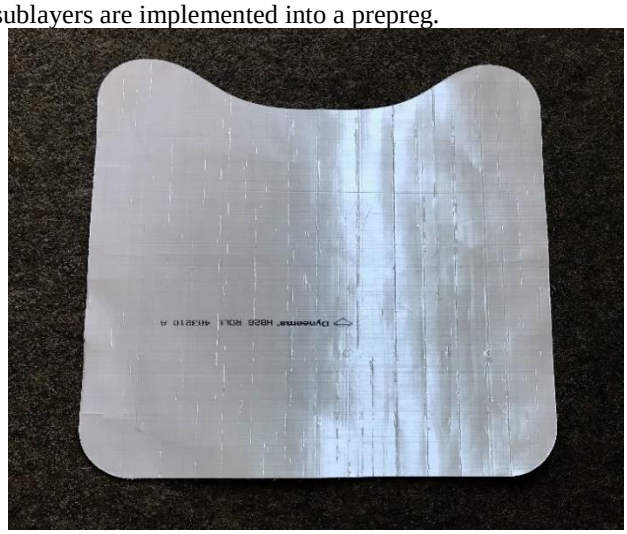


Figure 1. Main body pattern out of Dyneema® HB26. Width of 435 mm and a total height of 415 mm.

Figure 2. Side body pattern out of Dyneema® HB26. Width of 150 mm and a total height of 205 mm.

In this case Dyneema® HB26, HB50 and SB115 is used as a main material. As mentioned in Section II, the second matrix material is a flexible PUR.

A. Dyneema® HB26

It is shipped in a shape of a tape and as such, the matrix and fiber have already combined to a prepreg. The fibers are out of UHMWPE and the matrix out of PUR. The prepreg material consists of four sublayers of UHMWPE fibers, which are bidirectional orientated. The fiber direction per sublayer is turned by 90°, compared with the previous layer. This achieves an equal force dissipation in the prepreg material. Resulting from this structure, different geometric shapes of the material have anisotropic tensile properties, because of unequal fiber length.

The prepreg material has a density of approximately 0.85 g/cm³, due to a martial thickness of 0.3mm. UHMWPE fibers have a high molecular weight. Commonly they have an Intrinsic Viscosity (IV) from 8 IV up to 30 IV. The fiber

achieves high tensile strength (between 3.21 GPa and 5.99 GPa) and initial modulus (between 113GPa and 171 GPa) through long molecular chains out of methylene groups (CH_2). A characteristic of UHMWPE is that the intra molecular bounds are relative weak Van der Waals bounds. Though, the extreme long molecule chains leading to a significant overlap between the molecule chains. This results in many inter molecular Van der Waals bounds, which strengthen the overall intermolecular stability. Further developments have to consider that the fiber length is one of the most important constructional attributes, because of the link to the tensile strength of the material [8].

The matrix material of the prepreg has multiple tasks. Firstly, to protect the fiber against environmental conditions, pressure and kinks. Secondly, to hold the fibers in their position and direction. Finally, the matrix leads the forces into the fibers. Matrix material is often chosen from the group of thermoplastic elastomers, because of moldability and high elongation at break. These are good properties to support the fibers and the construction [1].

B. Dyneema® HB50

This UHMWPE material is an improved HB26 with the same basic physics. Also, the mechanical properties, especially the tensile strength (between 3.21 GPa and 5.99 GPa) and initial modulus (between 113GPa and 171 GPa), are the in same range as HB26. Section VII shows that these two mechanical properties of HB50 are (slightly) better than those of HB26, due to the improved ballistic results. It is quite possible that the tensile strength and initial modulus of HB50 is located in the upper area of the mentioned range. With 0.753 g/cm^3 HB50 has also a lower density compared to HB26 [6].

C. Dyneema® SB115

Dyneema® SB 115 is also a UHMWPE based fiber reinforced composite material. It is specially designed for soft armor applications, due to its flexible and soft matrix material and low areal density of 75 – 84 g/m². The prepreg material consist of two sublayers which are bidirectional orientated (0°, 90°), thus it is a cross ply. Additionally, the whole prepreg is covered with a coating, to protect the prepreg material against environmental damages, like chemicals and water [7]. Overall, one layer of the material has a thickness of 0.2 mm. This leads to a density of approximately 0.375 g/cm^3. The SB 115 consists of the fiber SK99. This fiber has a tensile stretch of 3.5 % and a tensile strength of 3.92 GPa. This high strength is based on the so called "Dyneema® Force Multiplier Technology" [9].

D. Cast Resin System - Second Matrix Material

The most important requirement for second matrix material is to preserve the flexibility of the prepreg. Therefore, the second matrix material has also to be flexible. As mentioned in Section II, the flexible PUR called R15GB-flex was selected. This material is a two-components cast resin system, consisting out of resin and hardener. With a mixing ratio of 100 parts resin and 25 parts hardener a Shore hardness (A) of ~40 is achievable. Compared with normal PUR cast resin systems, which have a Shore hardness (D) of ~65 (comparable to a Shore hardness (A) of ~105), the used PUR has a softer texture [10]. Another important property is the density of the second matrix material to keep a low overall weight. The material R15GB-flex has a density of 1.1 g/cm^3, comparable with the density of the prepreg material. The material has in mixed condition a medium viscosity and a processing time of approximately 15 min (100 g at 20 °C). These properties make the material suitable for the lamination process by hand. The maximum usage temperature is around 50 °C [8]. This fact will further be discussed in Section V [1].

IV. PROCESSING

The processing of the ballistic plates is divided into two stages:

- Cutting process for semi-finished parts;
- Lamination process for the finished composite material.

This will only be presented in detail for the plates, which were constructed for pre-tests and pre-ballistic test (Table I). The processing procedure for the plates used in the ballistic test are the same and can be abstracted. An overview of these plates is given in Table II.

A. Cutting Process

This project stage needs two patterns. First of all, the standard plate carrier inlay with rectangular, plane shape and dimensions of 300 mm x 250 mm. For the testing procedures one 5-layer, two 10-layer, one 15-layer and two 20-layer plates are produced (Table I). Overall, 80 layers of prepreg material in this shape are necessary.

Furthermore, patterns for a sweatshirt inlay, with a main body part (Figure 1) and side part (Figure 2) are used. On the whole, 30 main body part layers and 15 side part layers are produced (Table I).

The lamination process by hand makes it necessary to laminate every single layer. In this particular case, a fast-laser-cutting-process is unsuitable, because the individual layers would melt together. A single layer cut with the laser cutter is as well inefficient, due to the rectangular shape. For that reason, every layer has to be cut out by a special scissor for reinforced fibers.

To achieve a high precision, the more complex shapes of the sweatshirt patterns are cut by the laser. The patterns are replicated in a computer-aided design (CAD) program and saved as drawing exchange format (DXF) file. Lastly, the file is exported into the laser cutter program and executed (parameters: Power 120 W, frequency 1000 Hz, velocity 0.036 m/s). Due to the four-sublayer structure of the prepreg material, a homogeneous sublayer structure is achievable even without rotation between the prepreg layers [1].

B. Lamination Process

The hand lamination process requires following stages:

- Mixing;
- Coating;
- Hardening process.

TABLE I. OVERVIEW OF THE CONSTRUCTED TEST OBJECTS FOR THE PRE-TEST AND PRE- BALLISTIC TEST

Plate Number	Test Number	Material	Layers	Thickness [mm]	Weight [g]	Shape	Processing Method
1	1	Dyneema® HB26	10	4	295.8	Rectangular	Lamination
2	2	Dyneema® HB26	10	4.2	295.9	Rectangular	Lamination
3	3	Dyneema® HB26	15	6.8	448.3	Rectangular	Lamination
4	4	Dyneema® HB26	20	9.2	614.8	Rectangular	Lamination
5	5M	Dyneema® HB26	20	9	592.6	Rectangular	Lamination
	5R						
6	/	Dyneema® HB26	5	2	159	Rectangular	Lamination
7	7	Dyneema® HB26	15	10	/	Side Part	Loose
8	8	Dyneema® HB26	10	6.4	/	Main Body Part	Loose
9	9	Dyneema® HB26	20	13.3	/	Main Body Part	Loose

TABLE II. OVERVIEW OF THE CONSTRUCTED TEST OBJECTS FOR THE BALLISTIC TEST

Plate Number	Test Number	Material	Layers	Thickness [mm]	Weight [g]	Shape	Processing Method
12	12M	Dyneema® HB50	15	6.5	395.6	Main Body Part	Loose
	12L						
	12R						
13	3M	Dyneema® SB115	15	4.1	197.8	Main Body Part	Loose
	13L						
	13R						

In the mixing process, the mixing ratio is adjusted by the proportion of weight. For a low Shore hardness (A) the manufacturer recommends a mixing ratio of 100 parts resin and 25 parts of hardener. Proportion of weight of the two components, which is necessary for the different plate sizes, are shown in Table III. The two components are mixed with a wooden stick to a homogenous mixture. This mixture is equal spread over the prepreg layer with a lamination brush. For the coating, a slight film is sufficient. Pressure from the inside to the outside is applied to get a compact compound.

TABLE III. PROPORTION OF WEIGHT OF RESIN AND HARDENER FOR DIFFERENT AMOUNT OF LAYERS

Layer	Resin [g]	Hardener [g]	Final Thickness [mm]
5	42.6	11.55	2
10	85.2	23.1	4.2
15	127.8	34.65	6.8
20	170.4	6.2	9.3

Finally, surplus material is removed, and the plates are laid into a warm place for 24 h for hardening. After the hardening process, the plates are inherently stable [1].

V. HEAT AND SAG TEST

Besides the ballistic performance, further two primary material attributes of the new composite material have to be tested. Firstly, the heat stability in a heat test, because of possible high surface temperatures in area of use. Secondly, the flexibility of the material, as key functionality in a sag test.

Because of the nearly equal mechanical properties of HB26 and HB50, these tests were skipped for HB50 and will be done in future work. Based on the ballistic results, these tests will not be conducted for SB115.

A. Heat Test

Heat is a weak point of the second matrix material, due to its thermoplastic properties. Maximum temperature of usage is around 50 °C according to the manufacturer. Analyses of the possible area of use show that the average temperatures are moderate temperatures between 10 °C and 30 °C, which

are unproblematic for the composite material. Nevertheless, in desert areas, which are possible areas of use, surface temperatures can reach a maximum around 70 °C and daily surface temperatures around 60 °C [11]. To test the heat resistance of the composite material, one plate is faced a heat test.

The heat test proceeds as follows: an oven is preheated to a temperature of 30 °C. This 30 °C stage is used as a reference result. Every 30 min the temperature is set to a new value, first 50 °C, then 60 °C and finally 70 °C. The plate is left in the oven for 30 min. After 15 min and 30 min the plate is taken out and checked regarding: degeneration, flexibility, slip of layers and defects. A 5-layer plate (Plate No. 6) is chosen to ensure an even temperature distribution in material. Results are displayed in Table IV.

The results show that the composite material, and especially the second matrix material withstand temperatures up to 50 °C without any property changes. At higher temperatures above 50 °C an increased flexibility is recognizable. After 30 min at 70 °C the layers of prepreg are movable 3 mm against each other. After this movement, the layers go back to their initial state. At all temperature stages the plate shows no degeneration and defects. Following a cooling phase, the second matrix material solidified again. Summing up, the higher temperatures are uncritical to the composite material [1].

B. Sag Test

Flexibility is a key attribute of the used type of ballistic protection. Comparative values of the plate flexibility can be generated through a sag test. The aim of this test is to measure the sag of the plate under a certain load.

Therefore, a test procedure is created. The test setup is shown in Figure 3. The plates are laid onto two bars with a contact area of 10 mm x 250 mm on both sides. Moreover, the basic test setup consists of two rulers and a wooden baseplate (35 mm x 145mm), which are arranged as in Figure 3 shown. The sag test was conducted at 22 °C.

As a result of this arrangement, a basic load of 72.8 g lays up on the plate. Additionally, a 500 g block is used as a test weight. The center of the wooden baseplate is positioned in the middle of the plate at 150 mm x 125 mm.

TABLE IV. RESULTS OF THE HEAT TEST

Tempe-rature [°C]	Time [min]	Degene-ration	Flexibility	Slip of Layers [mm]	Defects
30	15	Non	Unchanged	Non	Non
	30	Non	Unchanged	Non	Non
50	15	Non	Unchanged	Non	Non
	30	Non	Unchanged	Non	Non
60	15	Non	Slightly increased	Non	Non
	30	Non	Slightly increased	Non	Non
70	15	Non	Increased	>1	Non
	30	Non	Increased	3	Non

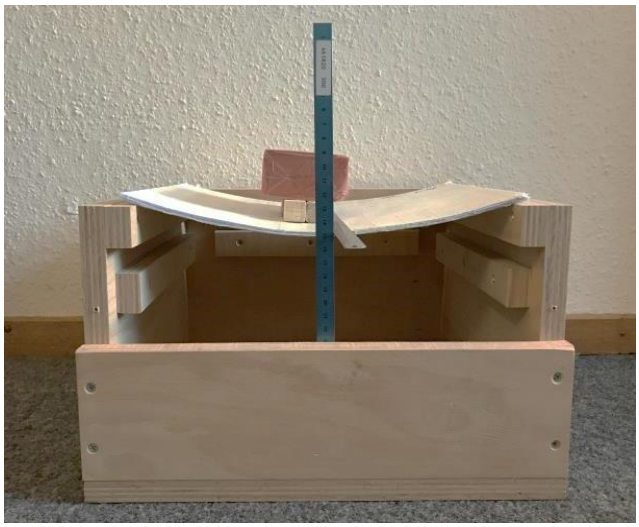


Figure 3. Sag test basic test setup with test weight.

One ruler is placed for measurements and the other as an indicator for it. The zero height is 120 mm.

First, the initial height of the plate is measured. In this phase the plate has an additional load of 72.8 g. After 30 sec under these conditions, the height is measured.

Subsequently, the plate is loaded with test weight. After 10 sec under these conditions, the height is measured again. Results are displayed in Table V. The results show that 20-layer plates have deficits in their sag values and thus in the overall flexibility.

They may be too inflexible for this particular application area. The 15-layer plate has a good ratio between number of layers and flexibility, compared with the 10-layer plate [1].

VI. PRE-BALLISTIC TEST

The pre-ballistic test is performed on the basis of the VPAM ("Association of test laboratories for bullet resistant materials and constructions") regulations. Especially following regulations are regarded: general basis for ballistic material, construction and product testing (APR) [12], ballistic protective vests (BSW) [13] and bullet resistant plate materials (PM) [14].

The aim of this project is to meet the regulations of VPAM 4. This level requires that the plate withstand a penetration of a .357 Mag. fired with a projectile velocity of 430 ± 10 m/s from 5 m distance [12]. A ballistic placement test provides an overview of the ballistic performance of the composite material. Therefore, a modified VPAM 3 (9 mm, 415 ± 10 m/s, 5m) and VPAM 4 level is tested.

TABLE V. RESULTS OF THE SAG TEST

Plate Number	Initial Height [mm]	End Height [mm]	Delta (Sag) [mm]
1	131	165	34
2	143	183	40
3	129	153	24
4	123	128	5
5	125	128	3

The shooting distance is increased to 10 m. Penetration and back face deformation are evaluated. Based on VPAM BSW No. 4.2 a maximal transmitted energy of 70 J is acceptable [13].

A. Preparation and Test Setup

Two components are necessary for the test setup: shooting-box and plasticine (Figure 4). The shooting box is built in consideration of the VPAM BSW [13]. The inner dimensions of the shooting-box are 300 mm width, 250 mm height and 150 mm depth. Especially the characteristic depth is important, because of compression effects with the rear panel. Based on the VPAM PM contact areas of 30 mm on three sides are built in [14]. The used plasticine is recommended by VPAM (VPAM BSW No. 5.2) [13]. The ballistic test is conducted at an ambient temperature of 21 °C.

A measurement of the plasticity is conducted as described in VPAM BSW No. 5.2.1. The mean imprint depth (d_m) of the plasticine is 19.8 mm. With this value and the maximal transmitted energy (E_{max}) of 70 J, the maximal volume (V_{max}) of the back-face deformation is calculated as shown in (1) [12].

$$V_{max} = (0.134 \cdot d_m - 1.13) \cdot E_{max} \quad (1)$$

Using the values of the mean imprint (19.8 mm) and maximal transmitted energy (70 J) in (1) leads to (2).

$$V_{max} = (0.134 \cdot 19.8 - 1.13) \cdot 70 = 106 \quad (2)$$

This leads to a maximal back-face deformation volume of 106 cm^3.

The gun, which is used for VPAM 3 testing is a SIG Sauer X-Five® with a 9 mm Luger 124 Gr. projectile. For the VPAM 4 testing an S&W 686 European Match® with a .357 Mag 158 Gr. projectile is used (Figure 5).

Figure 4. Shooting-box filled with plasticine. Inner dimensions: 250 mm x 300 mm x 15 mm.

Figure 5. Left side: .357 Mag 158 Gr. Right side: 9 mm Luger 124 Gr.

All shots are fired into the middle of the plates. The only exception is test number B5R, which get shot into the edge area. This edge shot has to hit the material in a distance of 30 ± 5 mm from an edge [13]. Test number B8 and B9 were shot without the shooting-box to check the penetration level of the loose layers [1].

B. Evaluation

After every shot the diameter (d) and the depth of the imprint (t) of the back-face deformation is measured (Table VI). With these values, and the volume equation of a circular cone the volume of the back-face deformation (V) is calculated (3).

$$V = \frac{1}{3} \cdot \pi \cdot \left(\frac{d}{2}\right)^2 \cdot t \quad (3)$$

Additionally, the transmitted energy (E) is calculated backwards with following equation [8]:

$$E = V/(0.134 \cdot d_m - 1.13). \quad (4)$$

The results show that test number 1, 3, 4, 5M, and 7 meet the requirement that the transmitted energy is lower than 70 J. Especially test number three (15 layers) (Figure 6) shows a good average transmitted energy, comparing to test number 1 (10-layers) and 5M (20 layers). Figure 7 shows the back-face-deformation in the clay after shot three. The deformation is wider and not as deep as in the other tests. This is a good result, because deeper back-face-deformation can more likely lead to internal injuries. For that reason, a flatter and wider imprint is strived. Test number 5R failed, because the projectile left the plate before it gets stuck in the plate. As expected, test number 2 got perforated by the .357 Mag projectile. This test was conducted to see the penetration behavior of the material and projectile, like layer and projectile movement, deformation of the projectile and damage to the plate. At test number 8 and nine the projectile got stuck in the fourth layer of the loose layers. The back-face deformation of these test numbers is negligible [1].

TABLE VI. RESULTS OF THE BALLISTIC TEST PHASE 1

Test Number	Layers	Caliber	Imprint -			Transmitted Energy [J]	Comments
			Depth [mm]	*Diameter [mm]*	*Volume [cm³]*		
1	10	9 mm	34.4	80	57	37.840	/
2	10	.357	/	/	/	/	Perforation
3	15	9 mm	26.0	70	33	21.897	/
4	20	.357	29.8	90	63	41.487	/
5M	20	9 mm	18.3	50	11	7.863	/
5R	20	9 mm	/	/	/	/	Leaving material
7	15	9 mm	40.4	70	51	33.940	/
8	10	9 mm	/	/	/	/	Stuck in fourth layer
9	20	.357	/	/	/	/	Stuck in fourth layer

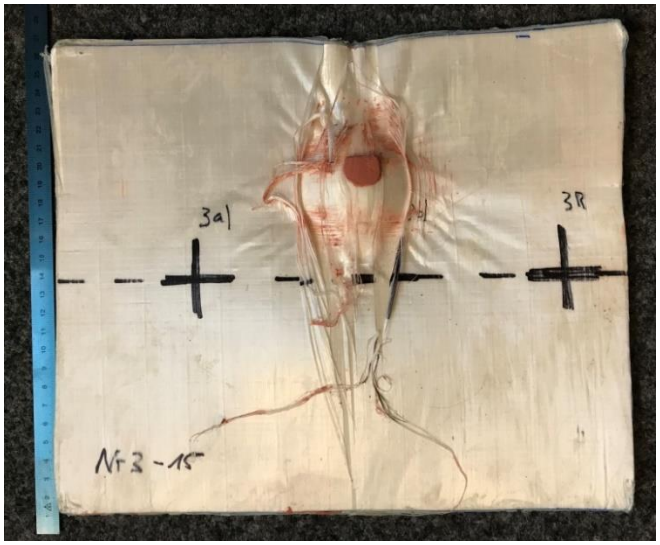


Figure 6. Test number 3 after the shot. Seeable is the back-face deformation and material behavior.

Figure 7. Test number 3 after the shot. Seeable is the volume of the back-face deformation in the clay.

VII. BALLISTIC TEST

The ballistic test is performed on the base of the VPAM APR [12] and VPAM BSW [13]. The aim of this test is to verify the shirt patterns for VPAM 3. This level is chosen, because there were no initial experiences with these materials. VPAM 3 requires that the plate withstand a penetration of a 9 mm Luger fired with a projectile velocity of 415 ± 10 m/s from 5 m distance [12]. Additionally, based on VPAM BSW, it has to withstand two additional shoots with a spacing of 75 mm arranged like an equilateral triangle [13].

A. Preparation and Test Setup

To get an idea of the back-face deformation while using the shirt patterns, a new test set up was constructed (Figure 8). This is necessary, because they do not fit into the shooting box, which was used in the pre-ballistic test. This time it is constructed completely based on VPAM BSW 5.1. The inner dimensions of the shooting-box are 400 mm width, 350 mm high and 150 mm depth. The shirt patterns are fix with Velcro fastener to the shooting-box. The used plasticine is recommended by VPAM (VPAM BSW No. 5.2) [13]. The ballistic test is conducted at an ambient temperature of 19 °C.

A measurement of the plasticity is conducted as described in VPAM BSW No. 5.2.1. The mean imprint depth (d_m) of the plasticine is 20.2 mm. Equation (1) is used to calculated the admitted maximal back-face deformation volume.

Figure 8. Test number 12M after the first shot. Seeable is the principle test setup, based on VPAM BSW.

$$V_{max} = (0.134 \cdot 20.2 - 1.13) \cdot 70 = 110.376 \quad (5)$$

This leads to a maximal back-face deformation volume of 110.376 cm^3.

The gun, which is used, is a H&K® MP5 with a 9 mm Luger 124 Gr. projectile. The velocities are measured with a light barrier measurement system. The average projectile velocity is 410.7 m/s, thus in the allowed range of the VPAM APR. The first shoot is fired into the middle of the of the plate (12M, 13M). The second shot is fired onto the left side of the first shoot, with a spacing of 75 mm (12L, 13L). The third shot is fired onto the right side of the second shoot, with a spacing of 75 mm, to get an equilateral triangle (12R, 13R).

B. Evaluation

After every shoot depth and diameter of the imprint is measured. With these values, the volume of the imprint with (3) and then the transmitted energy with (4) is calculated. The results are displayed in Table VII.

The results show that all shoots of test number 12 and the first shoot of test number 13 (13M) meet the requirements of VPAM BSW, because the transmitted energy is lower than 70 J. At test numbers 13R and 13L the projectile perforated the material, thus they do not meet the requirements of VPAM BSW. The front and the back side of test number 12 is shown in Figure 9, test number 13 in Figure 10. In the penetration process the kinetic energy of the projectile is partially dissipated into the deformation of the projectile (Figure 11). A flat deformed projectile shows an acceptable energy dissipation duo to the mechanical properties of the fiber. A nearly intact projectile shows that the fibers absorb none of the projectile energy.

TABLE VII. RESULTS OF THE BALLISTIC TEST

Test Number	Imprint -			Transmitted Energy [J]	Comments
	Depth [mm]	*Diameter [mm]*	*Volume [cm³]*		
12M	45.88	71	60	38.09	Stuck in ninth layer
12L	35.54	63	36	23.42	Stuck in ninth layer
12R	37.27	68	45	28.61	Stuck in ninth layer
13M	32.2	62	32	20.55	Stuck in fifth layer
13L	/	/	/	/	Perforation
13R	/	/	/	/	Perforation

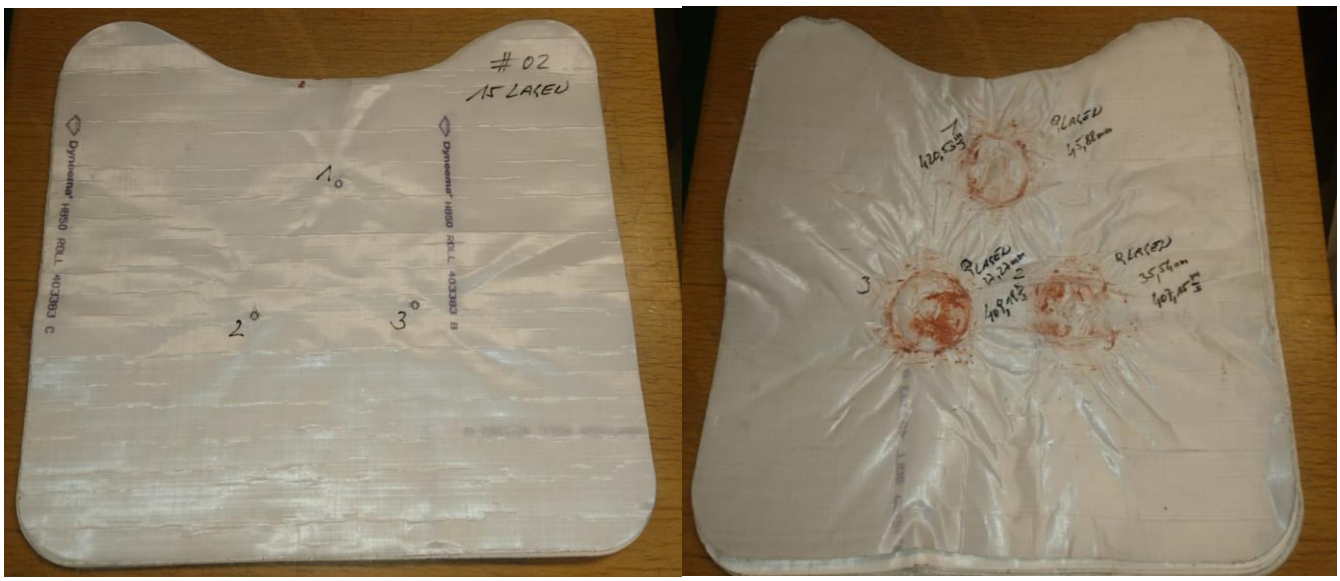


Figure 9. Left side: the frontside of test number 12. Right side: the backside after the shooting. Readable on the backside is the stuck layer, back-face deformation depth and projectile velocity.

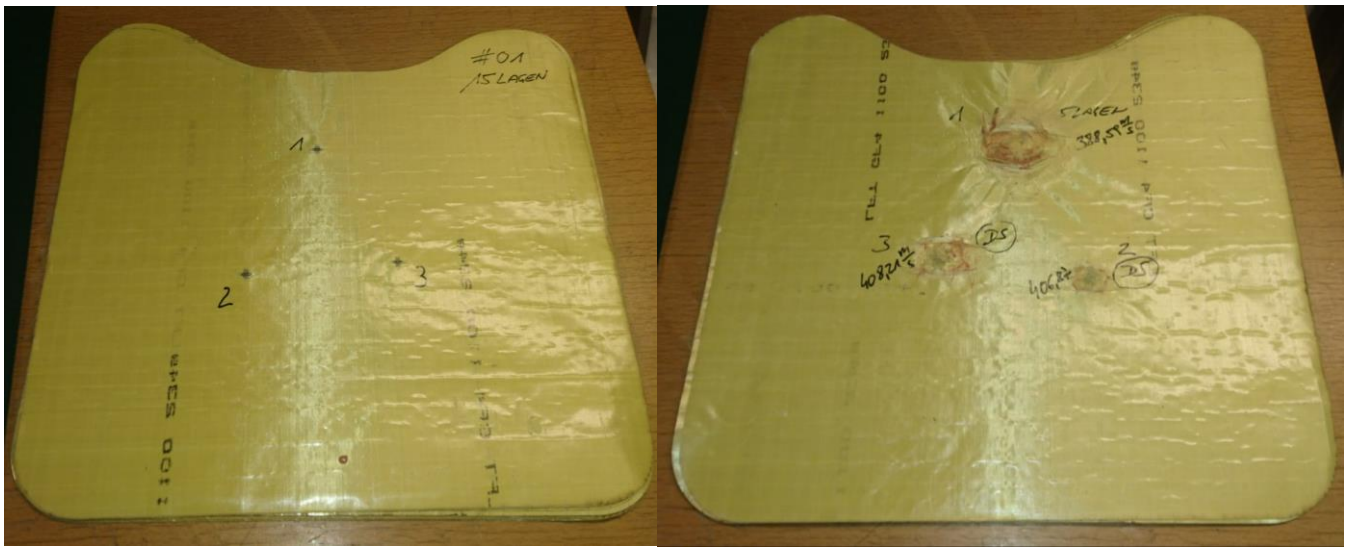


Figure 10. Left side: the frontside of test number 13. Right side: the backside after the shooting. Readable on the backside is the stuck layer (DS means perforation) and projectile velocity.

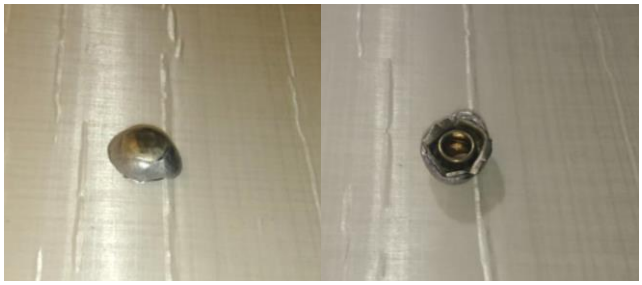


Figure 11. Deformed projectile of test number 12M.

VIII. DISCUSSION

The pre-testing and pre-ballistic test resulted in following findings. The heat-resistance-test shows that the composite plates are durable up to 70 °C. At higher temperatures a slip of layers is highly possible. Up to 60 °C no degeneration or melting is detectable (Table IV). Thus, the second matrix material and main material are suitable for the estimated conditions.

To examine the flexibility, a sag test is conducted. It shows that 20-layer plates are too stiff for this area of use. 10- and 15-layers plates show good values of flexibility and weight compared to other soft ballistic plates (Table V).

The ballistic plate has to meet the requirements of VPAM 3 and VPAM 4. This requirement allows a maximal transmitted energy of 70 J. In a first placement test all expected plates meets this requirement by far (Table VI). Also, the sweatshirt inlays were tested successfully. It seems that the extensive shape, which leads to longer fibers, increases the ballistic performance.

To summarize, the 15-layer plate is the best performing plate, because of the ratio between flexibility, weight and ballistic performance. 10-layer plates will not meet the VPAM 4 regulations, and 20-layer plates are too stiff and heavy for the area of use. These results will put the 15-layer plate into the focus of the studies. It has to be considered that these ballistic results are from a pre-testing series. They are just indicators for a relative rating between our ballistic plats. In further ballistic tests they have to be verified for the VPAM 4 level.

The ballistic test with the new materials confirms that the 15-layer Dyneema® HB solution meets the VPAM 3 and BSW regulations. In fact, the average amount of transmitted energy is 29.6 J. This is only 40 % of the allowed value and 12 % less transmitted energy compared to the HB26 plates. Thus, HB50 shows a promising ballistic performance to meet VPAM 4 regulations, which have to be tested in future work. Moreover, because of the lower density, the shirt patterns out of HB50 are 11.4 % lighter, compared to the predecessor material.

The SB115 solution showed good first-shoot-capabilities, with the lowest transmitted energy of the ballistic test but a perforation after the second and third shoot. Thus, SB115 missed the requirements of VPAM BSW and will no longer be considerate in the project.

IX. CONCLUSION AND FUTURE WORK

All in all, the aim of this project is to create lightweight and flexible ballistic body armor for the civilian market. Therefore, a UHMWPE prepreg as main material and a two-component flexible resin as second matrix material is used. Through a lamination by hand procedure, ballistic composite plates are produced. Because of this method, the flexible properties of the main material are sustained. Moreover, a temperature resistance of the composite material up to 60 °C is identified. The conducted pre-tests (heat and sag test) show that the 15-layer plate meet our expectations the best, in the area of flexibility and durability. This leaded to further ballistic tests with 15-layer thick plates. Especially, the ballistic results of the plate are promising. The patterns out of this material already meet the requirements of VPAM 3 and VPAM BSW. Furthermore, compared to other available

products on the market, the new plate has a reduced weight of ~150 g, is 1.2 mm slimmer and full flexible.

Further scientific work, will be focused on the HB50. This includes that 15-layer sweatshirt patterns made of HB50 will be extensive ballistic tested. Especially, VPAM 4 and VPAM BSW testing will be focused. A flatter and wider back-face-deformation imprint is strived. Moreover, the heat and sag test for this material will be done, in order to get sure about the material properties relating durability and flexibility. For all these tests, 15-layer sweatshirt patterns will be produced with the lamination process by hand. Afterwards, the finalization process will be begun. This process includes investigations of coatings and edge support, to protect the ballistic plates against environmental conditions and make them ready for mission.

REFERENCES

[1] H. Seeber and A. Ramezani "Development of Flexible and Lightweight Ballistic Body Armor, Constructional Contributions for Ballistic Components out of Ultra-High-Molecular-Weight Polyethylene", The Tenth International Conference on Advances in System Simulation (SIMUL2018) IARIA, pp. 46-52, Oct. 2018.

[2] Gun Violence Archive, "Past Summary Ledgers", [Online]. Available from: http://www.gunviolencearchive.org/past-tolls [09.2018].

[3] L.H. Nguyen, "The Ballistic Performance of Thick Ultra High Molecular Weight Polyethylene Composite". RMIT University, Dec. 2015.

[4] H.v.d. Werff and U. Heisserer, "High performance ballistic fibers: Ultra-High Molecular Weight Polyethylene (UHMWPE)" in "Advanced Fibrous Composite Materials for Ballistic Protection", First Edition. Woodhead, Oct. 2016.

[5] DSM Dyneema, "Product Specification Sheet HB26", Feb. 2014.

[6] DSM Dyneema, "Product Specification Sheet HB50", Feb. 2014.

[7] DSM Dyneema, "Product Specification Sheet SB115", Feb. 2014.

[8] T. Tam and A. Bhatnagar, "High-performance ballistic fibers and tapes" in "Lightweight Ballistic Composites: Military and Law-Enforcement", Second Edition. Amsterdam: Elsevier, May. 2016.

[9] M. Hendrix and M. Herzog, "Ballistic Behavior of Soft-Ballistic UHMWPE" ("Schutzverhalten von weich-ballistischen ultrahochmolekularem Polyethylen"), in "Wissenschaftliche Beiträge 2018", Technical University Wildau, Mar. 2018.

[10] HP-Textiles, "HP-R15GB-flex Data Sheet", Aug. 2017.

[11] EOSDIS Worldview, "Land Surface Temperature (Day)" [Online]. Available from: https://go.nasa.gov/2P3DxzW [09.2018].

[12] Association of test laboratories for bullet resistant materials and constructions (VPAM), "VPAM APR 2006 Version 2", Nov. 2014.

[13] Association of test laboratories for bullet resistant materials and constructions (VPAM), "VPAM BSW 2006", May. 2009.

[14] Association of test laboratories for bullet resistant materials and constructions (VPAM), "VPAM PM 2007 Version 2", Jan. 2014.

Real-Time Lighting of High-Definition Headlamps for Night Driving Simulation

Nico Rüddenklau, Patrick Biemelt, Sven Henning, Sandra Gausemeier and Ansgar Trächtler
Chair of Control Engineering and Mechatronics, Heinz Nixdorf Institute, University of Paderborn
Paderborn, Germany
Email: {nico.rueddenklau, patrick.biemelt, sven.henning, sandra.gausemeier, ansgar.traechtler}@hni.upb.de

***Abstract*—Introducing high-definition headlamp systems in the automotive industry opens up a wide range of possibilities for improving existing and developing new types of dynamic lighting functions. Due to the complexity and subjectivity of light distributions of modern headlamp systems, simulation-based development is indispensable. Strong restrictions regarding the time of day and weather conditions as well as the hardly reproducible traffic situations are further reasons to shift the testing of such systems as much as possible from reality to simulation. This contribution presents a first real-time simulation of high-definition systems in virtual environments. The simulation results are validated by measuring data and evaluated using validated software for simulating static light distributions at night. The performance of the implementation in terms of computational effort is also discussed. As it turns out, the presented implementation is well suited in terms of appearance and computational performance as a basis for a night driving simulation.**

Keywords–high-definition headlamp; real-time simulation; night driving simulation; dynamic light function; shader; lighting.

I. Introduction

The first implementation of a real-time capable simulation of high-definition (HD) headlamp systems with dynamically adjustable light intensity distribution was presented in [1]. This contribution extends the original paper by a sharper resolution of luminous intensity, a broader validation and a more detailed discussion of the implementation.

The introduction of glare-free high beam in 2010 gave a major boost to headlamp development and highlighted the benefits of situation-adaptive lighting functions [2]. It allows driving with permanent high beam, with the headlamp control unit masking the areas classified as glare-critical by an environment camera. Figure 1 shows the schematic function using the example of glare protection for oncoming traffic. Such lighting functions considerably increase driving safety and comfort at night. While this new functionality was initially achieved by mechanically swivelling headlamp components, the trend towards digitalization has made its way more and more into the automotive industry in recent years. Even though the idea of a pixel light was first presented at the PAL (today ISAL) conference in 2001, it took many years for it to be implemented due to technical challenges [3]. Since 2013, mechatronic headlamp systems have increasingly been replaced by light-emitting diodes (LED) matrix solutions [4]. These work without mechanically moved components and realize the dynamics of their light distribution through a considerably higher number of independently controllable light sources with sharply separated solid angle areas of their light cones. The light distributions of the individual light sources add up to the total light distribution weighted by their respective continuously selectable electrical currents. The resulting total light distribution can thus be freely selected via the current values within the limits of the available resolution.

Figure 1. Glare-free highbeam light function cutting off a sharp spatial angle to avoid glaring oncoming traffic. © HELLA

After the matrix headlamps initially equipped with approx. 30 LED segments, a system with 84 LEDs was introduced in 2016 [5]. The increased resolution of the HD84 system allows even more extensive and precise adaptive lighting functions to be realized. The headlamp systems of future vehicles will be equipped with even higher resolutions. A specific example of this is the Liquid Crystal Display (LCD) technology. Only a few LEDs are still used as backlight. The light distribution of these LEDs is then primarily adjusted by a downstream LCD, which filters the light emitted by the LEDs with high local resolution [6]. With this technology, resolutions of several tens of thousand up to a million pixels can be achieved.

Compared to other vehicle components, the development of headlamp systems is characterized by the multidimensional solution space in terms of possible light distributions and the highly subjective factor in their evaluation. These properties require a strongly test-driven development. Because practical testing requires a dark environment, Original Equipment Manufacturers (OEMs) take a huge amount of time and effort to build test tracks that are shielded from ambient light. The light channel of HELLA in Lippstadt, shown in Figure 2, comprises a straight, 140 m long, two-lane road with road markings, which is located in a hall.

Due to the high variability of the light distributions and the situational adaptability of nowadays headlamp systems, static tests such as can be carried out in the light channel are by far not sufficient to cover the available functional spectrum of today's systems. On the other hand, the development and testing of dynamic lighting functions by night driving is no alternative due to safety, time and cost aspects as well as the limited or non-existent reproducibility of environmental influences.

Motivated by this, the idea of a simulation-based virtual night drive was born. In addition to the simulator interface, this requires in a first step the physically precise simulation of

Figure 2. HELLA light channel in Lippstadt. © HELLA

the ego vehicle (vehicle, which defines the driver's view) and in particular the headlight emitted by it, the other road users, the scenery, which includes the road, roadside elements, urban buildings, etc., and the weather conditions in real-time. Based on such a solution, very fast virtual headlamp tests are possible in freely selectable scenarios with complete reproducibility. Even if virtual night drives cannot completely replace the real ones, they allow a considerable reduction of test effort and thus lead not only to a time and cost saving, but also to a safety gain.

The first implementation of a real-time simulation of headlamp systems was presented by Lecocq et al. [7]. This initial realization implements per-vertex-lighting, which means that the lighting model is only evaluated at the vertices of scene objects. Subsequently, the pixel colors are determined by interpolation over all pixels on the basis of their surrounding vertices. The quality of the simulated light distributions therefore depends strongly on the tesselation of the scene objects. Berssenbrügge et al. present an approach based on per-fragment-lighting that decouples tesselation from the resolution of light distribution [8]. In their contribution, the lighting model is evaluated for each pixel of the output device, whereby inaccuracies by interpolations can be completely eliminated. A comparable concept is presented in [9]. It utilizes a proprietary development for the simulation of night driving with additional functionalities. For example, they provide a bird view including a distance grid to assess the light distribution with regard to range, width, homogenity and contrast sensitivity. Based on such simulations, various publications exist for testing dynamic lighting functions. In [10], Kemeny et al. present the simulation of the cornering light function within the established driving simulation software SCANeR. Like Berssenbrügge et al. in [11], they implement predictive cornering light, which calculates the ideal light distribution on the basis of navigation data and vehicle speed. As it turns out, this concept leads to better results than making the tilting of the light cone dependent only on the steering angle. Next to the cornering light function, Berssenbrügge et al. are also testing an advanced leveling light system in their simulation environment [12]. Here, as well, it can be seen that the development and testing of new lighting functions and the test-driven optimization of their design parameters can be considerably accelerated by virtual night driving simulation. With active safety light Knoll et al. present the simulative testing of a new light function for highlighting possible escape ways in risky driving situations [13]. This contribution represents a particularly interesting application case for simulation-based development, as the dangerous driving maneuvers for testing active safety light can only be performed under enormous safety precautions, or in some cases not at all.

All of the implementations mentioned above have in common that the light distributions of the light sources used are static. In concrete terms, this means that they are independent of time. Dynamic functions are mapped exclusively by means of switching light sources on or off and changing their orientation angles. HD headlamps, however, realize lighting functions in a totally different way. In this paper we present the first real-time simulation and its underlying modelling of HD headlamp systems. Their outstanding features and their abstract modelling are described in Section II. In Section III the shader-based implementation of the light simulation is discussed. The first Subsection III-A transforms a mathematical representation of a luminous intensity distribution (LID) into concrete structures of computer graphics. Based on this, the light rendering procedure is divided into the determination of the total light distribution as a weighted sum of the distributions of the individual light sources in the headlamp (Section III-B) and the illumination of the scene based on the overall distribution (Section III-C). In Section IV, the simulation results are validated with measurement data (Section IV-A), evaluated using a validated reference simulation tool (Section IV-B) and their performance on the utilized test hardware is examined (Section IV-C). The last sections provide a conclusion as well as an outlook for future work.

II. HD Headlamp Systems

HD systems are characterized by a great number of independent controllable light sources. Their illumination areas concentrate on sharply defined solid angle intervals with small overlapping areas. The total light distribution of such a headlamp results from the superposition of all the individual light distributions. This architecture enables the generation of highly dynamic overall light distributions by independently controlling the electrical current of each individual light source

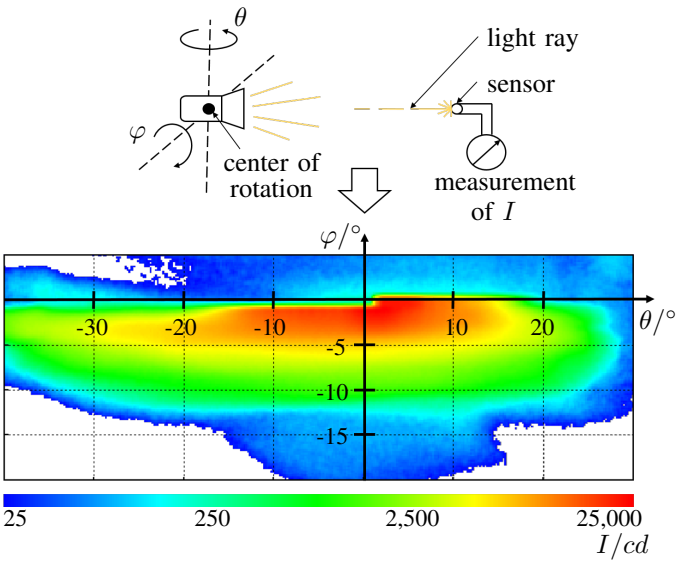


Figure 3. Low-beam distribution of luminous intensity I of left HD84-Matrix-LED headlamp measured in Candela [cd].

in the headlamp. The variety of the representable light distributions is limited only by the resolution of the headlamp, which can range from approx. 100 to several 10,000 pixels, and the permissible values of the electrical current depending on the light technology used [14].

For testing the simulation technique presented in this contribution, the HD84-Matrix-LED headlamp developed by HELLA is used. The actual HD component of this headlamp is realized by a matrix of 84 LEDs, which can be supplied individually with continuously adjustable electrical current. The individual light sources are mounted in three lines, with the lowest line (line 1) illuminating the close range in front of the vehicle, the middle line (line 2) focusing on the effective range of the low beam and the upper line (line 3) dedicated to functions in the high beam range (see upper area of Figure 4). To compensate for the relatively low resolution of 84 pixels, the illumination range of the HD module is limited to the solid angle range with the greatest variability requirements. Accordingly, additional light sources are provided to illuminate the vehicle front area, the sides and to support the high beam. Although the additional light sources are not visualized in Figure 4 for reasons of clarity, they are processed in the same way as the 84 LEDs of the matrix. In total, the HD84-Matrix-LED headlamp therefore contains 95 light sources.

To describe the characteristics of a light source, in lighting technology so-called luminous intensity distributions (see lower area of Figure 3) are used, which describe the luminous intensity depending on the direction of radiation [15]. The luminous intensity I measured in Candela $[cd]$ describes the radiation characteristic of a light source, by resolving the radiant power or luminous flux Φ (unit: Lumen $[lm]$) emitted by it in relation to the through-radiated solid angle Ω (unit: Steradiant $[sr]$) according to

$$I = \frac{\Phi}{\Omega} \quad [cd] = \left[\frac{lm}{sr}\right] \tag{1}$$

for a homogeneous luminous intensity within the considered solid angle. Using spherical coordinates the solid angle Ω is related to the polar angle φ and the azimuth angle θ by

$$\Omega = \int_{\theta_1}^{\theta_2} \int_{\varphi_1}^{\varphi_2} sin(\varphi) d\varphi d\theta, \tag{2}$$

whereby $[\varphi_1, \varphi_2]$ is the considered angle interval around the horizontal axis and $[\theta_1, \theta_2]$ corresponds to the interval around the vertical axis, which forms a square cutout on a spherical surface. [16]

In context of headlamps, the luminous intensity varies greatly with the radiation direction. By shrinking the solid angle to an infinitesimal area, the dependence of the luminous intensity I on the already mentioned angles φ and θ can be found by

$$I(\varphi, \theta) = \frac{d\Phi(\varphi, \theta)}{d\Omega} = \frac{d\Phi(\varphi, \theta)}{sin(\varphi) d\varphi d\theta}. \tag{3}$$

In the lower half of Figure 3, the low-beam distribution of the left HD84 headlamp is shown. The direction of radiation is defined by the polar angle φ and the azimuth angle θ. Their zero-position equals the direction of travel with regard to their mounting position in the vehicle. Even though Figure 3 does not show the entire angle range for better recognition, data for $\varphi \in [-29.9°, +19.9°]$ and $\theta \in [-89.9°, +89.9°]$ is available for all light distributions used in the simulation. The values of luminous intensity are color-coded with a logarithmic scale, as is usual for those diagrams. They vary over five orders of magnitude up to $25.000cd$. For high-beam distributions, the range is even greater. Therefore, this is also referred as high dynamic range information.

These luminous intensity distributions can be obtained for an existing headlamp by a goniophotometer-measurement [17]. In this case, the headlamp is mounted in a goniometer construction, by which it can be swivelled around its vertical and horizontal axis (see upper left area of Figure 3). The luminous intensity is usually measured by a fixed sensor at a distance of 25m (see upper right area of Figure 3). This distance serves in particular to maintain the photometric limit distance, under which the light source to be measured can no longer be approximated as a point source. The rotation angle around the horizontal and vertical axes of the goniometer construction correspond to the polar and azimuth angles φ and θ of the luminous intensity distribution.

Alternatively, the light distribution can be determined on the basis of computer models of the headlamp using ray tracing methods [18] and elaborate offline simulations, which were introduced at first by Neumann and Hogrefe in [19]. The quality of the resulting light distribution depends primarily on the model quality. Even though the characteristics of a real headlamp can never be exactly reproduced, this variant is particularly interesting in early development phases before the construction of a prototype.

In order to simulate the light distribution of a HD headlamp in any situation, the characteristics of the emitted light must be known for each individual light source in it. Therefore, the luminous intensity distribution is measured for each light source, in concrete terms 95 times, of the headlamp and especially for each LED of the HD84-Matrix. In the middle area of Figure 4, it can be seen exemplary intensity distributions of LED 1 (line 1, left), LED 45 (line 2, middle) and LED 84 (line 3, right). From the illustrations it becomes clear that each LED emits exclusively in a certain solid angle. As the lower line is responsible for the close range, the light intensity center is also found at negative polar angles. This regularity can also be applied to the other lines. Similarly, it becomes visible that the horizontal arrangement of the LEDs within the matrix corresponds to the horizontal arrangement of their light intensity centers. In addition, it is noticeable that the light distributions of the lateral light sources extend over larger areas than it is the case for light sources in the middle. Decisive for this are the especially high variability requirements placed on the central area in front of the vehicle. This area must therefore be resolved at a particularly high resolution in relation to the overall light distribution.

The individual light distributions are measured with running the corresponding LEDs at full power. During normal operation, the LEDs can be dimmed independently of each other in the range of 0-100% by specifying their electrical currents. In the lower left area of Figure 4 the low-beam distribution of the left HD84-Matrix headlamp is shown. The outline of the light distribution is also referred to as the light-dark boundary in the context of automotive headlamps. A characteristic feature of a low-beam light-dark boundary is the step in the range of $\theta = 0°$ in the light distribution. For negative values of θ, the luminous intensity in the direction of

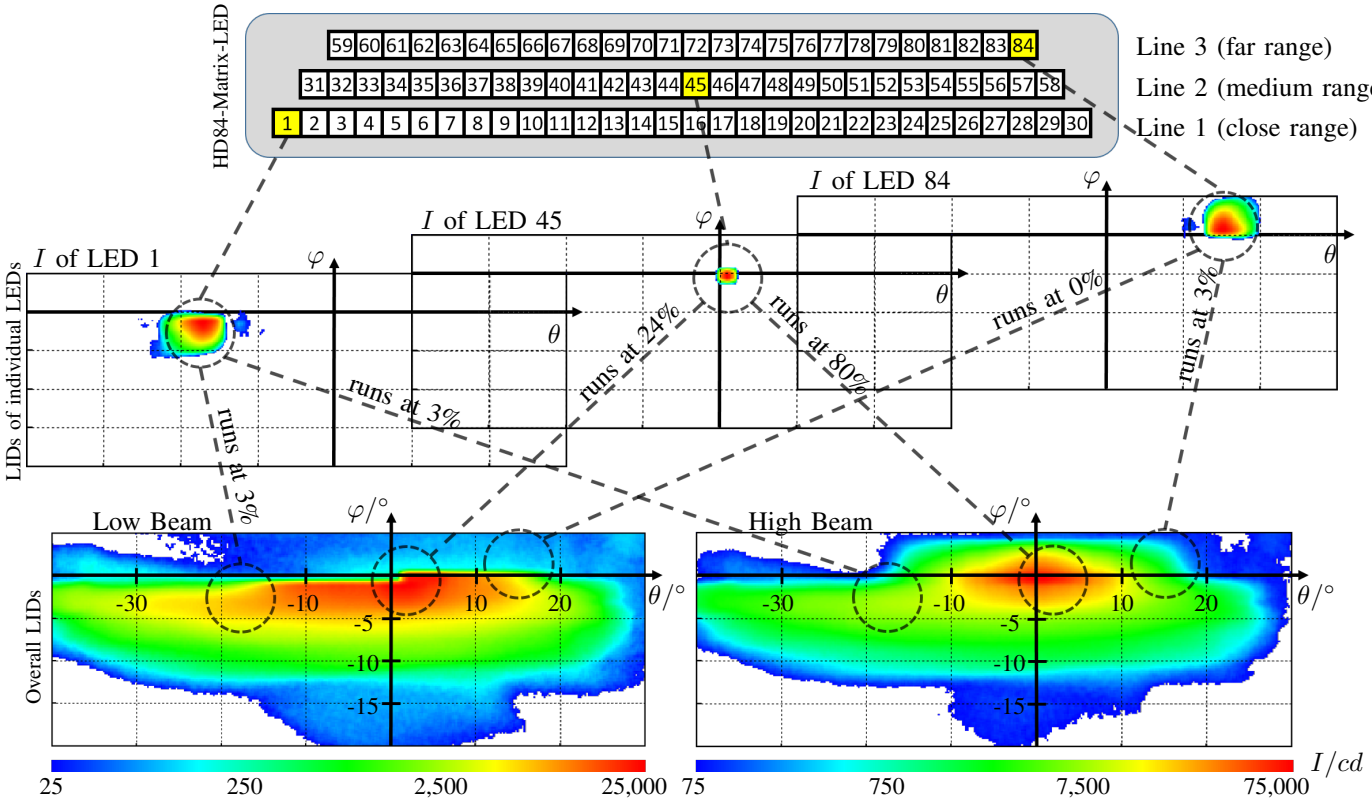


Figure 4. Luminous intensity distributions L_k of individual light sources (in this example for $k = 1, 45, 84$) and their electrical current weighted compositions to overall distributions, in example low-beam distribution (left) and high-beam distribution (right).

positive φ is cut off earlier than is the case for positive θ. In other words, the lane of oncoming traffic is illuminated less than the lane of one's own. This ensures that oncoming traffic is not glared, but simultaneously there is a good view of one's own lane.

In contrast to conventional headlamps, in which the geometry of the reflector shade makes it possible to achieve a low-beam light distribution using a single light source, the overall light distribution of an HD headlamp is obtained by adding all the light distributions of its individual light sources. In order to achieve the desired light-dark boundary of low beam, the headlamp control unit supplies all light sources with suitable dimming values. In the concrete example, all the LEDs in line 3, which is intended for the far range, are completely switched off. Therefore, the light distribution of LED 84 does not influence the overall light distribution of the low beam (see dotted lines between the individual and overall light distributions in Figure 4). In line 2, LEDs are primarily energized in the right half. LED 45, which is located directly at the step of the light-dark boundary, is operated at 24% of its maximum light output, for example. In line 1, all LEDs are energized, whereby the LEDs in the edge areas are dimmed more strongly than those in the middle area. This explains the low value of 3% of LED 1. When driving around bends, the center of the light distribution could be swivelled in the direction of the curve by adjusting the electrical current values to better illuminate the road.

The advantages of generating a light distribution from many small blocks can be easily seen. While the characteristic of the light distribution in conventional headlamps is fixed by the geometry of the reflector, it can be varied over a wide range in HD headlamps. Two factors limit the variety. On the one hand, this is the number of pixels and, on the other hand, the step size of the dimming value discretization, which plays a minor role. In the example of the HD84 module, the dimming values are coded by 6 bits. This results in a step size of about 1.6% relative to the respective maximum light output. The theoretical total number of light distributions is unimaginably high ($(2^6)^{84} \cdot 2^{16}$, 84 pixels in the HD84 module with 2^6 possible dimming values and 16 non-dimmable light sources).

The lower right part of Figure 4 shows the high beam distribution of the headlamp. The step in the light-dark boundary disappears and is replaced by a high beam cone which shifts the center of light intensity to positive φ. In addition, the light distribution is roughly symmetrical, since unlike in the case of low beam, no glare suppression of the opposite lane is desired. In addition, the headlamp shines brighter by a factor of 3, as can be seen from the false color scales at the bottom of Figure 4. Just as in the case of the low beam, the high beam distribution shown is generated exclusively by adjusting the dimming values. Line 3 is now active and the left half of line 2 is also energized. In addition, the light output of all LEDs is increased in order to achieve the overall higher light intensity. For example, the power of LED 45 is raised from 24% at low beam to 80% at high beam mode.

III. Implementation of HD Headlamp Simulation

After discussing the essential properties and functional principles of high-resolution headlamp systems, this section will focus on the real-time rendering of the headlamp light in the virtual scene. The implementation presented here sets the basis for virtual night drives with high-resolution systems and thus creates the possibility to develop and test these

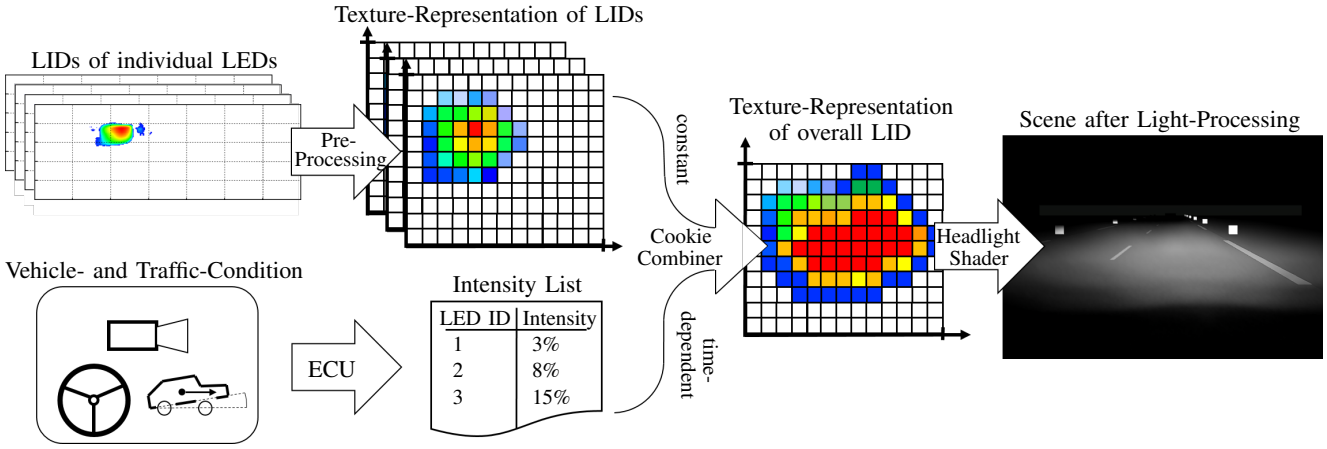


Figure 5. Functional flowchart of HD headlamp rendering.

simulation-based. For implementing the visualization of the night drive, we used the development environment Unity3D (Version 2017.3.1f1 [20]).

Figure 5 shows the logical flow of the presented implementation. The first step is to convert the previously measured luminous intensity distributions of the individual headlamp light sources into a data format that can be efficiently managed under the real-time requirements of driving simulation. Since this information does not vary over time, it is sufficient to perform this conversion as part of preprocessing. In concrete terms, each luminous intensity distribution is converted into a texture, as illustrated in the upper left area of Figure 5. This procedure only has to be carried out once for each headlamp type and is then available for any simulations. Details on this process will follow in Section III-A.

In addition to the individual light distributions, the relative current values of all light sources are necessary to determine the overall light distribution. The entirety of their values at a given point in time is referred to below as the intensity list. These are specified in the real vehicle by the headlamp control unit, depending on the traffic situation detected by the sensors (see lower left area of Figure 5). This updates the current values with a frequency of 50Hz. Within the framework of the implementation, the behaviour of the control unit is not simulated, since it will later be integrated into a hardware-in-the-loop simulation. In this respect, the current values can be assumed to be given, whereby it must be ensured that the overall light distribution can be determined within a reasonable time due to the high update frequency.

Once both, the individual light distributions and the temporary intensity list, are available, the total light distribution can be calculated. This can be achieved by adding the individual light distributions weighted by the relative current values. Due to the time dependence of the current values, the total light distribution also varies with time, so that it must be recalculated with the update rate of the intensity list. To be able to perform this calculation 50 times per second, it is carried out in a highly parallel manner using a shader on the GPU. All shaders presented here are implemented in Cg (C for graphics) [21]. This shader is called Cookie Combiner and is described in more detail in Section III-B. The output type of the Cookie Combiner corresponds to the types of the individual light distributions. It is also a texture that encodes the luminous intensity values with respect to the spatial angles.

Finally, all information is available to render the virtual scene. Here, deferred shading was used [22]. While the light calculations in the more established forward rendering are performed individually for each object of the scene, in deferred rendering all objects are first rendered in the base pass under exclusion of light influences in the so-called G-Buffer (Geometry Buffer). In the subsequent lighting pass, the light calculations are applied to the G-Buffer only once per light source, which generally requires considerably fewer shader executions than it is the case in forward rendering. Besides possible standard light sources of the scene, the lighting pass also activates the Headlight-Shader intended for displaying headlamp light (see last step of Figure 5). The execution of the Headlight-Shader is downstream of the standard shader for deferred shading and is integrated into the Programmable Rendering Pipeline by a Command Buffer [23]. This uses the total light distribution determined by the Cookie Combiner and, using a lighting model, determines the color pixel by pixel, which results from the object and light properties as well as the geometric relationships.

A. *Digital Representation of LIDs*

In order to be able to precisely reference the relevant elements of the light distributions of HD systems in the following sections, first of all these are described more formally. A discretized light distribution can be interpreted as a matrix whose dimensions depend on the considered angular range $[\varphi_l, \varphi_u]$ or $[\theta_l, \theta_u]$ and the corresponding resolution $\Delta\varphi$ or $\Delta\theta$ in horizontal or vertical direction, whereby $M = \frac{\varphi_u - \varphi_l}{\Delta\varphi}, N = \frac{\theta_u - \theta_l}{\Delta\theta}$ with $M, N \in \mathbb{N}$ must apply. The discrete value φ_m with $0 \leq m \leq M$ or θ_n with $0 \leq n \leq N$ then refers to the horizontal angle $\varphi_l + m \cdot \Delta\varphi$ or the vertical angle $\theta_l + n \cdot \Delta\theta$. The light distribution L_k of the light source k with $k \in \{1, K\}$ of an HD headlamp with a total of K individual light sources then has the form $L_k \in \mathbb{R}_{\geq 0}^{M \times N}$. In the concrete case, the values $\varphi_l = -29.9°, \varphi_u = +19.9°, \theta_l = -89.9°, \theta_u = +89.9°$ apply to the measured angular ranges. Resolving both angles with $\Delta\varphi = \Delta\theta = 0.2°$ results in matrices of $M = 250$ rows and $N = 900$ columns. In addition, the headlamp examined has a total of $K = 95$ light sources.

The entry $l_k(m, n)$ of row m and column n of the L_k matrix now contains the luminous intensity of the light source

k at full power in the discretized direction of the vertical angle φ_m and the horizontal angle θ_n in Candela. For example, the entry $l_{17}(100, 400)$ contains the luminous intensity of the light source with index 17 in the direction of the vertical angle $\varphi_{100} = \varphi_l + 100 \cdot \Delta\varphi = -29.9° + 100 \cdot 0.2° = -9.9°$ and the horizontal angle $\theta_{400} = \theta_l + 400 \cdot \Delta\theta = -89.9° + 400 \cdot 0.2° = -9.9°$.

After defining the light distributions of individual light sources, the aggregation of these to the total light distribution L can be formulated. For this purpose, a system is defined whose input variables represent the relative electrical current values $i_k \in [0, 1]$ of the individual light sources normalized to their maximum value. In contrast to the matrices $L_1, \ldots, L_K$, these values are time-dependent signals coming from the headlight control unit at 50 Hz. The output of the system is the resulting light distribution of the headlamp and thus a weighted composition of all individual light distributions. Formally, the output variable corresponds to a L matrix whose dimensions are identical to the dimensions of the individual distributions $L_1, \ldots, L_K$. The current overall light distribution can be formulated as

$$L(t) = \sum_{k=1}^{K} i_k(t) \cdot L_k \text{ with } L(t) \in \mathbb{R}_{\geq 0}^{M \times N}. \quad (4)$$

The time-dependent matrix L contains all information describing the light emitted by the headlamp at time t, which constitutes the basic information for simulation purposes.

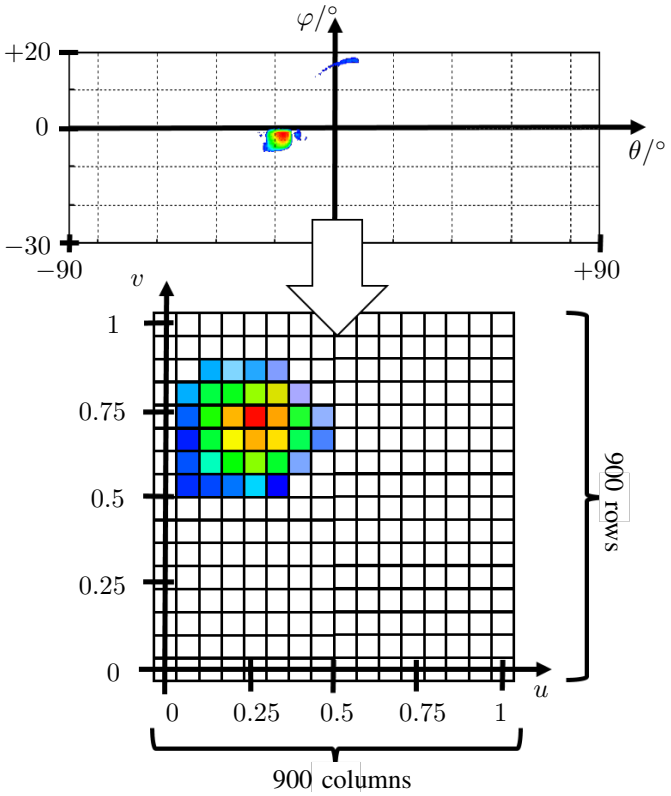


Figure 6. Digital representation of a luminous intensity distribution.

To represent the formally introduced matrices in the computer graphics field, textures are used. A texture is a data format established in computer graphics which was originally intended for the realistic coloring of scene objects. In this role it can be understood as an image which is placed like a sticker on the surface of a three-dimensional scene object. The individual entries of the texture encode the color at the respective place then typically in the 3 RGB channels (red, green, blue) and a 4th transparency channel (α channel). In this case, textures are used to encode light intensity distributions.

Now some technical details of the conversion of luminous intensity distributions to textures shall be described, which are relevant for the later explanation of Cookie Combiner- and Headlight-Shader. To minimize interpolation inaccuracies, we use textures with the same row and column numbers as we have for the discretized luminous intensity distributions (250 rows, 900 columns). Since the textures have to be square for technical reasons, it is necessary to increase the number of rows from 250 to 900. For this purpose, the vertical angle range is artificially increased from [-29.9, +19.9] to [-89.9, +89.9] and the luminous intensity values at the non-measured angles are set to 0. As a consequence, the texture contains 900x900 data points, which are also called texels. Using the same resolution as the measured luminous intensity distributions, each texel corresponds directly to a measuring point of the distribution. In classic color textures the values of the R, G, B and A channels of a texel are encoded with 8 bit fixed-point, whereby each texel contains 32 bit information (RGBA32 texture). A luminous intensity distribution, in contrast, contains only one dimensional information - the luminous intensity. Here the use of a RGBA32 texture would be inefficient, since a precision of 8 bit is on the one hand too low for the high dynamics of a luminous intensity distribution and on the other hand the remaining 3 channels remain unused. Instead, we use an RFloat texture format. This supports only one channel (R-Channel) and uses for this the entire 32 bit available to the corresponding texel. In addition, the value is coded in floating point format, which also benefits the high dynamics of the luminous intensity values.

Another important point is the indexing of textures. Figure 6 can be used for understanding. Texture coordinates or uv coordinates are always used normalized. In other words, a texel is not addressed by its absolute row and column number in the texture, but by the values divided by their total numbers. $u \in [0, 1]$ represents the horizontal axis (corresponds to the column number) and $v \in [0, 1]$ the vertical axis (corresponds to the row number). The texel at address $(u, v) = (0, 0)$ thus contains the luminous intensity in the direction of the vertical angle $\varphi = -89.9°$ and $\theta = -89.9°$ (respectively texel at $(u, v) = (1, 1)$ corresponds to the luminous intensity value at $\varphi = +89.9°$ and $\theta = +89.9°$).

Since u and v can be specified from a continuous value range, u, v pairs can also address places of the texture that do not match exactly to a texel but lie in the intermediate areas. Such accesses are possible and are answered with the bilinear interpolation between the surrounding texel values, which is visualized in Figure 7. In general, the return r of a texture access at the coordinates (u, v) can be traced back to the luminous intensity distribution as follows:

If the uv coordinates correspond exactly to a texel of the texture, its value can be returned unchanged. The condition for this is that $u \cdot M$ and $v \cdot N$ are elements of the natural numbers. In the other case, the neighboring texels of the access coordinates (u, v) must be found first, which are designated by t_{ll}, t_{lr}, t_{ul} and t_{ur} in Figure 7. As also noted in Figure 7, these

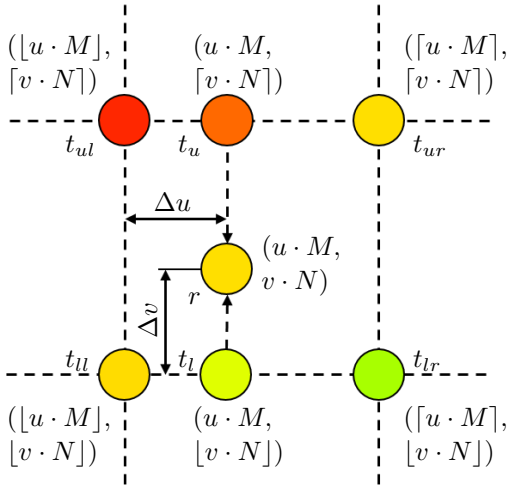


Figure 7. Bilinear interpolation on texture access at coordinate (u, v).

can be addressed by simply rounding $u \cdot M$ and $v \cdot N$ up and down. Afterwards, one interpolation is performed between the upper neighbor texels and one interpolation between the lower neighbor texels along the u-axis, yielding the values t_l and t_u. Formally these result to

$$\begin{aligned} t_l &= lerp(t_{ll}, t_{lr}, \Delta u) \text{ and} \\ t_u &= lerp(t_{ul}, t_{ur}, \Delta u) \\ &\text{with } \Delta u = u \cdot M - \lfloor u \cdot M \rfloor \\ &lerp(v_1, v_2, c) = v_1 \cdot (1 - c) + v_2 \cdot c, \\ &v_1, v_2 \in \mathbb{R}, c \in [0, 1] \end{aligned} \tag{5}$$

The return value r is now obtained from the bilinear interpolation by interpolating the interpolations along the u coordinate again along the v coordinate:

$$\begin{aligned} r &= lerp(t_l, t_u, \Delta v) \\ &\text{with } \Delta v = v \cdot N - \lfloor v \cdot N \rfloor \end{aligned} \tag{6}$$

Finally, in order to clarify the relationship between textures and the original luminous intensity distributions, the way back from a texture T to luminous intensity l in the direction of vertical angle $\varphi \in [\varphi_l, \varphi_u]$ and horizontal angle $\theta \in [\theta_l, \theta_u]$ is shown:

$$\begin{aligned} l(\varphi, \theta) &= T(u, v) \text{ with} \\ u &= \frac{\varphi - \varphi_l}{\varphi_u - \varphi_l}, \varphi \in [\varphi_l, \varphi_u] \\ v &= \frac{\theta - \theta_l}{\theta_u - \theta_l}, \theta \in [\theta_l, \theta_u] \end{aligned} \tag{7}$$

whereby $T(\cdot, \cdot)$ means a texture lookup at T at the u and v coordinates given in this order as operands.

B. Cookie Combiner-Shader

Once a way has been found to digitally map luminous intensity distributions to a graphics card compatible manner, the next step is to determine the total light distribution from the individual ones. The result can then be used to adjust the light intensity in the lighting pass, more precisely in the Headlight-Shader executed in it (see Section III-C), depending on the direction. Vividly, this realization is comparable to a dynamic transparency film, by which a homogeneous light source is filtered to produce the desired radiation characteristic.

Texturing of light sources to vary the luminous intensity in different beam directions is already established. Such light textures are called cookies, explaining the name of the Cookie Combiner-Shader. Its task is to combine the textures of the individual light distributions into a total light distribution texture according to Equation (4). Figure 9 illustrates the data flow of the combining procedure on a mathematical level.

The implementation of this calculation as a shader enables the highly parallel execution on the graphics card, which is necessary to fulfill the real-time requirements. Within the scope of this work, only vertex and fragment shaders are used. Vertex shaders process vertices and the information associated with them of the three-dimensional geometries in the scene. Afterwards the scene is rasterized object by object and transformed into a two-dimensional image. Then fragment shaders work on the fragments of this image and determine the pixel colors. Comprehensive information about shaders can be looked up in [23].

In the case of the Cookie Combiner, the vertex shader is trivial, since only two-dimensional data (luminous intensity distributions) is processed and no vertex operations are performed. Its whole logic is placed in the fragment program. The operations are visualized by the following pseudo code on the one hand and by Figure 8 in a graphical way on the other hand. While in the rendering pipeline a fragment program writes into the screen output and the pixels of the screen slip into the role of the fragments, we define a texture, which is reused in the rendering pipeline in a later step, as the render target for our application. More precisely, the render target of the Cookie Combiner-Shader is a texture T_{comb} with the same type and dimensions as the textures of the individual light distributions - a scalar floating point texture with 900x900 texels.

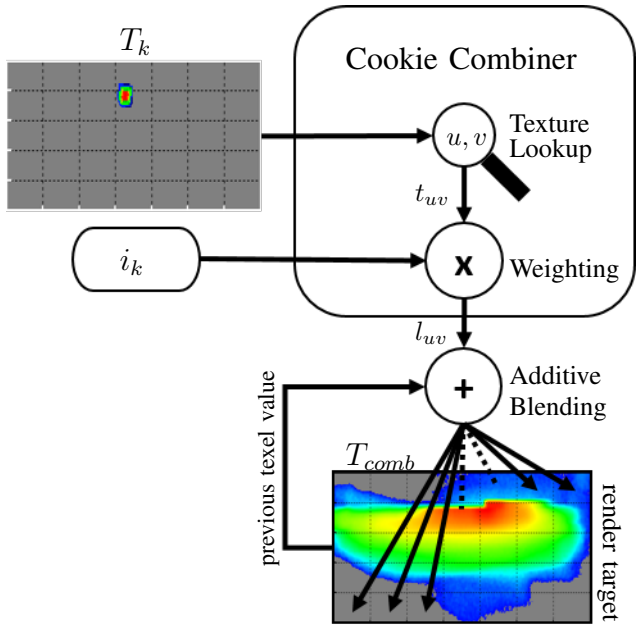


Figure 8. Logic diagram of the technical Cookie Combining Procedure.

Require: $u, v \in [0, 1]$ normalized coordinates of current render target texel, T_k scalar floating point texture corresponding to luminous intensity distribution L_k and relative current value $i_k \in [0, 1]$ of individual light source k

1: $t_{uv} = T_k(p_{target}.u, p_{target}.v)$ // texture lookup

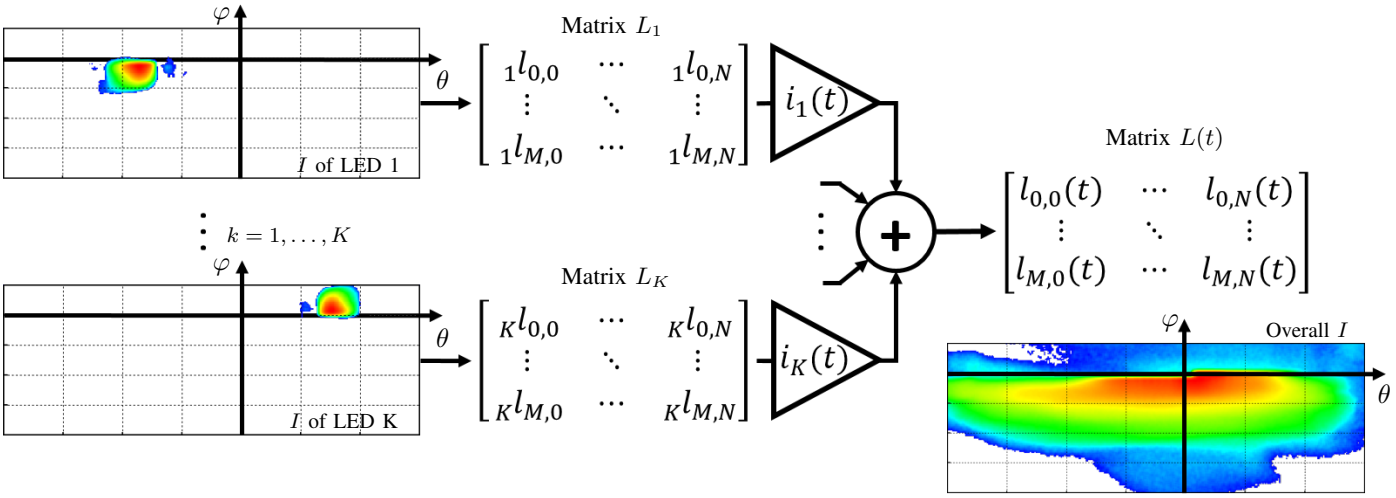


Figure 9. Data flow of Cookie Combining during a single frame.

2: $l_{uv} \leftarrow i_k \cdot t_{uv}$ // weighting
3: **return** l_{uv} // additive blending into render target

The fragment program of the Cookie Combiner-Shader expects two parameters (see upper left area of Figure 8). One is the texture T_k, which represents the light intensity distribution L_k of the individual light source k. The other parameter represents the relative intensity i_k with which this light source is currently operated. Its code is executed for each texel of the render target T_{comb} in parallel (see lower area of Figure 8), whereby the individual threads can be distinguished by further inherent parameters u and v constituting the normalized coordinates of the thread specific texel of the render target.

In the first step within processing, the shader reads the light intensity value t_{uv} of the given individual light texture T_k at position (u, v) corresponding to the texel to be written to in the render target (line 1). This simple method of addressing is made possible because both the individual light distributions and the target texture encode the light intensity distribution over the same angular ranges, even if their resolutions were different. Considering the intensity with which the light source k is currently operated, the multiplication of the maximum luminous intensity t_{uv} with the relative current value i_k takes place at line 2. Even if the relationship between the electric current and the luminous intensity is not linear, it can be assumed to be linear here by compensating the non-linearity with the real control unit. The product l_{uv} as actual luminous intensity in the direction represented by the u and v coordinates of the current texel according to 7 represents the output of the shader (line 3).

After completion of the shader operations for all texels, the target texture contains the contribution of the light source k within the total light distribution, whereby a single component of the sum in the data flow visualized by Figure 9 is mapped. In order to obtain the complete light distribution of the headlamp, all texels of the render target T_{comb} are first initialized with 0. Then the Cookie Combiner is applied to the target texture repeatedly with iterating through all individual light sources by changing the parameters T_k and i_k. The rendering is done with additive blending (see Figures 8 and 9), so the previous values in the render target are not overwritten by the returns of the shader but added to it [23].

After applying the Cookie Combiner to all individual light sources $k = 1, \ldots, K$, the render target T_{comb} contains the total luminous intensity distribution, in which the texels can still be interpreted as in Section III-A. Due to its dependence on the time-varying relative currents $i_1, \ldots, i_K$, this texture is only valid for the current time step. The K times execution of the Cookie Combiner must therefore take place with the update rate of the headlight control unit for each simulated headlamp, which is 50 Hz. The determined total light distribution represents a central parameter of the Headlight-Shader presented below.

C. Headlight-Shader

The implementation of the Headlight-Shader is much more extensive. Berssenbrügge et al. solved a similar problem for static light distributions in [8] by using a built-in spotlight and mapping the light distribution to its cookie scheme. In this contribution, the lighting is done by a custom shader using the deferred shading pipeline, whose implementation is oriented to Unity's built-in lighting shader for deferred rendering. Moreover, High Dynamic Range (HDR) Rendering is used. This technique allows a more detailed resolution of the color information by using higher memory resources, whereby especially with high-contrast images their details are preserved in the best possible way and come closer to the perception of the human eye [24]. Even if the color information has to be reduced to 255 brightness levels (8 bit per color channel) for output on a monitor, the colors can be dynamically scaled instead of just setting too bright areas to white, as is the case with Standard Dynamic Range (SDR). Especially in the context of headlamp simulation, the use of HDR colors leads to visibly higher quality images.

Figure 10 illustrates a very simplified pipeline run in deferred rendering. The concept of deferred shading is characterized by the strict division of rendering into a first step of projecting the 3-dimensional scene information to a 2-dimensional image (G(eometry)-Buffer) and the subsequent lighting on the basis of this G-Buffer. The main advantage here is that lighting only has to be applied to all pixels of the 2-dimensional G-Buffer instead of every object in the scene, as is the case with conventional forward rendering. Thus, the computational complexity of a scene with m objects and n lights can be reduced from $\mathcal{O}(m \cdot n)$ to $\mathcal{O}(m + n)$.

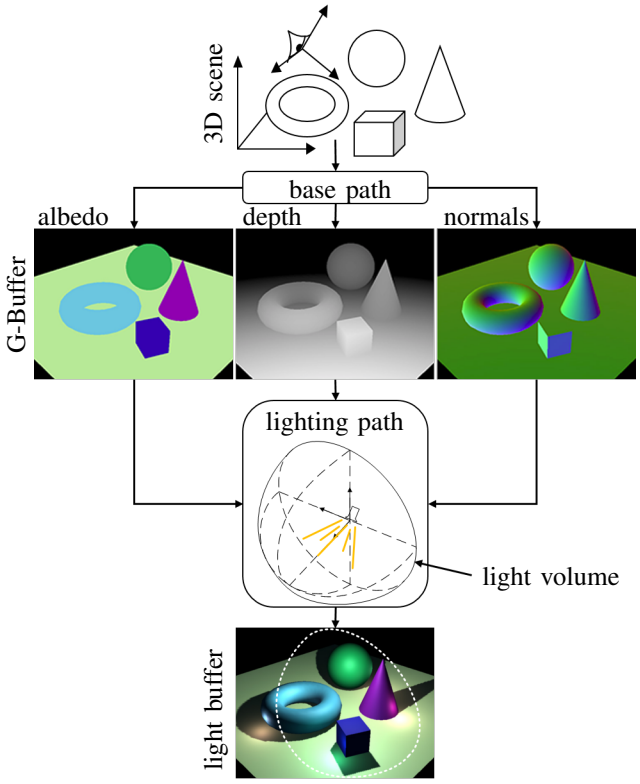


Figure 10. Deferred rendering pipeline.

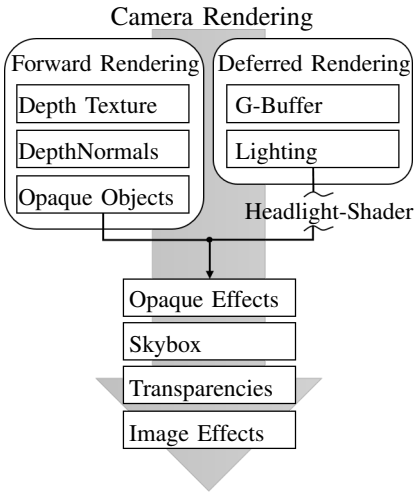


Figure 11. Injection of Graphics Command Buffer in Unity's rendering pipeline (created in accordance with [25]).

At first, the 3-dimensional scene objects are rendered in the base path into the G-Buffer, whereby the scene information is reduced to two dimensions. Since the lighting calculations are done in a deferred step, it is not possible to render directly the final pixel colors. Therefore, all relevant information for later lighting calculations has to be contained in the G-Buffer. In detail, this could be color information (often distinguished in colors for ambient, diffuse and specular components, see left image of middle left area of Figure 10), depth values in terms of the distance between camera and object (encoded by grayscale at center area of Figure 10) or surface orientations defined by normal vectors (normal directions color-coded at middle left area of Figure 10).

Given the G-Buffer, the lighting pass can be initiated. This is the pipeline step, where the integration of Headlight-Shader next to Unity's standard shader takes place. Using Unity's Graphics Command Buffer [26], its integration is done subsequently to the 'Lighting'-Block in Figure 11, where standard lighting by the built-in shader takes place. A Command Buffer holds list of rendering commands, which can be injected between all boxes in Figure 11 to the conventional rendering pipeline. Integrating the Headlight-Shader this way instead of modifying the built-in shader, preserves compatibility to all standard light sources in the scene.

Within the lighting pass, there is one shader called per light source. Their common render target is the light buffer, visualized in the lower area of Figure 10. After finishing all light shaders, the light buffer contains the final scene image, ready for displaying on the output device. While the standard shader is called for Unity's built-in lights (directional, point or spot light), the Headlight-Shader is called for each headlamp. In both cases, for each light source there is a predefined light volume describing the volume, which is potentially affected by it, relative to the light position and orientation. For this purpose, we choose a half sphere, approximated by a mesh of 65 vertices (see Figure 13), in contrast to a pyramid, used by Unity's spotlight. With reference to the luminous intensity distributions measured at an angular range of 180° in horizontal and vertical directions, a half sphere is much more suitable for this application. The light source is located in the spherical center and is oriented vertically to the plane half-spherical surface in the direction of the curvature (see center area of Figure 10). While processed by the vertex program of the Headlight-Shader, the light volume is rotated and translated to get its correct position in the scene. This ensures that the subsequent rasterization of the light volume can be understood as a marking of the pixels of the G-Buffer that may have been influenced by the light, which are illustrated by the white dotted outline in the light buffer of Figure 10. The following fragment program of the Headlight-Shader operates on all of these marked fragments of the light buffer, calculating their colors by accessing relevant information of the illuminated fragment stored in the G-Buffer. To consider all light sources, whose light volumes cover the current light buffer pixel, the return values of all light shaders are written in the light buffer by additive blending. After passing through all shaders in the lighting pass, the light buffer contains the completely rendered scene image.

At this point, the vertex and fragment program of the Headlight-Shader will be explained in more detail along its pseudo code. For better understanding, Figure 12 shows a very reduced scene and illustrates the essential components and coordinate systems for rendering. The relative positions of all components are described in the world space w. A further important space is the view space v (also common: 'eye space'), describing the scene from camera's point of view. All points in the scene, visible for the camera, can be limited by a pyramid frustum with the tip in the center of v and the floor and top area perpendicular to the viewing direction (see blue thick dashed lines in Figure 12). This geometry is known in computer graphics as view frustum. Its top area with distance n to camera position is called 'near clipping plane', while its floor area with distance f to camera

is termed 'far clipping plane'. With the lighting pass, the light source becomes relevant for the first time during the rendering process. Figure 12 shows the headlamp to be simulated with its local light space l. As already mentioned, the illumination range of a headlamp is described by a light volume in the form of a half-sphere, which is illustrated by yellow lines in Figure 12. Furthermore, important for the following considerations are the local spaces of all objects within the scene. As an example, Figure 12 shows a cylindrical object with the local coordinate system o.

According to the schema of lighting pass in deferred rendering, the vertices of the light volume are first processed by the vertex program. In this sense, during the lighting pass, the light volume slips into the role of a scene object in the base pass. Implementing per-fragment-lighting, the vertex program can be realized by a few lines of code, since all light calculations are done in the subsequent fragment program. Essentially, three parameters are generated. First, a vertex ${}_lp'$ of the light volume mesh, passed in homogeneous coordinates [27], is transformed from the light space l into the clip space c of the camera (line 1). The transformation from l to the world space w is done by the matrix L, from w to the view space v by the matrix V and from v to the clip space c by the matrix P (see Figure 12). c has the same origin as v, but distorts the coordinates in perspective for easier mapping to the screen. The vertex coordinates in the clip space ${}_cp'$ form the basis of the rasterization and must be included in the return of the vertex shader. Furthermore, the clip space position in the lines 2 and 3 is transformed so that the x and y coordinates are normalized to $[0, 1]$ for vertices in the view frustum in accordance with the perspective division in the fragment program. These values are then used as texture coordinates to correctly address the G-Buffer and form the second output of the vertex shader. Finally, the vector from the camera to the vertex in view space ${}_vp'$ is calculated by multiplying the vertex ${}_lp'$ in l by L and V (line 4). This value is needed next to ${}_cp'_{uv}$ to reconstruct the 3-dimensional position of the fragment in w within the fragment shader and forms the third return of the vertex shader. This way all vertices of the half sphere are processed by the shader.

Require: ${}_lp' \in \mathbb{R}^4$ vertex of light volume as homogeneus coordinates in l

1: ${}_cp' \leftarrow P \cdot V \cdot L \cdot {}_lp'$ // vertex in c
2: ${}_cp'_{uv}.x \leftarrow \frac{1}{2} \cdot {}_cp'.x + \frac{1}{2} \cdot {}_cp'.w$ // transform to screen space
3: ${}_cp'_{uv}.y \leftarrow \frac{1}{2} \cdot {}_cp'.y + \frac{1}{2} \cdot {}_cp'.w$ // transform to screen space
4: ${}_vp' \leftarrow V \cdot L \cdot {}_lp'$ // vertex in v
5: **return** ${}_cp', {}_cp'_{uv}, {}_vp'$

Afterwards the rasterization is effected based on the clip space coordinates ${}_cp'$. The remaining vertex program returns are bilinear interpolated to the fragments according to their distances to the vertices. The fragment shader processes all fragments in the light buffer covered by the light volume with the respective interpolation results ${}_cp'_{uv}, {}_vp'$ as parameters.

Require: ${}_cp'_{uv} \in \mathbb{R}^4$ transformed coords in c and ${}_vp' \in \mathbb{R}^3$ coords in v of vertex p' of light volume

1: ${}_vp'' \leftarrow \frac{f}{{}_vp'.z} \cdot {}_vp'$ // scale to far clipping distance (f)
2: $b_{uv} \leftarrow \frac{1}{{}_cp'_{uv}.w} \cdot {}_cp'_{uv}.xy$ // buffer coords of p' (same for p)
3: $z_{uv} \leftarrow G_{depth}(b_{uv})$ // norm. depth at screen position b_{uv}
4: ${}_vp \leftarrow z_{uv} \cdot {}_vp''$ // position of p in v
5: ${}_wp \leftarrow V^{-1} \cdot {}_vp$ // position of p in w
6: ${}_wv_{c,o} \leftarrow {}_wp_c - {}_wp_o$ // vector p→camera in w
7: ${}_wn_{c,o} \leftarrow \frac{{}_wv_{c,o}}{|{}_wv_{c,o}|}$ // direction p→camera in w

8: ${}_wl \leftarrow L[1:4, 4]$ // position of light in w
9: ${}_wv_{l,o} \leftarrow {}_wl - {}_wp$ // vector p→light in w
10: ${}_wn_{l,o} \leftarrow \frac{{}_wv_{l,o}}{|{}_wv_{l,o}|}$ // direction p→light in w

11: ${}_lp \leftarrow L^{-1} \cdot {}_wp$ // vector light→p in l
12: $a_{deg}.x \leftarrow \text{atan2}({}_lp.x, {}_lp.z)$ // horiz. and vert. angle be-
13: $a_{deg}.y \leftarrow \text{atan2}({}_lp.y, {}_lp.z)$ // tween light axis and ${}_lp$
14: $t_{uv} \leftarrow \frac{a_{deg}+90°}{180°}$ // Light-Cookie coordinates
15: $l_{uv} \leftarrow T_{cookie}(t_{uv})$ // light power in specific direction

16: $att \leftarrow \frac{1}{l.range^2} \cdot {}_wv_{l,o} \cdot {}_wv_{l,o}$ // light attenuation
17: $att \leftarrow att \cdot l_{uv}$ // consider light power
18: $l.color \leftarrow att \cdot l.color$ // attenuated light color

19: **return** lightingModel$(c_o, {}_wn_o, {}_wn_{c,o}, {}_wn_{l,o}, l.color)$

The shader code can be divided into five logical blocks. The first block (line 1 to 7) reconstructs the three dimensional surface point p of the scene object o visible on the current fragment. This reconstruction is enabled by the additional information provided by the vertex shader. Outgoing from p', which triggers the current fragment program call as it is part of the light volume surface, it can be seen, p' is mapped to the same screen position as p (see dashed line through p, p' and p'' in Figure 12). p in turn represents a surface point of a physical scene object for which lighting has to be performed in the following. According to the figure, ${}_vp'$ describes the cameras view direction to p in v. The G-Buffer created in the base path of deferred rendering can be used to determine the exact position of p on the corresponding line. It encodes the z coordinate (or depth value) in v for each fragment in addition to other data. To read the correct value from the depth buffer, the buffer coordinates must be determined first. Therefore, the vector ${}_cp'_{uv}$ is defined in the vertex shader, whose x- and y-coordinates lie in the interval $[0, {}_cp'_v.w = {}_cp'_{uv}.w]$ for points within view frustum. After the perspective division by the homogeneous component in line 2 the coordinates b_{uv} are in the value range $[0, 1]$ (line 3) used for texture/buffer access. The depth buffer encodes depth z_{uv} normalized on distance f to the far clipping plane in the interval $[0, 1]$. The coords of p in v result from scaling of ${}_vp'$ to the far clipping plane receiving ${}_vp''$ (line 1) and the subsequent multiplication with the normalized depth z_{uv} (line 4). Multiplication by the inverse of V transfers the object point p from v to w. For the evaluation of the lighting model, the normalized direction vector from the object point to the camera (eye vector) in world space ${}_wn_{o,c}$ (see Figure 12) is needed (line 6 and 7).

In addition to the eye vector, the incidence of light on the object plays a central role in the lighting model, too. The light vector is defined in lines 8 to 10. First, the position of the light source in world space ${}_wl$ is extracted from the matrix L (line 8). Since the Headlight-Shader only renders the mesh of the light volume into the light buffer, the transformation matrix L from current object space l to w is constant across all calls of the vertex program and contains the translation, rotation and scaling of the light volume mesh into w. In general, the world space coordinates ${}_ws$ of a point s can be achieved by multiplying its light space coordinates ${}_ls$ by the homogeneous transformation matrix L from l to w according to

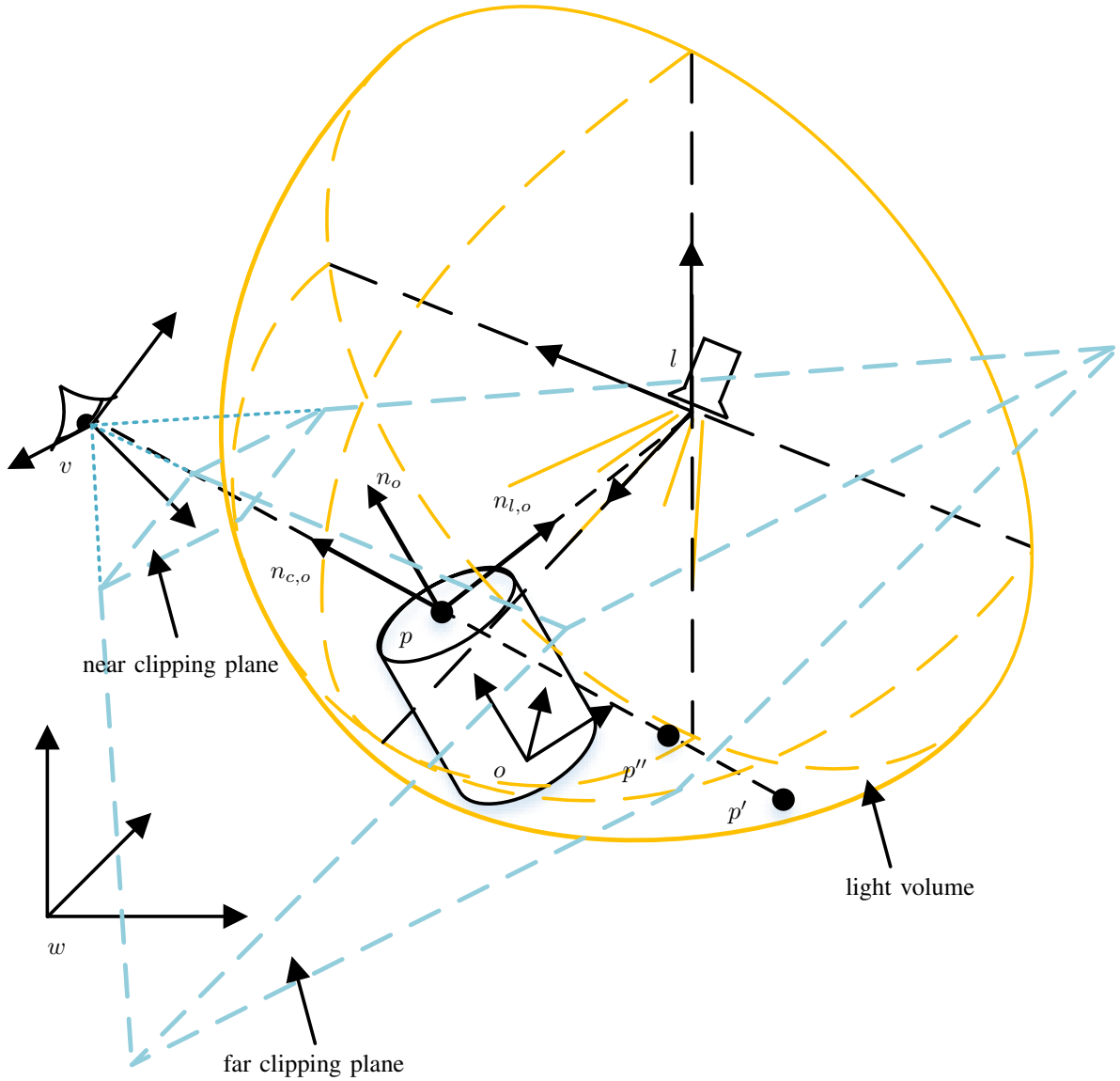


Figure 12. Simple scene to be rendered with coordinate systems for world space (w), view space (v), light space (l) and an exemplarily object with its local space (o).

$$ {}_{w}s = L \cdot {}_{l}s \text{ with } L = \begin{bmatrix} l_{11} & \dots & l_{14} \\ \vdots & & \vdots \\ l_{41} & \dots & l_{44} \end{bmatrix}, \qquad (8) $$

whereby l_{14}, l_{24} and l_{34} contain the translation from l to w along the x-, y- and z-directions [23]. Since the light source is in the coordinate origin of l, the translation in L corresponds to the position of the light in w and can be read from the fourth column of L. Now the normalized direction vector from the object point to the light source in world space ${}_{w}n_{l,o}$ (see Figure 12) can be determined by the lines 9 and 10 similar to the previous section.

In the lines 11 to 15, the luminous intensity distribution T_{comb} determined by the Cookie Combiner-Shader is taken into account. In order to determine the luminous intensity to be applied to the current fragment from the light distribution, the horizontal (θ) and vertical (φ) angles of the incident light beam with respect to the light center axis must be calculated. For this purpose, the object point belonging to the fragment is transferred to the light space l by multiplying with L^{-1}. The position of this point in l, in whose coordinate origin the light source is located and oriented along the z axis, simultaneously corresponds to the light beam ${}_{l}p$ from the source to the object. Its angle to the light center axis or z axis in l can then be determined with the $atan2$ function (line 12 and 13). The cookie T_{comb} must be addressed with normalized texture coordinates. Therefore, the angles moving in the range $[-90°, +90°]$ due to the used light volume are transformed by line 14 to texture coordinates t_{uv} according to Equation (7). Thereby, the applicable value can finally be read from the light distribution texture T_{comb} (line 15).

After the positions of the relevant elements and the luminous intensity of the light are already known, the distance to

the light source must be taken into account in the lines 16 to 18. The intensity of light decreases square with the distance to the shined object [16]. This square distance can be formulated most efficiently as a scalar product of the vector from object to light ${}_wv_{l,o}$ with itself. The distance between object and light is referred to a freely selectable positive light parameter $l.range$, which allows the user to adjust the range respectively the intensity of the light (line 16). It applies $att = 1$ if the distance between light source and object corresponds to the parameter $l.range$, and $att \to \infty$ for increasing distances. To take into account the directional luminous intensity l_{uv}, it is multiplied by att to the final light attenuation. The calculated attenuation can finally be mapped by the light color in line 18 through multiplying the color of the light defined by the user by att. Assuming a white light, this multiplication corresponds to a shift on the grayscale.

Finally, the lighting model can be evaluated. At this point, no separate solution has been implemented yet, but an Unity-internal local lighting model has been used. This receives the previously determined normals ${}_wn_{o,c}, {}_wn_{l,o}$, the surface normal ${}_wn_o$ (see Figure 12) and material data c_o of the object from the G-Buffer at buffer coordinates b_{uv} and the light color $l.color$, which already considers attenuation. Based on these data, the lighting model delivers the resulting color for the currently considered fragment and thus generates the finished image of the scene on the output medium.

IV. Results

After the implementation details, the images resulting from the rendering will be presented and discussed on their appearance. By selecting a suitable scene, it is in particular possible to validate the rendering results with respect to underlying measurement data (see Section IV-A). In the second part a comparison with the software LightDriver will be done. The LightDriver is an established tool for headlamp simulation from HELLA that has been in use for many years. It is successfully applied to support the development process and can therefore be regarded as validated. For this reason, it serves to validate the solution presented here using more realistic scenes as in Section IV-A. In addition, Section IV-C contains some remarks on the real-time capability of the simulation. Due to the dependence on countless factors (hardware and software configuration used, complexity of the scene and the depth of detail of the objects in it, number of simulated vehicles and other light sources, etc.), no general statement can be made here about the required calculation time per frame. The basic usability of the implementation for the night driving simulation should nevertheless be proven.

A. Validation

As described in Section II, the radiation characteristics of a light source can be defined by plotting the luminous intensity over the angles φ and θ in the spherical coordinate system. In this way, the luminous intensity can be represented in a plot, in which it is coded by false colors. In order to validate the rendered light based on the actual light distribution of the simulated headlamp, the artificial scene shown in Figure 13 is used.

It contains only three elements - a gray sphere, a left-side HD84 headlamp in the center of it and a camera defining the rendering perspective. Position and orientation of camera and light source are equal. They can be localized at the origin

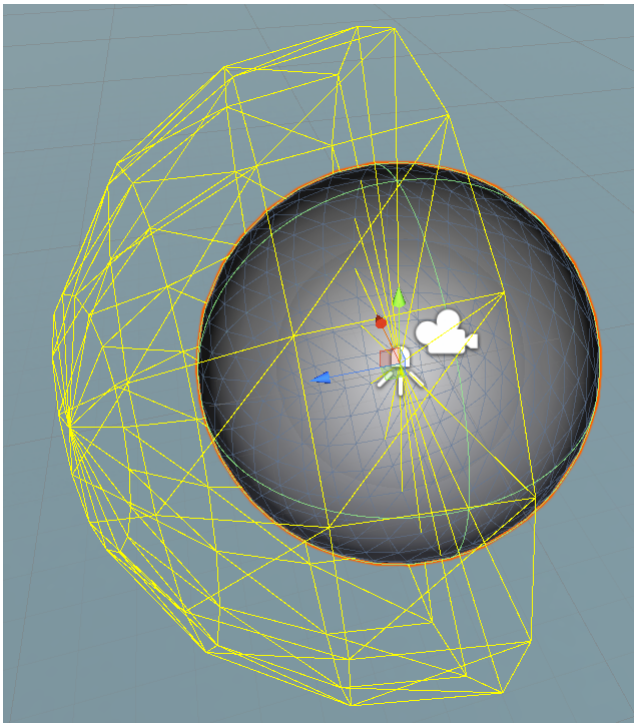


Figure 13. Artificial validation scene to compare the rendered headlight with measurement data.

of coordinate system represented by the blue, red and green arrows, whereby the blue arrow defines the viewing direction of the camera as well as the light axis ($\varphi = \theta = 0°$). The field of view of the camera is chosen to $60°$, which corresponds to a view frustum tip angle of $120°$ in the horizontal plane. The already mentioned half-spherical light volume approximated by a mesh of 65 vertices is visualized by the yellow wireframe.

According to this scene definition, the observer is in the center of the gray sphere and sees the light as it is emitted by the light source onto the inner wall of the sphere according to the implementation presented here. Running the headlamp with electrical current values to generate a low beam distribution, the rendering process produces the image shown in the upper part of Figure 14.

In rendered light, classic properties of a low beam distribution can already be perceived. In particular, the level of the light-dark boundary line along the vertical central axis should be mentioned here. The considerably greater horizontal spread of light compared with vertical spread can also be observed. For the complete validation of the implemented solution, however, the alignment with the measured data is necessary. For this purpose, it is shown in the lower part of Figure 14. In a direct comparison, the similarity of shape between rendering and measurement can be well demonstrated. At the same time, however, it is observed that the limited contrast of the light resolved linearly in gray scales is neither sufficient to perceive intensity differences in the bright area of the light distribution nor to perceive curtains in the edge area.

For this reason, the illuminance of the light on the elements of the scene was rendered logarithmically with another shader and coded in false colors. Luminous intensity I and

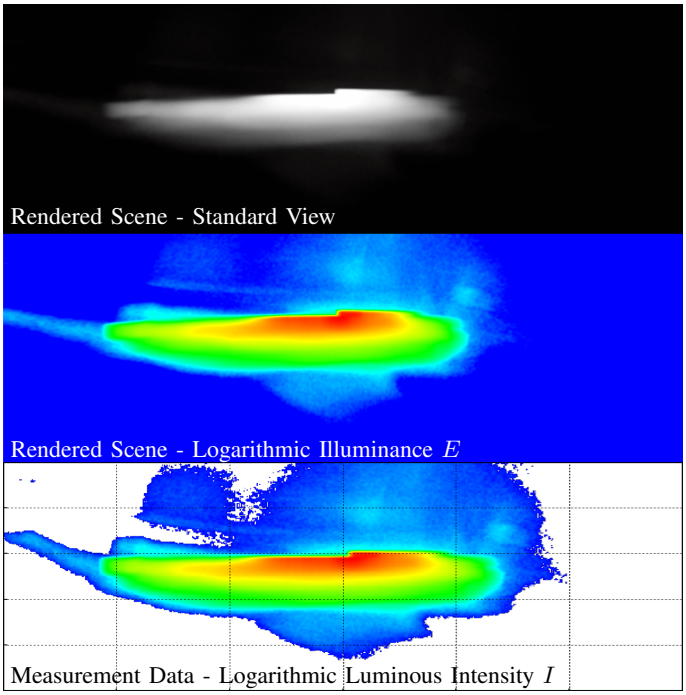


Figure 14. Comparison of rendered light (top) of left HD84 headlamp in low beam mode (false color coded illuminance on the spherical wall for better comparabilimty shown in the center area) and measured data of the same configuration (bottom).

illuminance E are related according to

$$E = \frac{I}{r^2} \cdot \cos(\epsilon), \tag{9}$$

where r is the distance between the light source and the illuminated object and ϵ is the angle between the light beam and the surface normal of the illuminated object [16]. In the specially selected scene, the distance between the light source and the illuminated object is constant at each pixel with the sphere radius. At the same time, each illuminated area is perpendicular to the incident light beam. This means that illuminance and luminous intensity have the same characteristics, as there is a linear relationship between them, making it particularly easy to compare the image produced with the measured data.

As it turns out, the correspondence of the simulated light distribution (center area of Figure 14) with the measured data (lower area of Figure 14) is even more convincing in the logarithmic false color representation. The gradations of intensity in the center of the light cone as well as the light curtains in the edge area can be found in both figures in the same way. These checks were also carried out for other light distributions and led to the same observations, so that the implementation can be considered as validated.

B. *LightDriver*

Even if the correctness of the implementation in terms of the mathematically correct reproduction of the luminous intensity distribution has already been carried out in Section IV-A, the appearance of the light distribution on the road should also be evaluated. In such an investigation, significantly more factors have an influence. The limitations of the output device (especially with regard to luminance and contrast), the modeling accuracy of the scene (textures, normal maps, reflection properties, ...) and the complexity of the lighting model used (reflections, ray tracing, ...) are just a few examples. Since the approximate reconstruction of a real environment is not an acceptable effort, the HELLA LightDriver (64 Bit Version built on July, 2017) should be used as a reference instead. This is HELLA's own development for night driving simulation, which has been used successfully for several years in the development process of new headlamp systems and lighting functions. In contrast to the implementation presented here, the LightDriver is not able to change the light distribution of a headlamp dynamically, as is necessary for the simulation of an HD system. Nevertheless, this fact does not limit its suitability for validating this implementation. The desired light distributions can be calculated in advance and then loaded into the LightDriver as a static total light distribution.

Figure 15 compares the low beam distribution of the HD84-Matrix-LED headlight (left and right headlamp) as calculated by Cookie Combiner- and Headlight-Shader (top) with the low beam distribution of the LightDriver as reference (bottom) in a simple street scene. In the right area of the figure, the scene is complemented by a white measuring wall with red control lines at a distance of 10m from the vehicle. This is a classic analysis tool for the evaluation of light distributions, as their shapes are more recognizable on this. The two vertical control lines are aligned with the mounting positions of the headlamps. While in the presented implementation only the electrical current values of the individual light sources belonging to this light distribution are specified, the LightDriver requires a complete light distribution as input, which is therefore calculated in advance. The scene for this comparison could not be taken directly from the LightDriver and was therefore recreated. As a consequence the textures and colors of scene objects are not exactly the same, but should suffice for a plausibility check.

If one compares the low beam distributions in the left area of Figure 15, they match well overall. A closer look reveals slight differences. On the one hand, the basic color or brightness on the street appears not quite uniform in both representations. In addition, it seems that the light curtains directly in front of the vehicle are of different brightness. The reasons for the deviations mentioned can be many and varied due to the large number of influencing factors already mentioned. However, despite its successful use in the development process, the LightDriver should not be regarded as an exclusive measure of optimality. The differences in the images result to a large extent from the use of HDR colors in the implementation presented, while the LightDriver uses classic SDR colors. In particular, the more clearly visible light curtains in front of the vehicle in this implementation are caused by this. Further differences may result from slightly different definitions of the lighting model and the scene objects.

In addition to these minimal differences, however, both simulations match, so that the rendering method presented here can also be checked for plausibility in realistic scenes. With the help of the road markings, the qualitative form equality of the illuminated road areas of both implementations can be recognized. In addition, the different light ranges for the left and right lanes are clearly visible in both representations. For this reason, the white control areas at the edges of the road on the right-hand side are illuminated at a greater distance than on the left-hand side. This characteristic is typical for low beam distributions and ensures the best possible illumination of one's own lane without glaring oncoming traffic.

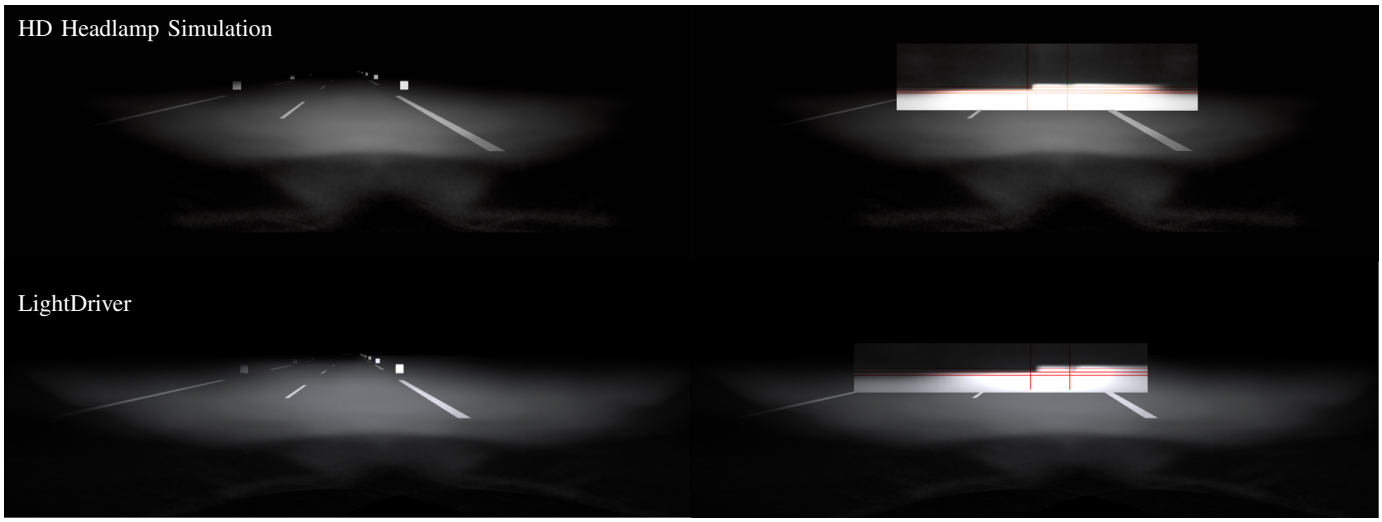


Figure 15. Comparison of the implementation presented (top) and HELLA LightDriver (bottom) simulating a low beam distribution in a simple street scene (left) and with a measuring wall in 10m distance (right).

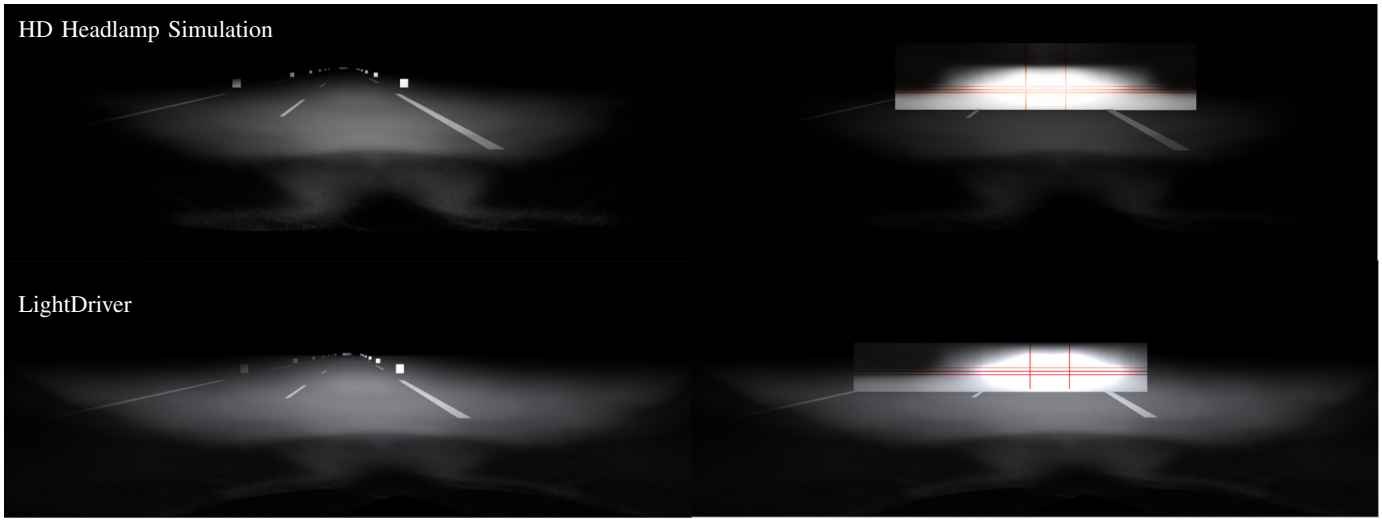


Figure 16. Comparison of the implementation presented (top) and HELLA LightDriver (bottom) simulating a high beam distribution in a simple street scene (left) and with a measuring wall in 10m distance (right).

Even if the light distribution on the road is the central evaluation criterion for the driver, the observation of light distributions on a vertical measuring wall has proven to be useful especially for comparison purposes. The contours of the light distribution, which are called the light-dark boundary in this context, become clear through the close projection of distant areas. The varying illumination distances of the lanes, which could already be observed through the control surfaces, become even clearer on the measuring wall. They can be found in the typical low beam distribution step in the upper middle area of the light-dark boundary. As can be seen from the simulation images, these steps form in the middle of the left and right headlamps, which can be identified by the vertical red control lines.

After comparing the low beam distributions in both simulations, Figure 16 shows a similar comparison for the high beam as a second light distribution of central importance. For the LightDriver the light intensity distribution was precalculated again and loaded as static total light distribution. As in the case of low beam, a good overall agreement can be observed in addition to differences in detail between the images.

As was to be expected, the illumination of the distant areas has increased in comparison with the low beam in Figure 15. The difference is particularly noticeable on the opposite lane and the control surfaces positioned along it. This effect is achieved by a fundamental change in the shape of the light distribution, as can be observed on the measuring walls on the right-hand side of Figure 16. The step in the light-dark boundary of the low beam disappears when high beam is used. Instead, a symmetrical light distribution is generated with a so-called high beam cone, which illuminates the distant areas in front of the vehicle.

C. Computational Effort

With regard to the application of the implementation presented here in an interactive night driving simulation, compliance with the real-time requirements must be considered. In order to give the viewer the impression of a dynamic scene, at least 30 frames per second (fps) should be calculated. The optimal case is 60 fps, which corresponds to the refresh rate

of most output devices [28]. A further optimization of the computing time does not lead to any benefit after this limit.

At the same time, however, it should be mentioned that a computer with a Windows operating system is not a real-time system. Due to the uninfluenceable scheduling of all running processes, measurements of calculation times are always rough approximations and can be subject to strong fluctuations.

In addition, the calculation times naturally depend heavily on the hardware and software configuration. The results discussed below were recorded on a mobile PC, whose specification can be found in Table I.

TABLE I. HARD- AND SOFTWARE-CONFIGURATION OF THE MEASURMENT PC

Operating System	Windows 10 Pro 64-bit (10.0, Build 16299)
Used Graphics Engine	Unity3D, Version 2017.3.1f1
Model	Dell Precision 7710
Processor	Intel(R) Core(TM) i7-6820HQ CPU @2.7GHz
Memory	16384 MB RAM
DirectX Version	DirectX 12
Graphic	NVIDIA Quadro M3000M
Video Memory	4062 MB VRAM (+ 8133 MB Shared)

In order to comply with real-time requirements, only those processes are relevant that are constantly being executed at runtime. In this application, these are Cookie Combining and lighting. Both can be examined in isolation from each other. The Profiler available in Unity is used to measure the calculation times. With this profiler, the calculation times of the CPU and GPU can be broken down frame by frame with regard to the operations performed.

First, the performance of the cookie combining is examined. It depends only on the number of single light sources and the resolution of the floating point textures describing their luminous intensity distribution (here: 95 32bit floating point textures with 900x900 texels). To exclude possible influences of the program start and caching, the total light distribution is calculated many times. All relative current values are randomly selected between 0% and 100% for each calculation.

The analysis in the profiler shows that combining the 95 floating point textures on the CPU requires an average of 0.45ms (min: 0.27ms, max: 0.64ms). With cookie combining, however, the CPU acts primarily as the client of the GPU. It instructs the graphics card to execute the Cookie Combiner shader by creating draw calls and defines the relevant context information, such as the render target or the current individual light distribution. So it is not surprising that the GPU has a significantly higher average calculation time of 4.61ms (min: 4.44ms, max: 4.73ms).

In addition to cookie combining, the calculation time of the headlight shader must also be examined. Due to the deferred rendering pipeline used here, this is called only once per light source and frame. The effort of the lighting therefore does not depend on the scene complexity, but only on the resolution of the output and thus of the light buffer, which is selected here as 1430x780 pixels. On the CPU, the lighting of a spotlight simulation requires an average of 0.02ms (min: 0.01ms, max: 0.04ms). The calculation time on the GPU measures approximately 0.097ms (min: 0.095ms, max: 0.114ms). In this respect, the effort of the actual lighting can be neglected compared to the calculation time of the Cookie Combining.

In comparison with the much simpler implementation of cookie combining, this result may seem astonishing. However, the GPU can perform the lighting calculations for all pixels of the light buffer in parallel. The Cookie Combiner shader, on the other hand, must be called repeatedly in one frame for all light sources with changing context information.

As a result, both shaders are performant enough to be used in the night driving simulation of a single vehicle with two headlamps. If the simulation of several vehicles with HD headlamps is intended, Cookie Combining will reach the limits of available computing power even for small vehicle quantities. If dynamic HD headlights are desired on all vehicles, more powerful graphics cards can be used to shift these limitations within certain bounds. On the other hand, the question arises whether dynamic lighting functions are required. If this is not the case, the precalculation of static light distributions (e.g. dipped beam) is a sensible alternative for third-party vehicles. In this way, they retain the basic light characteristics of an HD headlamp without making significant use of resources. The ego vehicle can still be simulated with dynamic lighting functions using cookie combining.

V. CONCLUSION

This contribution presents an approach for real-time simulation of dynamic HD headlamp systems and thus lays the foundation for the simulation-based development of high-resolution dynamic light functions. The implementation reproduces the real light distribution accurately and is also executable on average hardware for today's standards.

This contribution is motivated by the completely missing possibility of real-time simulation of high-resolution headlamp systems, which is indispensable for a systematic and verifiable procedure of development. In addition, there is the need for darkness and suitable weather conditions during real test drives, which cannot be completely eliminated by using a simulation, but can be significantly reduced.

Before the simulation of HD headlamp systems is presented, their technical structure and functionality are described in Section II. During the course, the light intensity distributions established in the context of headlamp measurement will also be introduced. They describe the luminous intensity emitted by a light source depending on the spatial direction. The HD84-Matrix-LED headlamp type under consideration here consists of a matrix with 84 LEDs and 11 further light sources (apron area lights, bend lights, additional high beam lights). The luminous intensity distribution is known for each of these light sources. They can be operated with current control, which results in corresponding scaling of the luminous intensity values. By locally separating their illumination areas, it is possible to build the overall light distribution summing up the individual distributions, whereby their shape variability is limited only by the resolution of the headlamp. An overview of this procedure is illustrated by Figure 4.

The following Section III describes the implementation, the overall scheme of which is shown in Figure 5. First of all, the digital representation of a luminous intensity distribution by a texture is defined in Section III-A. In particular, 32bit floating point textures are used to capture the high light dynamics in a night driving scene in detail. Since the underlying measurement data are available with an angular resolution of 0.2° over a range of 180° in horizontal and 50° in vertical direction, textures with 900x900 texels are selected. The measuring points can thus be mapped unchanged onto the texels (see Figure 6 and Equation (7)). Light intensity values between the measuring points are bilinearly interpolated according to

Figure 7 and Equations (5) and (6). In preprocessing, the measured luminous intensity distributions for all light sources in the headlamp can be converted into floating point textures and thus used for subsequent rendering.

The rendering of the headlight is divided into two steps. In the first step, the current total light distribution is determined as described in Section III-B. It is the weighted sum of the 95 individual light distributions of all adjustable light sources of the HD84-Matrix-LED headlamp. In the technical implementation, the light distributions are represented by textures and superimposed by blending the results of the Cookie Combiner-Shader applied to them as shown in Figure 8. Mathematically formulated, the total light distribution is formed as a linear combination of the matrices or textures of the individual light distributions (see Figure 9). Each individual light distribution is weighted with the present relative current value. In order to provide the optimum overall light distribution in every driving situation, the current values can be adjusted with a frequency of 50 Hz. The output of the Cookie Combiner-Shader is a floating point texture of the overall light distribution, which is structurally no different from that of the individual distributions.

The total light distribution initiates the second rendering step as input for the Headlight-Shader discussed in Section III-C. This is integrated into the deferred rendering pipeline by using Unity's Graphics Command Buffer. As Figure 11 illustrates, it is downstream of the standard operations in the lighting pass. In this way, compatibility with all standard lights in the scene is maintained. In addition, HDR rendering is used to reproduce the high contrasts of a night drive scene with the existing limitations of the output device as detailed as possible. Along the pseudo code the procedure in the headlight shader is discussed, whose final task is the determination of a color value for the respective pixel of the light buffer or the output under consideration of the current light source. The essential spaces, geometric correlations and vector operations can be understood by Figure 12. After finding all relevant information for the determination of the pixel color, this is transferred to a Unity-internal lighting model. This model finally returns the resulting color, which corresponds to the output of the Headlight-Shader.

Section V concludes the contribution by presenting the results. In the first step, the simulated luminous intensity distribution is validated on the basis of the measurement data. For this purpose, an artificial scene is created which serves this purpose exclusively. It contains only a sphere and in its center a HD84-Matrix-LED headlamp to be simulated. The background of this scene definition is the linear relationship between the luminous intensity of the headlamp and the illuminance on the spherical surface, as can be seen from Equation (9). In this way, in Figure 14, the simulated illuminance on the spherical surface and the measured luminous intensity distribution can be compared. The agreement of these data is convincing.

After the proof for the mathematically correct reproduction of the real light intensity distribution has been provided, the light impression in a more realistic scene is evaluated in Section IV-B. In such an evaluation, far more influences come into play than can be controlled within the framework of the implementation presented here. For this reason, the night driving simulation software LightDriver developed by HELLA serves as orientation for evaluating the rendering results. As a tool for headlamp development that has been established for years, it is suitable as reference, even if it is not an incontestable optimum. As Figures 15 (low beam) and 16 (high beam) show, the implemented simulation is very similar to the LightDriver. The differences can mainly be traced back to the unequal scenes and light models, as well as the higher luminous intensity resolution and HDR rendering used here. Consequently, they do not represent a quality defect of the presented implementation.

The performance analysis shows that the major share of the computing effort is attributable to Cookie Combining. Even if the simulation of a vehicle on the mobile PC with which the computing time measurements were carried out does not pose a problem, it should be noted that Cookie Combining quickly reaches the limits of computing power when simulating several vehicles. These can be moved upwards by using better hardware (in particular graphics card). As an alternative, the application of static light distributions for external traffic is proposed. In this case, the Cookie Combining must only be carried out for the ego vehicle.

VI. Future Work

In view of the good rendering results, future work will focus less on the further development of the presented implementation. Nevertheless, the lighting model, for which the Unity standrad BRDF model has been used so far, could be replaced by an own implementation in the future, depending on the resulting improvements. In addition, it should be checked whether Cookie Combining, which represents the bottleneck of the implementation in terms of computing time, provides potential for further performance improvement.

The presented work should serve much more as a basis for the night driving simulation, for the implementation of which various follow-up work is necessary. First of all, the previously manual setting of the current values for the individual light sources must be replaced by the integration of the headlamp control unit. This integration is divided into two steps. On the one hand, the outputs of the control unit must be received by the visualization system and transferred to an intensity list (see Figure 5) that is compatible with the Cookie Combiner. This step could already be performed at the current time with the necessary requirement of 50 Hz. On the other hand, the connection of the control unit only makes sense if it knows about the current traffic situation in order to select the light distribution based on it. For this purpose, the real sensors in the vehicle must be simulated by virtual sensors. The central role is played by the surrounding camera. Various approaches to implementation are currently being investigated, including machine learning methods on the images rendered from the point of view of the surrounding camera.

A third approach for further work is the use of analysis tools to assess the headlight in the scene. For this purpose, false color and iso-line representations have already been implemented. These can be applied to pixel brightness or illuminance. Their scaling to physical quantities and their adaptation to legal or customary standards will still have to take place in the future.

Acknowledgment

The authors would like to thank HELLA for providing the light distribution data of the HD84-Matrix-LED headlamp and the LightDriver simulation software. This paper is part of the EFRE.NRW (European Fonds for Regional

Development, North Rhine-Westphalia)-funded project 'Smart Headlamp Technology (SHT)'.

REFERENCES

[1] N. Rüddenklau, P. Biemelt, S. Henning, S. Gausemeier, and A. Trächtler, "Shader-based realtime simulation of high-definition automotive headlamps," IARIA SIMUL 2018, The Tenth International Conference on Advances in System Simulation, 2018.

[2] B. Fleury, L. Evrard, J.-P. Ravier, and B. Reiss, "Expanded Functionality of Glare Free High Beam Systems," ATZ Autotechnology, 2012.

[3] M. Enders, "Pixel light," Progress in Automobile Lighting (PAL), 2001.

[4] F.-J. Kalze and D. Brunne, "LED im Fahrzeug: Die Rolle der Matrixscheinwerfer und was sie leisten (LED in the Vehicle: The Role of Matrix Headlamps and what they perform)," Elektronik Praxis, 2018.

[5] C. Schmidt, B. Willeke, and B. Fischer, "Laser versus Hochleistungs-LED - Vergleich der Einsatzmöglichkeiten bei hochauflösenden Matrix-Scheinwerfer-Systemen (Laser versus High Power LED - Comparison of Application Possibilities for High-Definition Matrix Headlamp Systems)," VDI-Tagung Optische Technologien in der Fahrzeugtechnik, Karlsruhe, 2016.

[6] J. Roslak and C. Wilks, "Hochauflösende LCD-Scheinwerfer - Herausforderungen für Elektronikarchitekturen (High-Definition LCD Headlights - Challenges for Electronic Architectures)," Automobiltechnische Zeitschrift elektronik (ATZelektronik), 2017.

[7] P. Lecocq, J.-M. Kelada, and A. Kemeny, "Interactive Headlight Simulation," Driving Simulation Conference, 1999.

[8] J. Berssenbrügge, J. Gausemeier, M. Grafe, C. Matysczok, and K. Pöhland, "Real-Time Representation of Complex Lighting Data in a Nightdrive Simulation," 7. International Immersive Projection Technologies Workshop, 9. Eurographics Workshop on Virtual Environments, 2003.

[9] J. Löwenau and M. Strobl, "Advanced Lighting Simulation (ALS) for the Evaluation of the BMW System Adaptive Light Control (ALC)," International Body Engineering Conference & Exhibition and Automative & Transportation Technology Conference, 2002.

[10] A. Kemeny et al., "Application of real-time lighting simulation for intellignet front-lighting studies," Driving Simulation Conference, 2000.

[11] J. Berssenbrügge, J. Bauch, and J. Gausemeier, "A Virtual Reality-based Night Drive Simulator for the Evaluation of a Predictive Advanced Front Lighting System," Design Engineering Technical Conferences & Computers and Information in Engineering Conference, 2006.

[12] J. Berssenbrügge, S. Kreft, and J. Gausemeier, "Virtual Prototyping of an Advanced Leveling Light System Using a Virtual Reality-Based Night Drive Simulator," Journal of Computing and Information Science in Engineering, 2010.

[13] A. Knoll et al., "Evaluation of an Active Safety Light using Virtual Test Drive within Vehicle In The Loop," IEEE International Conference on Industrial Technology, 2010.

[14] "AutomobilIndustrie: Adaptives LCD-Licht mit 30.000 Pixeln (Automotive Industry: Adaptive LCD-Light with 30,000 Pixels)," 2017, URL: https://www.automobil-industrie.vogel.de/adaptives-lcd-licht-mit-30000-pixeln-a-629502/ [retrieved: 5, 2019].

[15] K. Reif, Ed., Automobilelektronik (Automotive Electronics), p. 301 ff. Vieweg+Teubner, GWV Fachverlage GmbH, Wiesbaden, 2009, ISBN: 978-3-8348-0446-4.

[16] F. Pedrotti, L. Pedrotti, W. Bausch, and H. Schmidt, Eds., Optik für Ingenieure - Grundlagen, 4. Auflage (Optics for Engineers - Basics, 4th edition), p. 13 ff. Springer, Berlin, 2007, ISBN: 978-3-540-22813-6.

[17] H.-H. P. Wu, Y.-P. Lee, and S.-H. Chang, "Fast measurement of automotive headlamps based on high dynamic range imaging," OSA Applied Optics Vol. 51, 2012.

[18] A. S. Glassner, Ed., An Introduction to Ray Tracing. ACADEMIC PRESS INC., San Diego, CA 92101, 1989, ISBN: 0-12-286160-4.

[19] R. Neumann and H. Hogrefe, "Computer simulation of light distributions for headlamp systems," SAE Technical Paper, 1991.

[20] "Unity Homepage," URL: https://unity3d.com/ [retrieved: 5, 2019].

[21] R. Fernando and M. J. Kilgard, Eds., The Cg Tutorial: The Definitive Guide to Programmable Real-Time Graphics. Addison Wesley Pub Co Inc., Feb. 2003, ISBN: 978-0321194961.

[22] T. Saito and T. Takahashi, "Comprehensible Rendering of 3-D Shapes," SIGGRAPH '90, Dallas, 1990.

[23] J. F. Hughes et al., Eds., Computer Graphics - Principles and Practice, 3th Edition. Addison-Wesley, Jul. 2013, ISBN: 978-0321399526.

[24] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, Eds., High Dynamic Range Imaging - 2nd Edition. Morgan Kaufmann, 2010, ISBN: 978-0-12-374914-7.

[25] "Graphics Command Buffer," 2018, URL: https://docs.unity3d.com/Manual/GraphicsCommandBuffers.html [retrieved: 5, 2019].

[26] "Unity Blog: Extending Unity 5 rendering pipeline: Command Buffers," 2015, URL: https://blogs.unity3d.com/2015/02/06/extending-unity-5-rendering-pipeline-command-buffers/ [retrieved: 5, 2019].

[27] A. Beutelspacher and U. Rosenbaum, Eds., Projektive Geometrie, 2. Auflage (Projective Geometry - 2nd edition), p. 63 ff. Vieweg, Wiesbaden, 2004, ISBN: 978-3-528-17241-1.

[28] A. Banitalebi-Dehkordi, M. T. Pourazad, and N. Panos, "Effect of High Frame Rates on 3D Video Quality of Experience," IEEE International Conference on Consumer Electronics (ICCE), 2014.

Material Requirements Planning Performance Improvement due to Safety Stock Relaxation

Klaus Altendorfer, Sonja Strasser, Andreas Peirleitner
Production and Operations Management
University of Applied Sciences Upper Austria
Steyr, Austria
e-mail: {klaus.altendorfer, sonja.strasser, andreas.peirleitner}@fh-steyr.at

Abstract—**Material Requirement Planning (MRP) is a broadly applied production planning method. One problem reported by practitioners and identified in research is that capacity constraints are not included in the planning algorithm. In this paper, the implementation of a simple capacity balancing function into the MRP run by allowing to temporarily relax the safety stock is investigated. Since such a safety stock relaxation method can be implemented in different ways, three specific implementations are developed and tested in a simulation study. For a simple production system structure with uncertainties in processing and customer demand, the performance improvement of the different safety stock relaxation methods is tested when a rolling horizon MRP planning is applied. A detailed analysis of planning parameter effects is presented and a broad set of scenarios provides further insights in the performance of the developed methods. In general, all three methods reveal a significant potential of improvement in comparison to MRP. Managerial insights are that too low production lot sizes and too low safety stocks should be avoided and the interaction between these two planning parameters cannot be neglected. Furthermore, very high and very low production system utilization reduce the improvement potential.**

Keywords – Material Requirements Planning; Rolling Horizon Planning; Discrete Event Simulation; Sensitivity Analysis; Safety Stock Relaxation.

I. INTRODUCTION

Material Requirement Planning (MRP) is widely applied for production planning due to its well comprehensible algorithm for scheduling production orders to satisfy material requirements. However, one problem in MRP application is that capacity constraints are ignored in the planning algorithm. In this paper, different methods for temporary safety stock relaxation within the MRP run to enable a capacity balancing are investigated. Note that this paper is an extension of [1] where preliminary results have been presented. Specifically, [1] is extended by a more thorough safety stock relaxation method presentation, a specific algorithmic explanation of its integration into MRP, a broad numerical study design and the specific formulation of managerial insights based on the numerical results.

MRP is studied a lot in literature, see [2] for its basic development, [3] for a detailed discussion, and [4] for parameter optimization. Specifically the problem of neglecting the limited capacity has often been addressed in literature [5][6][7]. Literature shows that neglecting capacity constraints leads to the generation of usually infeasible production plans by MRP, which require additional planning effort at the production control level [8][9][10]. In the last decades, several approaches have been developed to deal with the drawbacks of MRP. Especially for the integration of capacity constraints, there exist a set of different solution approaches [11]. One possibility is to react on capacity problems after the MRP run [8][9][12], although it is hard to solve these problems, which are generated at the higher MRP level. Some authors start before the MRP run and try to avoid capacity violations already at the Master Production Schedule (MPS) level [13][14]. Another approach is the formulation of an optimization problem with capacity constraints instead of the MRP run [15][16], or the including of a solution heuristic into the MRP algorithm [11]. In addition to the high computational effort for solving real world planning problems, the theoretical formulations limit the practical application of these approaches. The integration of a solution heuristic into the well-known MRP algorithm for tackling the capacity constraints is another possibility, which is more likely to be accepted for practical implementations. Different approaches can be found in [11], [17], [18], [19], or [20].

In [19], capacity planning is integrated into MRP by providing simple algorithmic measures, like the temporary relaxation of safety stock, load dependent dynamic planned lead times and lot size adaption heuristics. The concept developed in [19] is Material and Capacity Requirements Planning (MCRP), however, only a conceptual framework is provided, but details on the implementations are missing. In [21] the concept of MCRP is further detailed and some first insights on the overall performance of the MCRP algorithm are presented. However, details on the performance of different safety stock relaxation methods are still to be investigated.

The above introduced literature shows that the implementation of capacity limits into MRP is a relevant field of research. Practical requirements often imply that such solution heuristics should be easy to implement, to

enable a further real world implementation of the specific methods. Therefore, the implementation of simple safety stock relaxation methods into the MRP run for enabling capacity balancing already within the planning algorithm provide a significant contribution. In this paper, three safety stock relaxation methods for capacity load balancing are developed, based on the conceptual framework of [19]. Related to these methods, the following research questions are addressed:

- What is the performance improvement potential of the different safety stock relaxation methods in comparison to MRP?
- What is the influence of the planning parameters lot size and safety stock on the performance of the developed safety stock relaxation methods and how do these parameters interact?
- How do tardiness costs, production system utilization and setup effects influence the performance of the developed methods?
- What safety stock relaxation method has the highest improvement potential and can be applied for further research and in practical applications?

To answer these research questions, Section II provides the algorithmic extension of MRP and a detailed explanation of the different safety stock relaxation methods. For evaluating the performance of these methods, a simulation study is performed. The respective production system setup and the evaluated scenarios are introduced in Section III. To identify the general performance improvement potential of safety stock relaxation in comparison to MRP, the numerical results of a basic scenario are presented in Section III as well. The detailed planning parameter influence is evaluated in Section IV, where again the basic scenario is focused. The influence of different tardiness costs, production system utilization and setup effects are then evaluated in Section V with a broad numerical simulation study. Furthermore, the different methods performance is compared in this section in detail. In Section VI, concluding remarks summarize the main results and outline future research.

II. SAFETY STOCK RELAXATION

A safety stock within MRP is applied to reduce the negative effects of uncertainties in customer demand and production processes. From a planning perspective, the safety stock is never undershot in the original MRP algorithm (see netting in MRP algorithm, [3] and [2]) and is only used for unplanned occurrences. In the approach introduced in [19], safety stock is already applied in the planning algorithm for capacity load balancing, i.e., available safety stocks are used to temporary reduce the capacity needed. This leads to a shift in capacity consumption since this safety stock has to be refilled in later periods, which leads to a higher capacity consumption there. The basic idea behind that measure is that capacity shortages are only temporary and, therefore, some idle capacity is available further in the future, i.e., capacity load is balanced.

In Fig. 1 the MRP algorithm with the extension of the safety stock relaxation is presented. The MRP algorithm starts at Low Level Code 0 (LLC), which usually includes the sales parts, with the step *netting* for each material. The inputs are the gross requirements of LLC0 from customer orders or master production scheduling, the scheduled receipts from production orders currently processed, and the current inventory. After netting, the step *lot sizing* is applied, followed by the step *capacitating*.

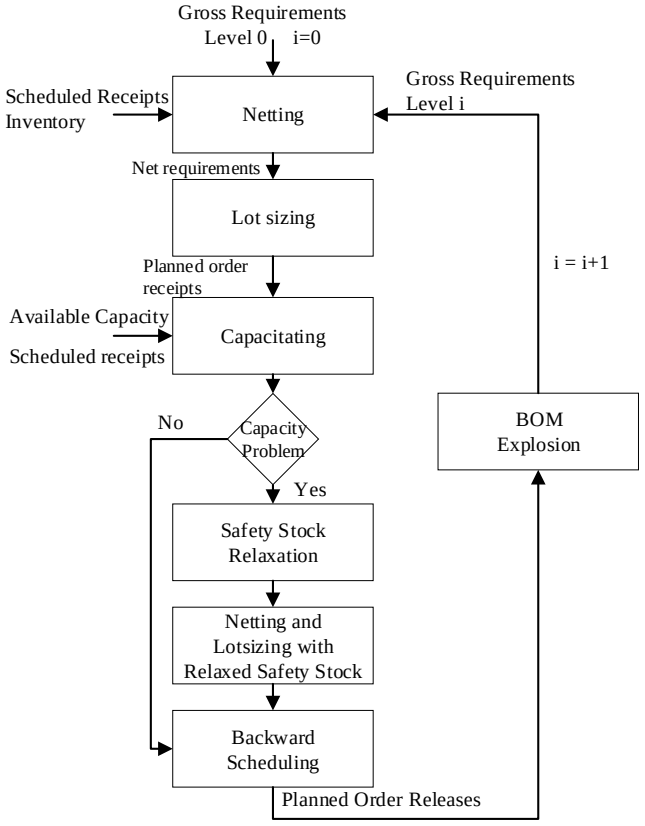


Figure 1. MRP Algorithm with Safety Stock Relaxation.

The capacitating step is fundamental for applying the safety stock relaxation because capacity constraint violations are determined by comparing *cumulated capacity available* with *cumulated capacity needed*. For each period within the MRP run, the capacity needed is calculated based on scheduled receipts and planned order receipts applying the corresponding processing and setup times for all materials and all machines at the current LLC. The available capacity is given by the work schedule applied for each machine. Whenever the *cumulated capacity needed* is higher than the *cumulated capacity available* within one planning period, a safety stock relaxation is applied. Note that the steps *capacitating* and *safety stock relaxation* are not part of the traditional MRP run. After *safety stock relaxation*, the steps *netting* and *lot sizing* are again performed with the relaxed safety stocks. These lower safety stocks lead to lower net

requirements and, therefore, influence the resulting production lot sizes. The next steps are *backward scheduling* and *bill of material* (*BOM) explosion*. The steps *backward scheduling* and *BOM explosion* are also executed if no capacity problem has been detected. In the following subsections, the different methods for safety stock relaxation are introduced.

Fig. 2 shows an example for the *capacitating* step, where a capacity problem is detected in period 5. Note that this calculation is applied for each machine within the production system. In this example the available capacity for each planning period, i.e., periods applied in the MRP run, is constant. This could be 8 hours capacity available for each day. The *cumulated capacity needed* includes all capacity demands from currently processed orders and new production lots resulting from the MRP step *lot sizing*. The capacity needed is scheduled at the planned end date of the order in this calculation. Whenever the *cumulated capacity needed* is higher than the *cumulated capacity available* a capacity problem occurs. In the example in Fig. 2, such a problem occurs in period 5. This capacity problem is the basis for the different safety stock relaxation methods discussed in the following subsections.

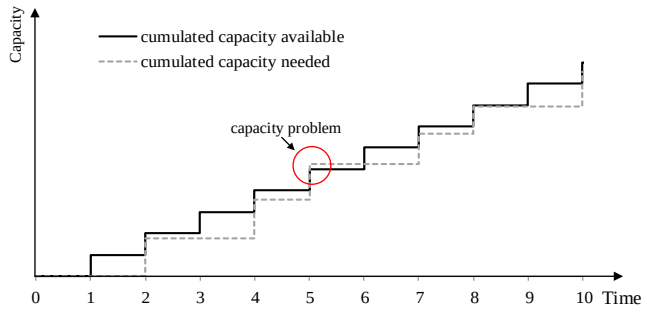


Figure 2. Cumulated capacitiy available vs. capacity needed.

A. *Safety Stock Relaxation Method 1*

In safety stock relaxation method 1, the safety stock for a specific material can only be reduced if there is a planned order receipt in the period of the capacity problem. Note that in the numerical study, the lot sizing rule FOP (Fixed Order Period) is applied, which summarizes net requirements of multiple periods, so there is not a planned order receipt for each material in every period. The safety stock is reduced to the level needed, so that the capacity problem is solved in the respective period; i.e., *cumulated capacity needed* after safety stock relaxation is equal or smaller than the *cumulated capacity available.*

Fig. 2 showed a capacity problem in period 5 of c=1.7 hours. If we assume that the processing time for a material, which has a planned order receipt in period 5 is p=2.55 minutes at the respective machine, the safety stock would be relaxed by r=40 pcs, i.e., $r=c/(p/60)$ (1.7 / (2.55/60)), for this material. Please note that the safety stock is relaxed for all periods of the current FOP lot size. To illustrate the specific safety stock relaxation methods, a numerical example is generated in which the capacity demand from Fig. 2 is generated by only two materials, which both apply the lot sizing policy FOP 3. Note that this assumption only applies for the simple illustrative example here but not for the numerical study presented later. The key results of the MRP run for this example are presented in Table I, where the period with the capacity problem is shaded in grey. Note that the *net requirements* include a current inventory and no projected inventory on hand is reported here to keep the example simple. The result in Table I shows that the safety stock relaxation leads to lower *planned order receipts after relaxation*, i.e., a lower production lot size for the respective lot and, therefore, lower cumulated capacity needed. The following production lot, i.e., the *planned order receipts* in period 8, include the *net requirements* and the amount needed to refill the safety stock.

TABLE I. MRP RUN FOR SAFETY STOCK RELAXATION METHOD 1

Period	1	2	3	4	5	6	7	8	9	10
Gross Requirements	100	90	78	129	72	87	100	30	84	80
Scheduled Receipts	300	0	0	0	0	0	0	0	0	0
Net Requirements	0	75	78	129	72	87	100	30	84	80
Safety Stock before relaxation	285	285	285	285	285	285	285	285	285	285
Planned Order Receipts before relaxation	0	282	0	0	259	0	0	194	0	0
Safety Stock after relaxation	285	285	285	285	**245**	**245**	**245**	285	285	285
Planned Order Receipts after relaxation	0	282	0	0	**219**	0	0	**234**	0	0

To account for uncertainties, a minimum safety stock level can be considered as lower bound for the relaxation. For the application of the safety stock relaxation method, the materials are ordered according to their capacity consumption per piece at the respective machine. The method starts with the material, which has the highest capacity consumption per piece and is performed for further materials until the capacity problem is solved.

B. *Safety Stock Relaxation Method 2*

Safety stock relaxation method 2 extends the set of materials for which the safety stock relaxation can be applied. In method 1 the safety stock relaxation can only be performed, if there is a *planned order receipt* in the period of the capacity problem. In method 2, this restriction is removed. A safety stock relaxation can also be performed, if there is a planned order receipt that covers net requirements (due to lot sizing policy FOP) in the period of the capacity problem. This allows that the safety stock for *planned order receipts* with end dates before the period of the capacity problem can be relaxed. Table II shows the results for the safety stock relaxation of the second material, which leads to the capacity demand from Fig. 2. This material has the same processing time and a *planned order receipt* in period 4. Note that this example assumes that the safety stock of the material from Table I has not been relaxed. Again, the *net requirements* calculation is skipped for simplicity reasons.

TABLE II. MRP RUN FOR SAFETY STOCK RELAXATION METHOD 2

Period	1	2	3	4	5	6	7	8	9	10
Gross Requirements	91	92	112	93	95	120	43	86	91	92
Scheduled Receipts	0	230	0	0	0	0	0	0	0	0
Net Requirements	0	0	0	67	95	120	43	86	91	92
Safety Stock before relaxation	285	285	285	285	285	285	285	285	285	285
Planned Order Receipts before relaxation	0	0	0	282	0	0	220	0	0	92
Safety Stock after relaxation	285	285	285	**245**	**245**	**245**	285	285	285	285
Planned Order Receipts after relaxation	0	0	0	**242**	0	0	260	0	0	92

The results in Table II show that in this case the safety stocks for periods 4 to 6 are relaxed and that the following production lot, i.e., the *planned order receipts* in period 7, include the *net requirements* and the amount needed to refill the safety stock. Note that this relaxation solves the capacity problem of this simple example since the lower capacity demand in period 4 reduces also the cumulative capacity needed in period 5.

C. Safety Stock Relaxation Method 3

Safety stock relaxation method 3 is an extension to method 2 and uses the same logic for the safety stock relaxation. However, in a rolling horizon planning, methods 1 and 2 do not store the relaxed safety stock numbers for the next MRP run. In a pure deterministic setting this would lead to a situation where the safety stock relaxation decision has to be taken in each MRP run until the capacity problem has passed. In comparison to methods 1 and 2, the relaxed safety stock numbers are stored in method 3 for the MRP run performed in the next period. The next MRP run is calculated with the predefined relaxed safety stocks. Method 3 has the effect that when a safety stock relaxation for a *planned order receipt* is made, it is never revised. The only exception is that the safety stock can be further relaxed to the minimum safety stock, if there is a new capacity problem. Note that in a stochastic setting where demands and shop floor behavior incur uncertainties, this method may lose some flexibility to react on short term influences. The MRP run from Table II, in which the safety stock was relaxed from period 4 to 6, is repeated one period later in Table III. Note that period 3 in Table III corresponds to period 4 in Table II, and the period with the capacity problem shifted to period 4.

TABLE III. MRP RUN FOR SAFETY STOCK RELAXATION METHOD 3

Period	1	2	3	4	5	6	7	8	9	10
Gross Requirements	92	112	93	95	120	43	86	91	92	83
Scheduled Receipts	230	0	0	0	0	0	0	0	0	0
Net Requirements	0	0	67	95	120	43	86	91	92	83
Safety Stock before relaxation	285	285	**245**	**245**	**245**	285	285	285	285	285
Planned Order Receipts before relaxation	0	0	**242**	0	0	260	0	0	175	0

The relaxed safety stock from the previous period is already stated in the *safety stock before relaxation* and if a further capacity problem occurs further safety stocks could be relaxed.

To test the behavior of these three safety stock relaxation methods in stochastic environments with rolling horizon planning, a simulation study is performed.

III. SIMULATION STUDY

In this section the modeled production system for the simulation study and the different scenarios are described, followed by the planning parameters investigated. For a basic setting, the performance of the different safety stock relaxation methods is compared to standard MRP. The generic simulation framework SimGen based on AnyLogic©, also used in [22] and [23], is applied for the simulation study. This framework allows to implement production planning simulation models efficiently. For details, see also [24].

A. Production System

The modeled production system structure applied in this paper is motivated by different automotive suppliers' production systems and similar to the production system presented in [22]. However, it is a very streamlined version (low number of products, simple BOM structure, only one machine per low level code) to not disturb the simulation experiment results unnecessarily, which are generated later on. Fig. 3 shows the resources, bill of material and work schedule applied.

The studied production system is a pure Make-to-Order (MTO) system. Eight final products (LLC0) are delivered to a set of different customers stating their orders with a random customer required lead time in advance of the respective due date. These final products consist of 1 piece of a semi-processed material on LLC1 and LLC2, whereby the raw materials on LLC3 are assumed to be always available. One machine is available for each processing step and the transformation from one low-level code to the next always includes one processing step. The lot sizing policy is FOP for all materials (see [3] for details).

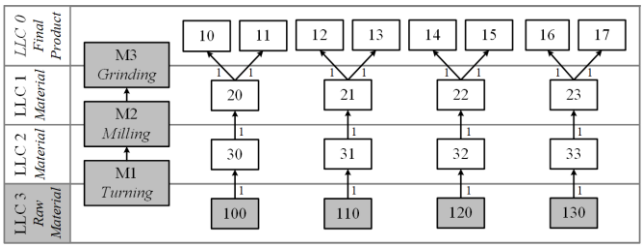


Figure 3. Production System.

B. Scenario Definition

To evaluate the performance of the safety stock relaxation methods in comparison to MRP, three different utilization levels are evaluated. This is necessary in order to study the effectiveness of the proposed methods, since for low work load, in comparison to the available capacity, methods for capacity balancing are getting pointless. The utilization factors evaluated in this study are 90%, 95% and 98%. The time that is spent for setup in comparison to the total capacity needed can also have an influence on the performance of the different safety stock relaxation methods. Therefore, two different percentages of setup activities (5% and 10%) and two different setup scenarios are investigated. In the standard setup scenario, called *always setup*, setup times occur for every order. In the second setup scenario, named *setup at material change*, setup times only occur if the next product produced is not the same as the previously produced. The different utilization scenarios are generated by varying customer demands. Starting with customer demands (log-normally distributed with a coefficient of variation of 1), which result in a shop load of 100% without setup, the utilization scenario is generated by multiplying the monthly demand with the utilization factor minus a predefined percentage of setup activities (5% and 10%). The resulting demand for, e.g., final product 10 with initial demand of 1,059 pcs/month, a utilization factor of 95% and percentage setup of 5%, is 953 pcs/month.

Applicable Customer Required Lead time (CRL) values are estimated in a preliminary simulation study. Summing up the average production lead times for each processing stage delivers a basic lead time value. The average CRL for our simulation study is determined by multiplying this basic lead time value with a CRL factor of 3. To model stochastic effects in CRL a log-normal distribution with a coefficient of variation of 0.5 is applied. In our simulation model, all customer orders are accepted. Due to an average utilization below 100% in the considered production system, short term overloads can be balanced in future periods or by covering customer orders with safety stocks.

Overall costs, consisting of holding and tardiness costs are selected as performance measure. The holding costs per piece and day are 1 CU for final products, 0.5 CU for semi-processed materials and the tardiness costs for final products are 19 CU per piece and day for the basic scenario. In the sensitivity analysis, tardiness costs of 9 and 99 CU per piece and day are investigated as well. In the simulation study, 5 years are simulated, where the first year is considered as the warm-up period and therefore excluded from the analysis. Due to the stochastic effects in demand and CRL, each iteration is evaluated with 10 replications.

C. Planning Parameters

Applied lot sizing rules, safety stock levels and planned lead times are important planning parameters for MRP [3]. In our simulation study, we choose Fixed Order Period (FOP) as lot sizing policy and the number of periods, for which the demand is accumulated into one production lot, as a planning parameter. To examine the influence of different safety stock levels, a safety stock factor is introduced as planning parameter. The actual safety stock is the initial value of safety stock multiplied with the safety stock factor. The fixed planned lead time of MRP is introduced as a factor, which is multiplied by the basic lead time values. These values are generated in the preliminary study, which is already used for setting customer required lead time values (see Section B. Scenario Definition). The initial value for safety stock of a product type is its average demand per day, i.e., a safety stock factor of 4 means that the average demand of 4 days is kept on safety stock.

For the safety stock relaxation methods, defined in Section II, a lower bound for the safety stock is introduced as an additional planning parameter. This minimum safety stock is again implemented as a factor that is multiplied with the applied safety stock. In order to get reasonable planning parameters for the safety stock relaxation methods, as well as for MRP, a grid search procedure is applied. Table IV shows the specified values for all planning parameters with respect to the different utilization factors.

TABLE IV. PARAMETER SETTINGS WITH RESPECT TO DIFFERENT UTILIZATION FACTORS

	Utilization Factor		
Parameter	90%	95%	98%
FOP periods	{1,2,3,4,5,6,8,10}	{4,5,6,8,10,12,14,16}	{4,6,8,10,12,14,16}
Safety stock factor	{0,1,2,4,6,8}	{0,1,2,4,6,8,16}	{0,1,2,4,6,8,16}
Planned lead time factor	{0,0.5,1,1.5,2}	{0,0.5,1,1.5,2,2.7}	{0,0.5,1,1.5,2,2.7,3.4}
Minimum safety stock factor	{0, 0.25, 0.5, 0.75}	{0, 0.25, 0.5, 0.75}	{0, 0.25, 0.5, 0.75}

D. Improvement potential and best parameters for basic scenario

The basic scenario is defined as the setting with 95% utilization, always setup and tardiness costs of 19 CU. However, since the percentage setup leads to different production systems both 5% and 10% setup are included into this basic scenario. The optimized planning parameters are found by identifying the parameter combination that leads to minimum overall costs for each method of safety stock relaxation and for MRP. Table V shows the results for this basic scenario with 5% and 10% setup. For both settings, 5% and 10% setup, all methods for safety stock relaxation reduce the overall costs significantly.

For 5% setup, method 3 delivers the best result and leads to a cost improvement of 25% in comparison to MRP. In this 5% setup setting, the number of FOP periods and the planned lead time factor are similar for all methods, only the safety stock factor is higher for method 2 and 3.

TABLE V. OPTIMAL SETTINGS FOR UTILIZATION 95%

Setup		MRP	Method 1	Method 2	Method 3
5%	Minimum overall costs	10426.6	8319.2	8112.1	7595.7
	Relative Improvement	-	-18.1%	-20.2%	-25.3%
	FOP periods	6	5	5	5
	Safety stock factor	4	4	16	8
	Planned lead time factor	1.5	2	1.5	1.5
	Minimum safety stock factor	-	0	0	0.25
10%	Minimum overall costs	10163.6	8498.1	9528.0	9760.7
	Relative Improvement	-	-16.4%	-6.3%	-4.0%
	FOP periods	6	8	6	6
	Safety stock factor	6	4	8	8
	Planned lead time factor	1	1.5	1	1
	Minimum safety stock factor	-	0	0.5	0.5

In the cost minimum solution for 5% setup, the introduced minimum safety stock factor is only applied for method 3.

In the setting with 10% setup, method 1 leads to the best result. The selected parameters show that methods 2 and 3 demand for a higher safety stock and a minimum amount of this safety stock, which must not be used for relaxation. Again, FOP periods and planned lead time factors do not reveal major differences for the applied methods. An interesting result concerning the comparison of safety stock relaxation methods is that method 1, i.e., having less safety stock relaxation occurrences but recalculating these each MRP run, leads to similar cost reduction potentials independently of the setup times. However, methods 2 and 3, i.e., allowing the safety stock to be reduced more often, do not perform that well if setup times are high. This might be related to the fact that safety stock reduction sometimes implies a new production lot to refill the safety stock after finishing a lot with reduced safety stocks. The negative impact of this unintended behavior is higher if setup times are higher.

To understand the influence of the planning parameters on the inventory, tardiness and overall cost more in detail, the following section discusses respective effects.

IV. PLANNING PARAMETER EFFECTS FOR BASIC SCENARIO

In this section, the influence of the two MRP parameters FOP periods and safety stock factor is investigated in detail to create a comprehensive understanding of how the three introduced safety stock relaxation methods behave in comparison to MRP. The influence on the performance, as well as the interrelationship of these parameters, is analyzed. Note that this analysis is performed for the basic scenario with 95% utilization, always setup and tardiness costs of 19 CU. The effects of the other parameters are discussed in the scenario analysis in Section V.

A. The Influence of FOP Periods on Performance

The application of four different methods and two different percentages of setup lead to eight different cases in this basic scenario, which are examined separately. For each specified value of the number of FOP periods (see Table IV), we select the combination of the other planning parameters, which results in minimal overall costs. Additionally we show the amount of inventory and tardiness costs and the minimum cost from MRP in Fig. 4.

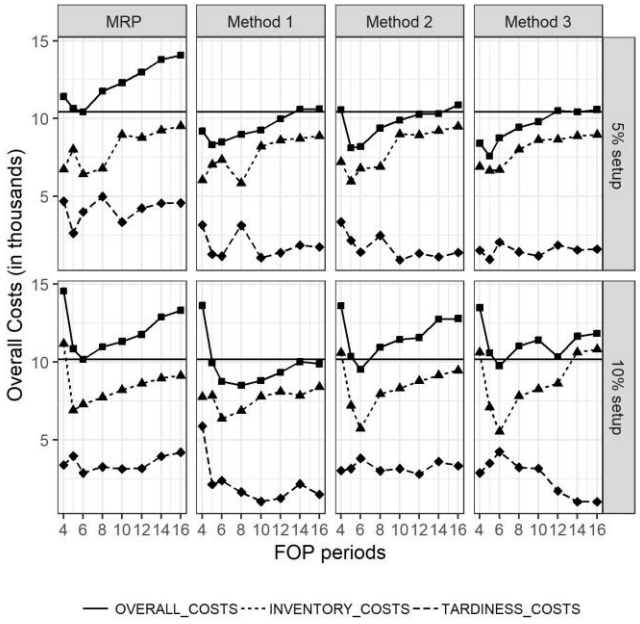


Figure 4. Influence of FOP periods on costs.

All cases show a more or less convex function for overall costs with respect to FOP periods with just a few outliers. As already mentioned in Section III, the optimal value for FOP periods are almost the same for all four methods. A low number of FOP periods leads to significantly higher overall costs in the 10% setup setting, whereas a higher number leads to a moderate increase in costs. The reason is that lower lot sizes lead to significantly higher setup times and, therefore, higher overall utilization in the 10% setup case in comparison to the 5% setup case. For all numbers of FOP periods, optimal inventory costs exceed optimal tardiness costs considerably. Apart from some outliers for small number of FOP periods, the inventory costs show a convex behavior with respect to the FOP periods. These results are in line with analytical production system findings without capacity balancing [7].

A detailed comparison of the optimal costs for safety stock relaxation methods with MRP shows that for the 5% setup setting, all safety stock relaxation methods lead to lower overall costs for a broad range of FOP lot sizes. This means that for lower setup times the negative effects of too high lot sizes can be mitigated by the safety stock relaxation methods. For 10% setup, only the safety stock relaxation method 1 leads to lower costs for a broad range of FOP lot sizes. This means that methods 2 and 3, which allow more frequent safety stock relaxation occurrences, are no more able to benefit from the capacity balancing if lot sizes become higher. This result fosters the finding from the previous section that these higher number of safety stock relaxation occurrences leads to some additional small production lots that reduce the overall performance.

B. The Influence of Safety Stock Factor on Performance

For the safety stock factor, the same analysis as for the FOP periods is performed and the results can be found in Figure 5. Note that the potential to apply safety stock relaxation for capacity balancing is linked to the amount of safety stock available. This subsection, therefore, identifies how much safety stock is needed for relaxation and how well additional safety stock is used by the methods.

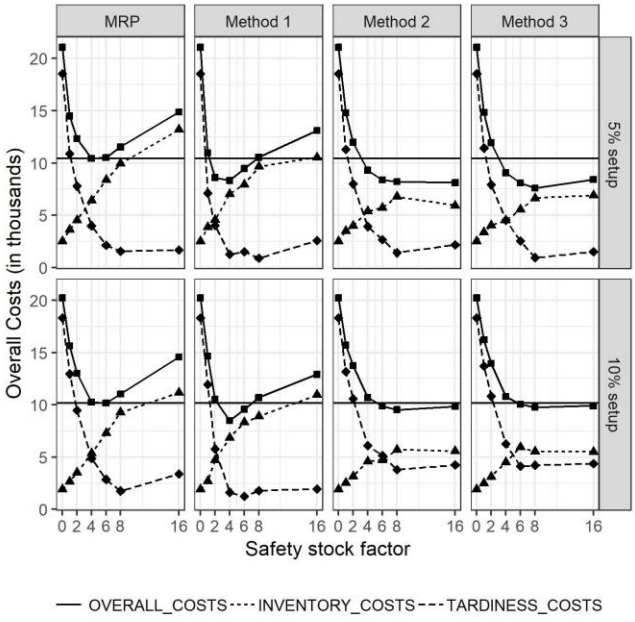


Figure 5. Influence of safety stock factor on costs.

The curves for overall costs show a clear convex shape with respect to safety stock factor, again with significant higher cost values for low safety stock values. For practical applications, this means that it is preferable to choose a higher safety stock when using safety stock relaxation, instead of selecting a safety stock that is too low. Small safety stock factors lead to high tardiness costs in comparison to inventory costs because the ability to balance capacity demands is limited. When safety stock is increased, also inventory costs increase and exceed the tardiness costs. The results show that method 1, with a lower number of safety stock relaxation occurrences, is much more sensitive on defining the right safety stock, similar to MRP. On the contrary, methods 2 and 3, i.e., more safety stock relaxation occurrences without/with memorizing this decision, can also benefit from higher safety stocks. Looking at the inventory costs shows that methods 2 and 3 also have lower inventory costs at higher safety stocks in comparison to method 1 and MRP. This implies that in methods 2 and 3 the average safety stock is lower which is intuitively clear since more safety stock relaxation occurrences are expected with these methods. Looking at the safety stock configurations for the relaxation methods that lead to lower costs than the optimal MRP setting shows that, contrary to the FOP influence, here methods 2 and 3 have a broader range of better parameters.

C. The Influence of FOP Periods on Safety Stock Factor

To explore the relationship between the parameters FOP periods and safety stock factor, for each value of FOP periods, the optimal safety stock factor is displayed in Fig. 6. This means, that for a fixed number of FOP periods, all other parameters are varied in the predefined grid (see Table IV) and the safety stock factor, which leads to the minimal overall costs is selected. Again, the 5% setup and 10% setup settings are shown for the basic scenario. The optimal parameter settings presented in Table V are marked by a star.

In general, a lower number of FOP periods, i.e., higher overall shop load due to setup times, leads to a higher optimal safety stock factor (apart from one outlier for method 2 at 5% setup). This shows that specifically for high shop congestion, the safety stock relaxation methods demand for more safety stock in order to balance capacity better. The result for method 3 in the 10% setup scenario is interesting and shows a further increase in safety stock for a high number of FOP periods. Note that in this scenario method 3 performs significantly worse than method 1 (see also Fig. 4). This implies that memorizing the safety stock reduction decision might in situations with high setup efforts and high lot covering ranges lead to system instabilities, which entail high safety stocks.

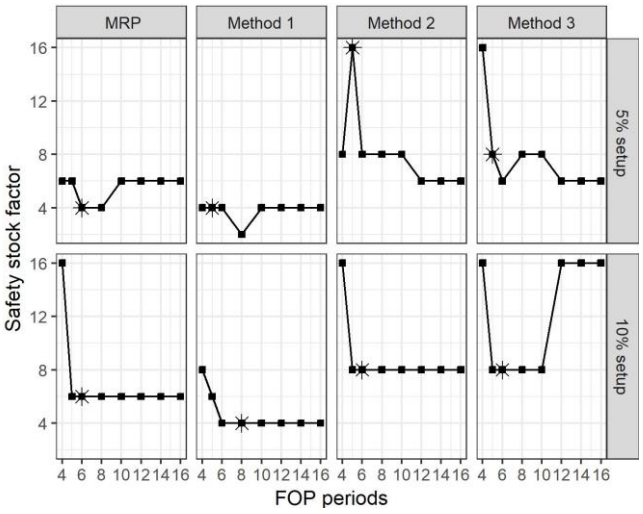


Figure 6. Influence of FOP periods on safety stock factor.

D. The Influence of Safety Stock Factor on FOP Periods

In this section we fix the safety stock factor and determine the number of FOP periods, which result in minimal overall costs. The results for all methods and scenarios are displayed in Fig. 7. In six of the eight cases, the number of FOP periods show a concave shape with respect to the safety stock factor. Only for methods 2 and 3 in the 10% settings there seems to be no influence of the safety stock on the optimal value of FOP periods. This is an interesting result since these are exactly the two scenarios where safety stock relaxation only leads to a rather small cost reduction potential (see Table V).

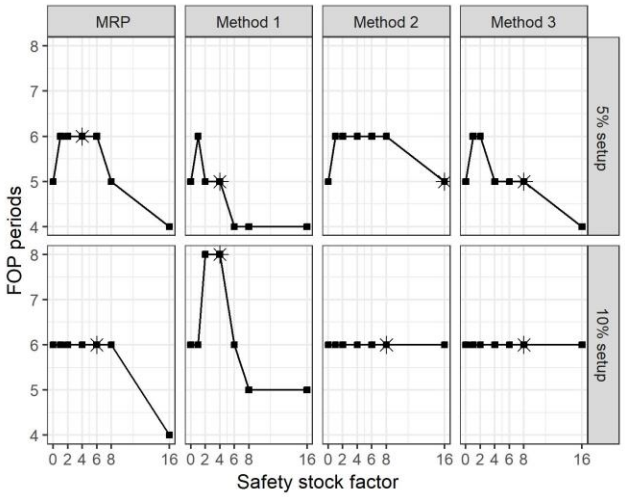


Figure 7. Influence of safety stock factor on FOP periods.

Low safety stock values lead to the situation that flexibility related to the customer demand can only be gained by lower production lot sizes. These situations still lead to high costs because no capacity load balancing is possible (see Fig. 5). For medium safety stock levels, a slight increase in lot size leads to a lower overall shop load (and capacity balancing by safety stock relaxation is already possible). This lower overall utilization combined with the capacity balancing leads for most cases also to the lowest overall costs. For very high safety stock factors, high inventory costs and low tardiness costs result, i.e., customer orders can always be fulfilled from the safety stock. Therefore, lower lot sizes (lower lot covering ranges) provide a possibility to slightly decrease the inventory costs.

The in depth discussion of the functionality of the different safety stock relaxation methods and the effects of different planning parameters on the performance of these methods indicates that these methods are promising for further research and practical application.

V. SAFETY STOCK RELAXATION COST PERFORMANCE FOR DIFFERENT SCENARIOS

After the in depth discussion of the functionality of the different safety stock relaxation methods in the last section, this section provides an analysis of the cost performance for a broader range of scenarios. Since three different methods for safety stock relaxation are presented in this paper, the current section shows which of them perform best in different cases and can be suggested for practical application and further research.

A. Tardiness Cost Effects

Since production systems face customers, which have different tardiness perceptions, this subsection investigates the methods performance for tardiness costs of 9 and 99 CU/day in comparison to 19 CU/day in the basic scenario. These values are selected, because based on simple inventory models and in combination with inventory costs of 1 CU/day, they correspond to a service level target of 90%, 95% and 99%. Table VI shows the results for tardiness cost of 9 CU/day and an intuitive result is that overall costs for all methods are lower than in the basic scenario. For 5% setup, the cost reduction potentials are similar to the basic scenario, however, method 1 and method 3 lead to nearly the same cost reduction potential. In this case the result for method 1 is interesting since it needs only very few safety stock. For 10% setup, the cost reduction potential is in this scenario significantly lower than in the basic scenario, specifically method 1 performs worse since here a 5.3% cost reduction can be realized in comparison to 16.4% in the basic scenario. In general the results for tardiness costs of 9 CU/day show a lower cost improvement potential with safety stock relaxation. Note that the 10% setting is the only one in the broader numerical study in which method 2 shows the best performance.

TABLE VI. OPTIMAL SETTINGS FOR UTILIZATION 95% WITH TARDINESS COSTS 9

Setup		MRP	Method 1	Method 2	Method 3
5%	Minimum overall costs	8,225.7	6,476.1	6,970.3	6,497.5
	Relative Improvement	0.0%	-21.3%	-15.3%	-21.0%
	FOP periods	6	5	5	5
	Safety stock factor	2	2	16	8
	Planned lead time factor	1.5	1.5	1.5	1.5
	Minimum safety stock factor	0	0	0	0
10%	Minimum overall costs	7,684.5	7,277.1	7,175.4	7,482.5
	Relative Improvement	0.0%	-5.3%	-6.6%	-2.6%
	FOP periods	6	6	6	6
	Safety stock factor	4	6	6	4
	Planned lead time factor	1	1	1	1
	Minimum safety stock factor	0	0.5	0.5	0.75

For higher tardiness costs, i.e., more impatient customers, Table VII shows the results and in general, a much higher cost reduction potential is observed. For 5% setup, method 3 performs best and for 10% setup method 1, which is consistent with the results of the basic scenario. However, method 2 shows a lower performance than in the basic scenario. A further interesting finding is that higher safety stocks are applied by most of the methods. This result is in line with simple analytical planning parameter optimization models without safety stock relaxation opportunity, which also show an increase in safety stock if tardiness costs increase [4]. The minimum safety stock factor is not significantly higher than in the basic scenario meaning that all safety stock is available for capacity balancing. Summarizing the general results for tardiness costs shows that higher tardiness costs lead to a better performance of safety stock relaxation. For practical application, this means that capacity balancing is more important if customers are more impatient or service level sensitive.

TABLE VII. OPTIMAL SETTINGS FOR UTILIZATION 95% WITH TARDINESS COSTS 99

Setup		MRP	Method 1	Method 2	Method 3
5%	Minimum overall costs	17,262.1	12,381.5	12,612.4	11,607.9
	Relative Improvement	0.0%	-28.3%	-26.9%	-32.8%
	FOP periods	6	5	12	5
	Safety stock factor	8	6	8	8
	Planned lead time factor	1.5	2	2.7	1.5
	Minimum safety stock factor	0	0.25	0	0.25
10%	Minimum overall costs	17,771.2	12,436.8	16,706.9	15,290.3
	Relative Improvement	0.0%	-30.0%	-6.0%	-14.0%
	FOP periods	5	10	6	16
	Safety stock factor	16	6	16	16
	Planned lead time factor	0.5	2	1	2.7
	Minimum safety stock factor	0	0	0.5	0

B. *Utilization Effects*

The comparison between 5% setup and 10% setup in the basic scenario already provides the insight that overall system utilization has a big impact on the performance of safety stock relaxation methods. In this subsection the influence of lower overall utilization, i.e., 90%, and higher overall utilization, i.e., 98%, is studied.

The results for lower utilization are shown in Table VIII and a general intuitive finding is that lower utilization leads to lower overall costs for all methods including MRP. For the 5% setup case, no safety stock is optimal for MRP and methods 1 and 2, i.e., these methods do not lead to a performance improvement. Method 3 leads to a performance improvement of 3.9%, which is far less than in the basic scenario. For 10% setup, method 3 leads to a cost reduction potential of 10.1%, which is higher than in the basic scenario. An interesting finding here is that for 10% setup method 1 and method 2 lead to higher costs than MRP. This means that for low utilization, the additional disturbances, which are caused by capacity balancing and relaxing safety stocks, have a higher negative influence on the overall performance than the positive effect of avoiding capacity shortages. The good performance of method 3 shows that especially in such lower utilization cases it is important to memorize the safety stock decisions to avoid additional disturbances. In general the result for 90% utilization shows that low production system utilization has only few needs for capacity balancing and only low improvement potential.

TABLE VIII. OPTIMAL SETTINGS FOR UTILIZATION 90%

Setup		MRP	Method 1	Method 2	Method 3
5%	Minimum overall costs	4,310.1	4,310.1	4,310.1	4,143.2
	Relative Improvement	0.0%	0.0%	0.0%	-3.9%
	FOP periods	3	3	3	3
	Safety stock factor	0	0	0	1
	Planned lead time factor	1.5	1.5	1.5	1.5
	Minimum safety stock factor	0	0	0	0
10%	Minimum overall costs	5,307.3	5,405.0	5,398.0	4,769.0
	Relative Improvement	0.0%	1.8%	1.7%	-10.1%
	FOP periods	6	6	6	4
	Safety stock factor	1	2	1	4
	Planned lead time factor	1.5	1.5	1.5	1.5
	Minimum safety stock factor	0	0	0	0

For high utilization cases, Table IX shows that capacity balancing has only a lower improvement potential and that all safety stock relaxation methods lead to similar results. High safety stocks are needed by all methods including MRP and a high minimum safety stock factors is needed for the different methods. The high utilization leads to a system that is near to instability, i.e., a lot of planning parameter combinations lead to a theoretical utilization above 100% and, therefore, to an instable system. An interesting finding is that all methods (including MRP) lead to optimal lot sizes, which are below the maximum lot size of FOP 16. Based on the production system and customer demand uncertainties, higher lot sizes imply that sometimes short term demands

occur for which additional production lots have to be issued. Hence, one explanation for this medium optimal lot size in this case is that this medium lot size provides a trade-off between too many setup operations based on low lot sizes and too many setup operations based on short term demands or safety stock refill orders.

In general, the results concerning utilization show that the best performance for capacity balancing can be gained for medium to high system utilizations. If the system utilization is too low, capacity balancing is not needed and if the system utilization is too high, there is only very few room for balancing the capacity, i.e., it is difficult for safety stock relaxation methods to refill the relaxed safety stock.

TABLE IX. OPTIMAL SETTINGS FOR UTILIZATION 98%

Setup		MRP	Method 1	Method 2	Method 3
5%	Minimum overall costs	17,096.2	15,904.4	15,915.0	15,975.5
	Relative Improvement	0.0%	-7.0%	-6.9%	-6.6%
	FOP periods	6	6	6	6
	Safety stock factor	16	16	16	16
	Planned lead time factor	0	0.5	1	1
	Minimum safety stock factor	0	0.75	0.75	0.75
10%	Minimum overall costs	15,538.6	14,316.1	14,365.0	14,319.6
	Relative Improvement	0.0%	-7.9%	-7.6%	-7.8%
	FOP periods	8	6	6	6
	Safety stock factor	8	16	16	16
	Planned lead time factor	1	0.5	0.5	0.5
	Minimum safety stock factor	-	0.75	0.75	0.75

C. Setup Effects

In the scenarios above, each new production order leads to a setup at the respective machine. This setting is chosen since the production system is a simplified setting with a low number of materials in comparison to real world systems. However, the results from Section IV and the two subsections above indicate that methods 2 and 3 might lead for some specific planning parameter combinations to small production lots from refilling the safety stock and less efficient capacity balancing. Therefore, this last investigation allows production lots to be put together and be produced without setup. In detail, a setting where setup occurs only if the material changes is studied. If there are more production orders waiting at a machine, orders with the same material as produced last are preferred as long as no other order has an earlier due date. This implementation mimics the behavior of workers who try to minimize their setup effort at the machine. Note that for this simplified production system structure this leads to less setup operations needed since only few materials are produced at each machine and the respective positive effects might be overestimated.

Table X shows that for all safety stock relaxation methods as well as for MRP, this setting leads to lower overall costs. Furthermore, method 3 has the highest cost reduction potential in this scenario for 5% and 10% setup. This means that being able to add smaller production lots that just refill the safety stock to other ones that already exist, significantly improves the performance of method 3. Looking at the optimal planning parameters, shows that lower safety stocks, lower production lot sizes and slightly higher planned lead times are optimal in comparison to the basic setting. An intuitive result is that the cost improvement potential improves in comparison to the basic scenario for 10% setup since there the setup operations show the highest influence.

TABLE X. OPTIMAL SETTINGS FOR UTILIZATION 95% WITH SETUP MATERIALCHANGE

Setup		MRP	Method 1	Method 2	Method 3
5%	Minimum overall costs	7,755.5	6,832.8	7,019.8	6,305.8
	Relative Improvement	0.0%	-11.9%	-9.5%	-18.7%
	FOP periods	6	4	6	4
	Safety stock factor	2	8	4	8
	Planned lead time factor	2	1.5	2	1.5
	Minimum safety stock factor	-	0	0	0
10%	Minimum overall costs	8,225.1	7,107.9	7,199.2	6,949.9
	Relative Improvement	0.0%	-13.6%	-12.5%	-15.5%
	FOP periods	8	6	6	6
	Safety stock factor	2	8	6	6
	Planned lead time factor	1.5	1.5	1.5	1.5
	Minimum safety stock factor	-	0	0	0.25

D. Overall performance comparison

Overall, 12 scenarios have been tested and the planning parameters for four methods, i.e., MRP and safety stock relaxation methods 1 to 3, have been optimized by search space enumeration. This broad numerical study shows that in all scenarios, the safety stock relaxation for capacity balancing leads to a considerable cost reduction potential. A managerial insight is, therefore, that using safety stock to balance capacity should be considered for improving production planning performance and rather simple heuristics already perform well.

Comparing the different safety stock relaxation methods shows that method 2, i.e., having more safety stock relaxation occurrences but not memorizing them, shows the worst performance and is only in 1 of the 12 scenarios the best option. Interestingly, methods 1 and 3 show in general a similar performance, i.e., method 1 leads in 5 scenarios to the best result and method 3 in 6 scenarios. Also concerning the average improvement potential, method 1 leads to an average cost improvement of 13.2% and method 3 to 13.5%. However, their performance in different scenarios differs significantly. For example, method 1 shows a significantly better performance for 10% setup and the basic scenario as well as for tardiness costs 99 CU/day. However, method 3 shows a significantly better performance at utilization 90%, where method 1 shows even a slight cost increase for 10% setup. From a managerial perspective, this means that both methods perform well but it depends on the specific production system structure, which one might be better to apply. For further research this means that both methods have potential to be further investigated and their sensitivity to planning interactions has to be further studied.

VI. CONCLUSIONS

In this article, three methods for temporary relaxing safety stock as an extension to traditional MRP are investigated. Since MRP neglects capacity constraints, heuristics for balancing capacity demand can improve the performance of the production system. The results of the simulation study show that all methods for safety stock relaxation lead to significant improvement in overall costs in comparison to MRP. For a broad range of numerical scenarios, the relative cost improvement potential of the best respective safety stock relaxation method is between 4% and 33%. Concerning planning parameter effects, one finding with practical relevance is that a higher safety stock is advantageous when relaxing safety stock, because there is only a small increase in inventory costs while decreasing tardiness costs due to capacity balancing. Opposite to this, a safety stock, which is too low, leads to considerably higher overall costs. Also for production lot sizes, too low lot sizes have shown a significantly lower performance than too high lot sizes. With respect to tardiness costs, the results indicate that higher tardiness costs lead to a better performance of safety stock relaxation. Concerning utilization we find that most improvement potential is gained for medium to high utilization. However, a very high utilization leaves only little space for capacity balancing and, therefore, a lower improvement potential is reported.

The performance comparison of the three developed safety stock relaxation methods shows that method 1 and method 3 perform similar while method 2 shows the worst performance. Even though, the performance of method 1 (leading to fewer safety stock relaxation occurrences without memorizing them) and method 3 (implying more relaxation occurrences but memorizing them) are similar, they lead for different system settings to different results. For further research this implies that both methods could be applied and combined with further actions for capacity balancing.

Limitations of this study are the selected ranges for the planning parameters for the grid search, which cannot guarantee an optimal solution. Furthermore, the simulation study is applied to a simple manufacturing structure. In further research, the safety stock relaxation methods have to be tested in more complex production structures or real production systems to get better estimates for the improvement potential in real world manufacturing systems. The robustness of the solutions, with respect to changes in utilization or machine failure behavior, could also be investigated. Additionally, other methods for capacity load balancing, e.g., lotsize adaption or alternative routings could be implemented and the their performance could be compared to the safety stock relaxation.

ACKNOWLEDGMENT

This paper was partially funded by FFG Grant 858642 and the FH OÖ grant MCRP.

VII. REFERENCES

[1] S. Strasser, K. Altendorfer, A. Peirleitner, Sensitivity Analysis of Simulation Study Results on Safety Stock Relaxation in Material Requirement Planning, in: SIMUL 2018: The Tenth International Conference on Advances in System Simulation, Nice, France, Nice, France, 2018, pp. 23–28.

[2] J. Orlicky, Material requirements planning: The new way of life in production and inventory management, McGraw-Hill, New York, 1975.

[3] W.J. Hopp, M.L. Spearman, Factory Physics, 3rd ed., Mc Graw Hill / Irwin, Boston, 2008.

[4] K. Altendorfer, Effect of limited capacity on optimal planning parameters for a multi-item production system with setup times and advance demand information, International Journal of Production Research 57 (6) (2019) 1892–1913. https://doi.org/10.1080/00207543.2018.1511925.

[5] T. Rossi, M. Pero, A simulation-based finite capacity MRP procedure not depending on lead time estimation, International Journal of Operational Research 11 (3) (2011) 237. https://doi.org/10.1504/IJOR.2011.041343.

[6] L. Sun, S.S. Heragu, L. Chen, M.L. Spearman, Comparing dynamic risk-based scheduling methods with MRP via simulation, International Journal of Production Research 50 (4) (2012) 921–937. https://doi.org/10.1080/00207543.2011.556152.

[7] K. Altendorfer, Influence of lot size and planned lead time on service level and inventory for a single-stage production system with advance demand information and random required lead times, International Journal of Production Economics 170 (2015) 478–488. https://doi.org/10.1016/j.ijpe.2015.07.030.

[8] M. Taal, J.C. Wortmann, Integrating MRP and finite capacity planning, Production Planning & Control 8 (3)

(1997) 245–254. https://doi.org/10.1080/095372897235307.

[9] P.C. Pandey, P. Yenradee, S. Archariyapruek, A finite capacity material requirements planning system, Production Planning & Control 11 (2) (2000) 113–121. https://doi.org/10.1080/095372800232315.

[10] N.A. Bakke, R. Hellberg, The challenges of capacity planning, International Journal of Production Economics 30-31 (1993) 243–264. https://doi.org/10.1016/0925-5273(93)90096-4.

[11] T. Rossi, R. Pozzi, M. Pero, R. Cigolini, Improving production planning through finite-capacity MRP, International Journal of Production Research 55 (2) (2017) 377–391. https://doi.org/10.1080/00207543.2016.1177235.

[12] T. Wuttipornpun, P. Yenradee, Development of finite capacity material requirement planning system for assembly operations, Production Planning & Control 15 (5) (2004) 534–549. https://doi.org/10.1080/09537280412331270797.

[13] A.M. Ornek, O. Cengiz, Capacitated lot sizing with alternative routings and overtime decisions, International Journal of Production Research 44 (24) (2006) 5363–5389. https://doi.org/10.1080/00207540600600106.

[14] A.R. Clark, Optimization approximations for capacity constrained material requirements planning, International Journal of Production Economics 84 (2) (2003) 115–131. https://doi.org/10.1016/S0925-5273(02)00400-0.

[15] J. Maes, L.N. van Wassenhove, Capacitated dynamic lotsizing heuristics for serial systems, International Journal of Production Research 29 (6) (1991). https://doi.org/10.1080/00207549108930130.

[16] L. Özdamar, T. Yazgac, Capacity driven due date settings in make-to-order production systems, International Journal of Production Economics 49 (1) (1997) 29–44.

[17] P.J. Billington, J.O. McClain, L.J. Thomas, Heuristics for Multilevel Lot-Sizing with a Bottleneck, Management Science 32 (8) (1986) 989–1006. https://doi.org/10.1287/mnsc.32.8.989.

[18] D.L. Woodruff, S. Voß, A Model for Multi-Stage Production Planning with Load Dependent Lead Times, System Sciences (2004) 88–96. https://doi.org/10.1109/HICSS.2004.1265247.

[19] H. Jodlbauer, S. Reitner, Material and capacity requirements planning with dynamic lead times, International Journal of Production Research 50 (16) (2012) 4477–4492. https://doi.org/10.1080/00207543.2011.603707.

[20] J.J. Kanet, M. Stößlein, Integrating production planning and control: Towards a simple model for Capacitated ERP, Production Planning & Control 21 (3) (2010) 286–300. https://doi.org/10.1080/09537280903363209.

[21] S. Strasser, A. Peirleitner, K. Altendorfer, C. Jenewein, H. Jodlbauer, The effect of safety stock relaxation and dynamic planned lead time within the MCRP algorithm in a simple manufacturing structure, unpublished.

[22] T. Felberbauer, K. Altendorfer, Comparing the performance of two different customer order behaviors within the hierarchical production planning, in: Proceedings of the Winter Simulation Conference 2014, Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, 2014, pp. 2227–2238.

[23] K. Altendorfer, T. Felberbauer, H. Jodlbauer, Effects of Forecast Errors on Optimal Utilization in Aggregate Production Planning with Stochastic Customer Demand, International Journal of Production Research 54 (12) (2016) 3718–3735. https://doi.org/10.1080/00207543.2016.1162918.

[24] T. Felberbauer, K. Altendorfer, A. Hübl, Using a scalable simulation model to evaluate the performance of production system segmentation in a combined MRP and kanban system, in: Proceedings of the Winter Simulation Conference 2012, Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, 2012, pp. 1–12.

A Verification Framework for Business Rules Management in the Dutch Government Context

Koen Smit
Digital Smart Services
HU University of Applied Sciences Utrecht
Utrecht, the Netherlands
koen.smit@hu.nl

Martijn Zoet
Optimizing Knowledge-Intensive Business Processes
Zuyd University of Applied Sciences
Sittard, the Netherlands
martijn.zoet@zuyd.nl

Matthijs Berkhout
Digital Smart Services
HU University of Applied Sciences Utrecht
Utrecht, the Netherlands
matthijs.berkhout@hu.nl

***Abstract*—Since an increasing amount of business decision/logic management solutions are utilized, organizations search for guidance to design such solutions. An important aspect of such a solution is the ability to guard the quality of the specified or modified business decisions and underlying business logic to ensure logical soundness. This particular capability is referred to as verification. As an increasing amount of organizations adopt the new Decision Management and Notation (DMN) standard, introduced in September 2015, it is essential that organizations are able to guard the logical soundness of their business decisions and business logic with the help of certain verification capabilities. However, the current knowledge base regarding verification as a capability is not yet researched in relation to the new DMN standard. In this paper, we re-address and - present our earlier work on the identification of 28 verification capabilities applied by the Dutch government [1]. Yet, we extended the previous research with more detailed descriptions of the related literature, findings, and results, which provide a grounded basis from which further, empirical, research on verification capabilities with regards to business decisions and business logic can be explored.**

Keywords-Business Rules; Business Rules Management; Verification; Capabilities; Dutch Government

I. INTRODUCTION

Business rules (BR's), as part of business logic, are increasingly being utilized in enterprises as building blocks for (automated) decision making, for example, supporting execution of various types of e-services like applying for an insurance product and applying for social benefits and automated fraud detection at financial organisations. As a result, organizations employ various methods and processes to manage these BR's, often referred to as Business Rules Management (BRM) [2]. An important part of BRM comprises quality control, which focuses on reducing errors in the syntax and intended behavior of the business rules. Thereby improving the quality of the defined and executed BR's [3]. This particular capability is referred to as verification. A capability in this paper is defined as an ability that an organization, person, or system, possesses. It, therefore, defines what an organization, person or system does or can do but not how it accomplishes it. In practice, a capability can be implemented in different ways, for example manually, partly- or fully automated.

In September 2015, the Object Management Group (OMG) released a new standard for modelling decisions and underlying business logic, the Decision Model and Notation (DMN) [4]. As the adoption of the DMN standard increases, the need for verification of BR's, which are a significant component of the decision logic layer in DMN, increases as well. Therefore, in this paper, we adhere to the DMN 1.1 standard [5] and aim to explore which verification capabilities are relevant in the verification process of decisions and underlying BR's.

Verification, as a capability in general software development, is an established research field and has received a lot of attention from researchers in the previous decades (Ackerman, Buchwald, & Lewski, 1989; Van der Aalst, 1999; Vermesan & Coenen, 2013). In literature, verification of business rules is a capability, executed by a specific component, of expert systems, knowledge management systems, knowledge engineering systems, or knowledge based systems. Regarding these research domains, different scholars and practitioners identify different types of verification capabilities, for example, the work [9] on verification capabilities for expert systems, in the work of [10], [11] and [8] on verification capabilities for business process models, and in the work of [12] and [7] on verification capabilities for Knowledge Based Systems. Another contribution within the research domain of business logic is the work of Von Halle and Goldberg [13], which presents multiple principles that refer to capabilities that are applicable when performing verification of business logic, containing business rules. An example included in their work is the normalization and integrity verification capability.

However, in current literature on business logic, no uniform overview exists. Additionally, the current knowledge base predominantly focuses on theory forming by means of deductive research methods, while inductive

research methods to explore the spectrum of the verification capability seem almost non-existent to the knowledge of the authors.

This paper aims to define, from practice, the spectrum of capabilities required for the verification of business logic which can be designed and specified with DMN. To be able to do so, we addressed the following research question: "*Which verification capabilities are useful to take into account when designing a business rules management solution?*" Five large Dutch government institutions participated in a three-round focus group to derive verification capabilities applied in practice. The results form a framework of capabilities regarding the verification of business rules.

The remainder of this paper proceeds as follows. First, we provide, in short, insights into what verification comprises in the context of BRM and how it relates to the other BRM capabilities. This is followed by the research method utilized to identify the verification capabilities applied in the Dutch governmental context. Furthermore, the collection and analysis of our research data are described. Subsequently, our results, which led to our framework containing 28 verification capabilities are presented. Finally, we discuss which conclusions can be drawn from our results, followed by a critical view of the research method and techniques utilized and propose possible directions for future results.

II. Background and Related Work

With increasing investments in BRM, organizations are searching for ways to guide the design of business rules management solutions. A business rule is defined as "*a statement that defines or constrains some aspect of the business intending to assert business structure or to control the behavior of the business*" [14]. A business rules management solution enables organizations to elicitate, design, specify, verify, validate, deploy, execute, evaluate and govern business rules, see Figure 1 [15], [16]. When a business rules management solution is designed, each of the nine previously mentioned capabilities need to be designed, implemented and governed. The manner in which way these capabilities are realized depends on the actual situation in a specific organization. This paper is part of a research project in which the focus was to evaluate all nine capabilities of five government institutions. In this paper, we focus on the verification capability as other studies (i.e., [17]–[20]) already focused on the exploration and definition of the other BRM capabilities.

As stated in the introduction section, no uniform overview exists with regards to verification capabilities in the context of BRM, and more specific, in relation to DMN. Literature in neighboring fields often define one or more verification capabilities, however, they do not present a uniform overview. Furthermore, the verification capabilities described in neighboring fields are often based on or related to a specific language and therefore less generalizable. For example, regarding software development verification, [21] and [6] describe several verification capabilities, but do not aim to be complete as their work define the boundaries of verification in general and a process to execute verification. Furthermore, for example, with regards to Business Process Management and process modeling. The work of [22] and [23], describe verification as a capability for process model checking. However, they do so in a technical and non-uniform manner. From the literature we find that verification capabilities, in a general sense, are often mentioned or described as part, thus often a sub-goal, of a research study, to evaluate the conformance with certain guidelines. To contribute to the current knowledge base, we aim to define the verification capability with regards to BRM utilizing an inductive approach.

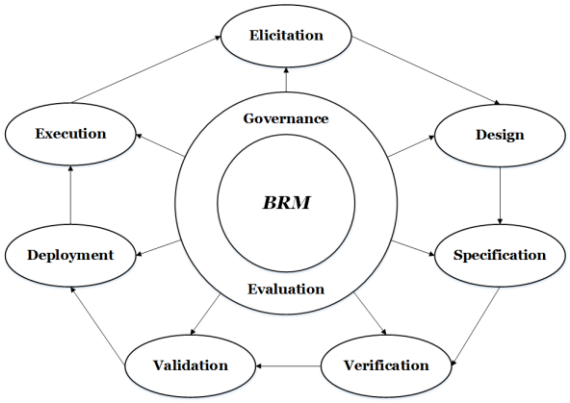


Figure 1. BRM Capability Framework

A detailed explanation of each capability can be found in [24]. However, to ground our research, a summary of the elicitation, design, specification and verification capabilities is provided here. The purpose of the elicitation capability is twofold. On the one hand, the purpose is to determine the knowledge that needs to be captured from various legal sources to realize the value proposition of the business rules. Many possible legal sources from which this knowledge can be derived exist, for example, laws, policies, guidelines, regulations, expert hearings, research outcomes, case law, and internal documentation. Depending on the type of knowledge source(s), different methods, processes, techniques and tools to extract the knowledge are applied [25]. The output of the elicitation capability is the collection of knowledge that is required to design the Decision Requirements Diagram (DRD), which is the highest level of abstraction with respect to decision modelling in DMN. The DRD layer recognizes four concepts: 1) a decision, 2) business knowledge, 3) input data, and 4) a knowledge source. When no DRD exists, elicitation information is collected to specify the four. On the other hand, when a DRD is already in place, an impact analysis is performed to identify the modifications that need to be processed to the existing structure and underlying business logic in the design and specification capabilities. The DRD consists of a combination of business decisions. A DRD is a collection of business logic, in terms of business rules and fact types. The relationship between different decisions is depicted in a derivation structure. The DRD is the high-level output which

the design capability needs to realize. After the DRD is created, the content (business rules and fact types) of each individual decision need to be specified in the specification capability. The purpose of the specification capability is to create the business rules and fact types needed to make the decision, the Decision Logic Level (DLL). The decision logic level has multiple key concepts which are described in two languages the Friendly Enough Expression Language (FEEL) and the Simple FEEL variant (SFEEL). SFEEL is a subset of FEEL, tailored for simple expressions in conjunction to be utilized in decision tables. However, the same concepts of SFEEL and FEEL can be expressed in multiple other languages. For example, Camunda also supports the use of other languages to define business logic with, such as Javascript, Groovy, Python, Jruby, and Juel [26]. The language selected to represent the decision logic does not influence the decision requirements level. The output of the specification capability is a specified context design that contains decisions, business rules, and fact types. After the DRD and DLL are created, they are verified, comprising the evaluation to eliminate syntax errors in both abstraction levels. This is defined as the verification capability. The purpose of the verification capability is to determine if the decisions, business rules, and fact types adhere to predefined criteria, for example, conformance to language guidelines, and are logically consistent. If errors are identified, two scenarios can occur. First, the verification issues are resolved in a revision of the designed and specified business knowledge. Second, the verification issues are ignored and the decisions, business rules, and fact types are deployed based on the current elicitated, designed and specified business logic. However, verification errors not properly addressed could result in the improper execution of the value proposition in the execution capability later on in the BRM processes, thus posing a possible risk to the organization that employs the business logic [24].

III. RESEARCH METHOD

The goal of this research is to identify verification capabilities that are being utilized in practice. Currently, research is conducted on business rules (management), however, the existing knowledge base is rather old and mostly from a theoretical perspective [15], [24], [27]. Additionally, most of the research that is conducted on business rules (management) embraces a deductive approach, while little is known on how verification is applied in practice. Therefore, research studies utilizing an inductive approach could lead to further theory refinement by taking into account verification capabilities are applied in practice. An appropriate focus of research with an inductive approach is on identifying new constructs and establishing relationships between identified constructs from a practical context (e.g., Edmondson and McManus, [28]). Therefore, through grounded theory based data collection and analysis, in our research we explore verification capabilities applied in practice and combine them into a framework to, on the one hand, guide organizations in the design and execution of the verification capability as part of business rules management, while on the other hand strengthen the currently available knowledge base with insights derived from practice.

To explore a range of possible solutions with regards to a complex issue group based research techniques are adequate [29]. Group-based research techniques are useful when the collection of possible solutions need to be combined into one view, backed by empirical evidence that is not present in the body of knowledge. Examples of group based techniques are brainstorming, nominal group techniques, focus groups and Delphi studies. Group based research techniques can be differentiated by the type of approach they utilize, face-to-face versus non-face-to-face approaches to gather research data. Of course, both the face-to-face and non-face-to-face approaches are characterized by their advantages and disadvantages; i.e., in face-to-face meetings, participants can provide (additional) feedback directly. On the other hand, face-to-face meetings are characterized to be restricted with regard to the number of participants as well as the possible existence of group or peer pressure.

For this study we selected a face-to-face approach to be more appropriate, also facilitating peer-discussion regarding the application of the verification capability at the selected governmental organizations. Earlier experiences of the researchers with similar approaches showed that participants will trigger each other to elaborate more in-depth on why and how a specific capability is applied.

IV. DATA COLLECTION AND ANALYSIS

The data for this study is collected over a period of three months, between January 2014 and March 2014. The collected data has initially been categorized based on the beta version of the DMN standard that was published in August 2013. Since no significant changes between the beta and the final version of the DMN standard occurred, we refer to the final 2016 version of the DMN standard in this paper [5]. The data collection was conducted through three rounds of focus groups. Between each individual focus group, the researchers consolidated the results.

When designing a focus group, a number of situational characteristics need to be considered: 1) the goal of the focus group, 2) the selection of representative participants, 3) the number of participants, 4) the selection of the main facilitator and research team, 5) the information recording facilities, and 6) the protocol of the focus group. The goal of the focus group was to identify the current verification capabilities being applied in practice. The selection of participants should be based on the group of individuals, organizations, information technology or community that best represents the phenomenon studied [30]. In this study, organizations and individuals that deal with the verification of a large amount of business rules represent the phenomenon studied; examples are financial and governmental institutions. Taking this into account, multiple Dutch government institutions were invited to participate. The organizations that agreed to cooperate with the focus group meetings were the: 1) Dutch Tax and Customs Administration, 2) Dutch Immigration and Naturalization Service, 3) Dutch Employee Insurance Agency, 4) Dutch Education Executive Agency, and 5)

Dutch Social Security Office. We believe that this is a representative selection due to several reasons; 1) they apply all degrees of automation to their decision making (i.e., human, a human supported by a machine, a machine supported by a human, and fully automated), 2) they design and execute a large variety of rule types (i.e., derivation, calculation, constraints, process, validation, and decision rules), and 3) they are required to indisputably implement the laws and regulations for all Dutch citizens and organizations.

Based on the written description of the goal and consultation with experienced employees of each government institution, participants were selected to take part in the focus group meetings regarding verification of business rules. In total, ten participants took part in the focus group rounds which fulfilled the following positions: One legal advisor, two BRM project managers and seven business rule analysts. All involved subject-matter experts had a minimum of five years of experience with the verification of business rules. Delbecq & van de Ven [29] and Glaser [31] state that the facilitator of the focus groups should have an appropriate level of experience with regards to the topic. Also the facilitator should have experience with the workings of face-to-face group based research techniques. The facilitator in this research project has a Ph.D. in BRM and has conducted nine years of research with regards to BRM. Furthermore, the facilitator has conducted research while utilizing many face-to-face research techniques before. Additionally, three researchers were supporting the facilitator during the focus group meetings. One researcher was the 'back-up' facilitator. The back-up facilitator monitored whether each participant provided equal input. When necessary, the back-up facilitator involved specific participants by asking for more in-depth elaboration on the subject at hand. The other two researchers acted as minute's secretary, taking notes. All focus group meetings were video and audio recorded. The duration of the first focus group was 192 minutes, the second 205 minutes and the third 207 minutes. All three focus group meetings followed the same overall protocol, starting with an introduction and explanation of the purpose and procedures of the focus group at hand, after which verification capabilities were generated, shared, discussed and/or refined.

Prior to the first round, the research team informed the participants with regards to the purpose of the research and meetings at hand, after which the participants were invited to submit their current documentation with regards to verification capabilities regarding business decisions and business logic. Prior to the first focus group meeting, the research team already consolidated similar verification capabilities that were derived from the received documentation. This was to ground and start up the discussion of the first focus group meeting. During the first focus group meeting, participants first explained their submitted documentation and why their verification capabilities were relevant in their context. For each capability, the group discussed whether it was related to business rules management processes in general or not, for example, some of the mentioned results focused more on the verification of process models or data types. The second and last part of the focus group meeting was committed to defining new or missing capabilities where participants thought they were missing from the already identified selection of capabilities. For each proposed capability, it's ID, label, description, rationale, classification, and example(s) were discussed and noted, see Table I.

Table I. Identical business rules verification

capability ID:	B4.
capability label:	Identical business rules verification.
capability description:	Identical business rule verification checks if a business rule occurs more than once in the exact same appearance in the same business rule set.
capability rationale:	Identical business rule verification is needed as redundant rules account for extra maintenance burden on top of the negative impact they have on performance.
capability classification:	Decision logic level verification
capability example: (underlined business rules are identical)	*Decision: Rights for Child Benefits* *1 – The Age of the Child is between 16 and 17* *2 – The Child is registered as part of => 1 household* *3 – The Child has the right to receive study benefits* *4 – The Child is registered as part of => 1 household* 5 – *The Registration Status of the Child is Household of 1*

When the first focus group meeting was finished, the researchers started analysis to consolidate the results. Consolidation of the results comprised the detection of incomplete and redundant capabilities. Next, the results of the analysis by the research team were sent to the participants of the focus group meeting fourteen days in advance before the next meeting. During these fourteen days, the participants assessed the consolidated results in relationship to three questions: 1) "*Are all capabilities described correctly*?" (in terms of the capability label and accompanied examples), 2) "*Do I want to remove a capability*?", and 3) "*Do we need additional capabilities*?"

During the second focus group, the participants discussed the derived capabilities. The group started to discuss their usefulness, and, again, whether all capabilities were described correctly. Furthermore, the participants were asked to validate whether the capabilities that were identified to be redundant from the consolidation by the research team needed removal from the selection of relevant capabilities. For example, one of the participants submitted the capability 'illegal value', while another capability labelled 'domain violation' already existed in the results of the first focus group round, which is an equivalent capability. As the end of

the second focus group meeting showed signs of saturation amongst the participants. For this reason, the third focus group was redesigned as a validation session in which we solely wanted to validate the results that were derived from the first two focus groups. The discussion in the last focus group therefore focused on further refinement of the existing capabilities in terms of their descriptions, classification, and goals.

V. RESULTS

In this section, based on our data collection and analysis, we present our results with the help of a case which is further specified in [32]. This case does not adhere to the business rules provided by the governmental agencies. The reason for this is that per discussed capability, different case examples were adhered to in the focus groups by the participants. The reader therefore would get snippets of business rules from each agency, lacking the discussion of an integrated case. Therefore the results of the focus groups and Delphi study have been mapped onto an integrated case that is presented in an integrated manner. The case we selected comprises the determination of risk of malnutrition regarding a hospitalized patient, see Appendix A.

Table II. Business logic expressed in the DMN formal decision table format

Decision: Determine Malnutrition Risk			
Business rule #	**Input**	**Output**	**Annotation**
	Malnutrition Risk Points of the patient	Malnutrition Risk of the patient	
1	<= 3	1	Malnutrition of the patient is 1 (low)
2	]3..6[	2	Malnutrition of the patient is 2 (moderate)
3	>= 6	3	Malnutrition of the patient is 3 (high)

As stated in the background and related work section of this paper, the DMN standard features two levels of abstraction, the DRD and the DLL. For the demonstration of the business logic used in our examples we choose not to use, due to page size constraints, the DMN formal decision table format, but a simpler and compact representation; structured English. Business logic expressed in decision tables is easily transformed into structured English business logic, see for example Table II (decision tables) and Table III (structured English).

In total, the consolidated framework for the verification of business decisions and business logic consists of 28 verification capabilities, see Figure 2. Due to space limitations, we present each capability by its label, description, and example.

Table III. Business logic expressed in the structured English format

Decision: Determine Malnutrition Risk
BR1 - Malnutrition Risk of the patient must be equated to 1 IF Malnutrition Risk Points of the patient <= 3
BR2 - Malnutrition Risk of the patient must be equated to 2 IF Malnutrition Risk Points of the patient is >3 AND <6
BR3 - Malnutrition Risk of the patient must be equated to 3 IF Malnutrition Risk Points of the patient >= 6

To further structure our derived capabilities, the abstraction layers of the DMN standard are utilized for categorization as some verification capabilities are only relevant in the context of a certain abstraction level of business logic. Lastly, as some derived verification capabilities are relevant in a generic sense, the generic category has been added.

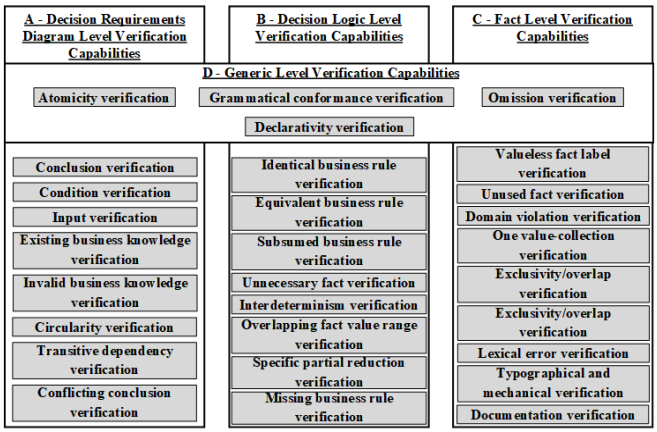


Figure 2. BRM verification capability framework.

A. *Decision requirements level verification*

Regarding the highest level of abstraction, the decision requirements level, eight verification capabilities were identified.

Conclusion verification

Conclusion verification checks if the conclusion fact of an individual decision is used as a condition fact in another decision. In a DRD, this situation can only legitimately occur once, namely with the top-level decision. Additional occurrences indicate an error in the logical completeness. For example, let's examine the decisions "E - Calculate Body Mass Index Risk Points" and "F - Calculate Body Mass Index". The conclusion fact "Body Mass Index" of decision F is used as a condition fact in decision E. If this would not be the case decision E would be floating and an error would occur.

Condition verification

Condition verification is a reversed form of conclusion verification. It first checks if a condition fact is a ground-fact or derived-fact. If a fact is a derived-fact, the test checks if the fact is the conclusion fact of another decision. Let's examine the same example as presented in the conclusion verification. The condition fact "Body Mass Index" of the decision "E - Calculate Body Mass Index Risk Points" is the conclusion of the decision "F - Calculate Body Mass Index". If this would not be the case, decision E would be executed, but an error would occur.

Input verification

Input verification checks if the conclusion fact of an underlying decision is used as a condition fact in the parent decision. Contrary to conclusion verification and condition verification, input verification checks if there are no unnecessary decisions in the DRD. For example, let's examine the decision: "C – Calculate Weight Loss Risk Points". For this example, this decision has two connected decisions: "D - Calculate Weight Loss" and additionally, "Assess Weight Pattern." However, only the conclusion fact of decision D is applied in decision C. Indicating that the decision "Assess Weight Pattern" is unnecessary and should be removed from the DRD.

Existing business knowledge verification

Existing business knowledge verification checks if a decision is accompanied with specified business knowledge. For example, let's examine the same decisions as used in the example of input verification: "D – Calculate Weight Loss" and "Assess Weight Pattern." In this specific instance, no business knowledge is specified for decision "Assess Weight Pattern", while the business knowledge is required to execute the parent decision "C – Calculate Weight Loss Risk Points."

Invalid business knowledge verification

Invalid business knowledge verification checks if each fact value of the condition fact of a decision is also present as a fact value of the linked conclusion fact of the underlying decision(s). For example, let's examine the two decisions: "E - Calculate Body Mass Index Risk Points" and "F - Calculate Body Mass Index." The conclusion fact of decision F is used as a condition fact in decision E. In this example, the conclusion fact of decision F contains a value range from 10 to 80 BMI points, while the condition fact in decision E contains a value range of 20 to 70 BMI points.

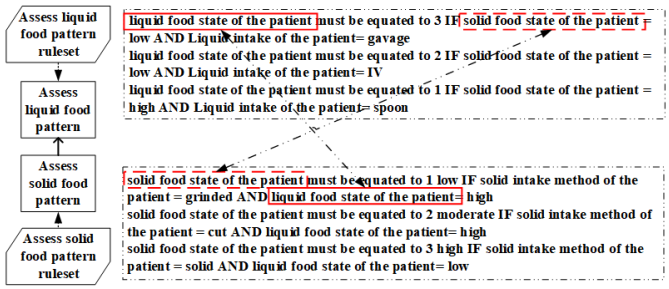


Figure 3. Example based on [32] depicting a circularity error

Circularity verification

Circularity verification checks if a conclusion fact of the parent decision is used as a condition fact in the underlying decision while at the same time, the conclusion fact of the underlying decision is used as a condition fact in the parent decision. For example, see Figure 3.

Transitive dependency verification

Transitive dependency verification checks if the same condition fact occurs twice in a set of three decisions that are connected to each other. For example, take into account a situation where there are three decisions: "Decision A", "Decision B" and "Decision C". "Decision A" applies the conclusion facts of "Decision B" and "Decision C" as condition facts. In addition "Decision B" applies the condition fact "intake solid food of the patient" to derive a conclusion. Additionally "Decision C" also applies the condition "intake solid food of the patient" in addition to an extra condition fact to derive a conclusion. Therefore the condition fact "intake solid food of the patient" is applied twice to eventually derive the conclusion of "Decision A". This does not necessarily have to be an error, but usually implies an error in the business logic, which can be solved by removing either "Decision B" or the condition fact "intake solid food of the patient" in "Decision C", see also Figure 4.

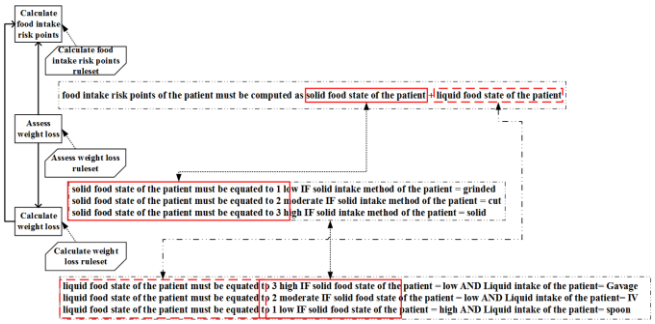


Figure 4. Example based on [32] depicting a transitive dependency error

Conflicting conclusion verification

Conflicting conclusion verification checks if there are conclusion facts that are established using different business rules and facts. For example, the conclusion fact "Food Intake Risk Points" of decision G could be calculated by the following business rules: "BR18 - Food Intake Risk Points of the patient must be equated to 4 IF Solid Intake of the patient > 5 days AND Age of the patient > 18 AND Liquid Intake of the patient <= 1 days" and "BR19 - Food Intake Risk Points of the patient must be equated to 4 IF Solid Intake of the patient > 5 days AND Age of the patient > 18 AND Liquid Intake of the patient <= 1 days ". In this case, the conclusion fact is differently calculated and therefore the results of both business rules could contradict each other.

B. Decision logic level verification

Regarding the decision logic level, eight verification capabilities were identified.

Identical business rules verification

Identical business rule verification checks if a business rule occurs more than once in the exact same appearance in the same business rule set. Take for example the following two business rules: "Malnutrition Risk of the patient must be equated to 2 IF Malnutrition Risk Points of the patient is >3 AND <6" and "Malnutrition Risk of the patient must be equated to 2 IF Malnutrition Risk Points of the patient is >3 AND <6". See also the completely elaborated capability description and rationale in Table I.

Equivalent business rules verification

Equivalent business rule verification checks for business rules which are expressed different, but have the same outcome. Take for example the following two business rules: "Food Intake Risk Points of the patient must be equated to 4 IF Solid Intake of the patient > 5 days AND Age of the patient >= 18 AND Liquid Intake of the patient > 1 days" and "Food Intake Risk Points of the patient must be equated to 4 IF Solid Intake of the patient >= 6 days AND Age of the patient > 17 AND Liquid Intake of the patient >= 2 days." Both business rules have the same outcome, but are written differently.

Subsumed business rules verification

Subsumed business rule verification checks if business rules exist that are more comprehensive compared to a business rule with the same outcome. Take for example the following business rule: "Food Intake Risk Points of the patient must be equated to 4 IF Solid Intake of the patient > 5 days AND Liquid Intake of the patient > 1 days". The previous business rule concludes exactly the same fact as the following business rule, which utilizes one more condition: "Food Intake Risk Points of the patient must be equated to 4 IF Solid Intake of the patient > 5 days AND Age of the patient <= 18 AND Liquid Intake of the patient > 1 days".

Unnecessary fact verification

Unnecessary fact verification checks for facts that is included in a business rule, but is not required to calculate or derive the outcome. Take for example the business rule: "Body Mass Index of the patient must be computed as Current Weight of the patient / Square (Height of the patient)". Additionally, we present the facts for this ruleset: 1) Body Mass Index, 2) Current Weight, 3) Height, and 4) Shoe Size. In this example the Shoe Size of a patient is not used in the business rule to calculate the outcome and is therefore unnecessary.

Interdeterminism verification

Interdeterminism verification checks if there are two business rules with the same condition facts but with a different conclusion. For example the business rules: "Malnutrition Risk of the patient must be equated to 2 IF Malnutrition Risk Points of the patient is >3 AND <6" and "Malnutrition Risk of the patient must be equated to 4 IF Malnutrition Risk Points of the patient is >3 AND <6". The conditions of these two business rules are exactly the same, but the conclusion fact value is different, which leads to conflicting results.

Overlapping fact value range verification

Overlapping fact value range verification checks if condition fact value ranges in a business rule overlap each other which may lead to inconsistent business rule output. For example the business rules: "Body Mass Index Risk Points of the patient must be equated to 1 IF Patient Body Mass Index of the patient is >18.5 AND <20" and "Body Mass Index Risk Points of the patient must be equated to 2 IF Body Mass Index of the patient is >19 AND <22." In this case, the ranges of the condition fact Body Mass index are 18.5-20 and 19-22, which consist of partly overlapping values.

Specific partial reduction verification

Specific partial reduction verification checks if two ranges in business rules can be combined. Take for example the following business rules: "Food Intake Risk Points of the patient must be equated to 6 IF Solid Intake of the patient > 5 days AND Age of the patient > 18 AND <= 30 AND Liquid Intake of the patient > 1 days" and "Food Intake Risk Points of the patient must be equated to 6 IF Solid Intake of the patient > 5 days AND Age of the patient > 30 AND <40 AND Liquid Intake of the patient > 1 days." The conclusion is the same in this example so the two business rules can be combined into one business rule.

Missing business rules verification

Missing business rules verification checks if there are situations in which a particular inference is required, but there is no rule that succeeds in that situation and produces the desired outcome. Missing business rules can be detected when it is possible to enumerate all possible scenarios in which a given decision should be made or a given action should be taken. For example, when the first business rule of the decision 'C - Calculate Weight Loss Risk Points' is missing, it is impossible to conclude a total amount of risk points other than the fact values '1' and '2'.

C. Fact level verification

Regarding the decision fact level, eight capabilities regarding verification were identified. In the verification capabilities above references are made to specific decisions that are part of the case that is utilized to demonstrate the capabilities. From this point on, this is not relevant as fact types and fact values can be and are usually independently managed and applied across different collections of business logic (like business rules or decision tables).

Valueless fact label verification

Valueless fact label verification focuses on the label of the fact in the fact vocabulary. It checks whether each fact type label is expressed without any fact values. For example, the fact type "Body Mass Index >=20" is a binary question, thus only comprising two fact values, "yes" and "no". In this case the fact value "20" should be removed from the fact label and formulated in two separate rules.

Unused fact verification

Unused fact verification focuses on facts that are present in the fact vocabulary, but not utilized in any BR. Unused facts, especially at large amounts, can decrease efficiency as these unused facts need to be maintained just like the facts that are utilized in BR's. Such errors are often caused by the removal of a BR without checking whether the facts are still utilized in other BR's. For example, the following facts and BR's are expressed: 1) "Body Mass Index", 2) "Body Mass Index Risk Points", and 3) "Standardized Body Mass Index", BR's: "Body Mass Index Risk Points of the patient must be equated to 0 IF Body Mass Index of the patient is >= 20" and "Body Mass Index Risk Points of the patient must be equated to 2 IF Patient Body Mass Index of the patient < 20". In this case three fact types are available, but the given BR set does not utilize the fact type "Standardized Body Mass Index".

Domain violation verification

Domain violation verification focuses on how fact values are expressed, in terms of its format, against how they should be expressed. This is important as it influences if the executability of the BR of which the fact types are part of. For example, consider a BR that utilizes the following fact: "Current Weight". The weight of the patient is usually expressed as an integer, for example "122". However, when the data type of the fact "Current Weight" is designed as a string the weight can only be expressed as: "one hundred and twenty-two kilograms".

One value-collection verification

One value-collection verification focuses on collections and the amount of fact values a fact contains. Less than two fact values in a collection can be caused by 1) not enough fact values are created in the specification process or 2) due to changes to laws and regulations that result in the removal of fact values as part of the collection. For example, the fact type "Malnutrition Risk" normally comprises three fact values "1", "2" and "3". However, when both the first and second fact values are removed, the fact comprises a collection with only one value.

Fact value overlap verification

Exclusivity/overlap verification focuses on the detection of fact values that are not exclusive, thus overlapping each other. This verification capability is only applicable for a fact that comprises a collection of fact values. For example, the fact "Body Mass Index Risk Points" consists of several fact values, from which some are categories of values, which indicate a certain risk level of a patient. In this case, the following fact values are present: "> 20", "18.5..20", "19..21" and "< 22". Both the fact values "18.5..20" and "19..21" are not exclusive as they partly overlap. This could result in a situation where two fact values are valid at the same time, potentially resulting in conflicting conclusions.

Lexical verification

Lexical error verification focuses on the usage of a wrong fact type in BR's. Take for example a wrong synonym or a complete other word than required. Another example within a BR: "Food Intake Risk Points of the patient must be equated to 0 IF Hard Intake of the patient <= 5 days AND Age of the patient <= 18 AND Fluid Intake of the patient <= 1 days." In this case, 'Solid' and 'Liquid' are replaced by 'Hard' and 'Fluid'.

Typographical and mechanical verification

Typographical and mechanical verification focuses on spelling, capitalization and punctuation errors in facts and fact values as part of business rules. An example of a typographical error would be: "Fod Intake Risk Points of the patient". Moreover, an example of a mechanical error would be: "Food Intake Risk Points of the patient."

Documentation verification

Lastly, documentation of each fact should be available in the fact vocabulary. If a fact is added to the fact vocabulary without any documentation, business rule analysts cannot utilize the fact vocabulary as a single point of truth, as double or conflicting facts could exist. For example, when the following two facts exist in the fact vocabulary: 'patient weight' and 'weight'. When no documentation is added for one of these facts the difference between both is hard to distinguish. However, automated verification is only able to check whether documentation is available, and not if it is semantically correct.

D – Generic verification level

Regarding the generic verification level, four capabilities regarding verification were identified.

Grammatical conformance verification

Grammatical conformance verification checks that all individual components in the business rule set are verified on whether they are designed according to the language-related guidelines. Take for example (regarding decision logic-level verification), the business rule "Body Mass Index Risk Points of the patient must be equated to 2 IF Patient Body Mass Index of the patient <= 18.5," which is a business rule to determine "E. Calculate Body Mass Index Risk Points". A syntax requirement of a business rule language could state that the combination of elements for a business rule needs to be consequent, in this case: fact (Body Mass Index) plus operator (=<) plus operand (18.5).

Declarativity verification

Declarativity verification checks whether there is no implicit or explicit order in which decisions, business rules, or facts are executed or evaluated. For example (regarding decision logic-level verification), we take the combination of two business rules: "BR14 - Food Intake Risk Points of the patient must be equated to 2 IF Solid Intake of the patient <= 5 days AND Age of the patient > 18 AND Liquid Intake of the patient <=1 days" and "BR15 - Food Intake Risk Points of the patient must be equated to 4 IF Solid Intake of the patient <= 5 days AND Age of the patient > 18

AND Liquid Intake of the patient > 1 days." This part of the ruleset may not imply any sequentially as it does not matter if business rule 15 is executed before business rule 14. The same holds for the sequence in which the business rule evaluates the condition facts.

Omission verification

Omission verification checks if required components on all three layers are missing. For example (decision requirements-level verification), decisions in a DRD modelled without a source or input data, or missing operands (i.e., =, >, =<), condition facts, conclusion facts, and fact values. Moreover, in another example (decision logic-layer verification), the operator and fact value of the conclusion fact are missing in the following business rule: "BR16 - Food Intake Risk Points of the patient must be equated to <…> IF Solid Intake of the patient <…> 5 days AND Age of the patient > 18 AND Liquid Intake of the patient > 1 days".

Atomicity verification

Atomicity verification focuses on the atomic design principle. This means that all components need to be normalized in such a state that no further normalization is possible. Therefore it checks whether all components are expressed and modelled in their atomic state. For example (decision requirements-level verification), it checks whether each decision is expressed in its atomic state, from which no other decisions could be derived, thus the decisions are fully normalized. Take the decision "Calculate Weight and Circumference Loss", which actually comprises both the Weight and Circumference loss of a patient. As both decisions are calculated using different business rules and facts, the decision should be further normalized into two unique decisions, one comprising only the weight loss and one comprising the circumference loss of a patient.

VI. CONCLUSIONS

Business rules, as part of business logic, are increasingly being utilized in organizations as building blocks for (automated) decision making. In this research we aimed to find an answer to the following research question: "*Which verification capabilities are useful to take into account when designing a business rules management solution?*" To accomplish this, we have conducted a three round focus group with five large Dutch governmental institutions. Our rounds of data collection and analysis resulted in a collection of 28 capabilities that, depending on the situation, must be taken into account when designing the verification capability as part of a BRM solution, see Figure 2. The BRM verification capability framework resulted from this study features capabilities for 1) the business decisions level, 2) decision logic level, and 3) the fact level. Additionally, our results presented a fourth category, 4) generic level capabilities with regards to verification.

From a theoretic perspective the verification framework provides a framework for further development and research into the verification capability and possible relationships with other BRM (sub)capabilities. This is important because practice often seems to confuse validation with verification and vice versa, however, both are different, see also (self-reference 2017). Further theoretical maturation is needed as the current Decision Management and Business Rules Management body of knowledge offers few means to structure verification of business decisions and business logic.

From a practical perspective, the verification framework offers practitioners a set of building blocks that could make up the verification in their organizations tooling. Depending on the situation, i.e., the language used to formulate the business decisions and business logic and the maturity of the tooling utilized, the framework offers verification types that can be implemented in different levels (fact-level, business logic level, and business decision level).

VII. DISCUSSION

As is generally true with empirical research, our results are subject to interpretation and are limited to the data and its context that was analyzed. Four threats to the validity of the conclusion are identified. First, the sample of organizations included is solely drawn from governmental institutions. Although we believe that governmental institutions adequately represent organizations that apply BRM, we lack the inclusion of commercial organizations in this study.

Second, regarding the research method and techniques utilized, our study included a sample of ten verification subject-matter experts. Despite the expertise levels of these ten subject-matter experts was substantive, this could be seen as a low number of participants. We argue that, given the maturity of the body of knowledge on verification with regards to business decisions and business logic, the qualitative approach with a lower number of participants was more suitable. This approach offered a more in depth view of the 'verification' phenomenon at these organizations and provides the body of knowledge with a basis to continue more 'broad-focused' research on the topic.

Third, with regards to the completeness of the results, this study allowed us to identify a large chunk of relevant verification capabilities in the Dutch governmental context. While we believe that the included organizations represent the Dutch governmental agencies fairly, the Dutch government houses many more governmental agencies that could not participate or were not involved in this study due to the scope and other practical matters. Of course we cannot claim that the verification capabilities presented in this paper represent the full spectrum of verification capabilities possible or utilized in practice.

Fourth one additional relevant factor with regards to our results might be the importance of each capability in practice, which we yet have to find an answer to. We stress that future research should focus on finding answers to such knowledge gaps.

VIII. FUTURE RESEARCH

This study and its results revealed multiple possible and interesting research directions. Undoubtedly the first research direction would be the shift from a narrow and deep focus of qualitative research to a more broad focus including both qualitative and quantitative research methods and data collection & analysis techniques. This allows for the inclusion of more participants, thus larger sample sizes, and therefore improves the generalizability of future results. The latter is important to both the body of knowledge and practice as it ensures validity of verification as a capability and drives adoption in practice. Additionally, future studies should include both governmental and commercial organizations and possibly identify best practices for both types of organizations. It would be interesting to find out whether verification is implemented significantly different in a commercial setting. On the one hand, this is possibly due to the fact that a governmental organization has to 'lead by example' rather than take risks on purpose. On the other hand, concerning the availability and expectations of a commercial organization's services (being very dependent on correctly implemented and executable business decisions and business logic), the possible risks could be much higher.

One dimension that could be explored in future research are the situational factors that drive the verification goals. For example, the type of organization, language utilized or type of services delivered could affect which verification capabilities are implemented. Such factors help organizations to easily identify where to focus on when designing and implementing verification capabilities.

REFERENCES

[1] K. Smit, M. Zoet, and M. Berkhout, "Verification Capabilities for Business Rules Management in the Dutch Governmental Context," in *Proceedings of the fifth International Conference on Research and Innovation in Information Systems (ICRIIS), Langkawi, Malaysia*, 2017.

[2] M. Zoet, *Methods and Concepts for Business Rules Management*, 1st ed. Utrecht: Hogeschool Utrecht, 2014.

[3] J. Boyer and H. Mili, *Agile business rule development: Process, Architecture and JRules Examples.* Springer Berlin Heidelberg, 2011.

[4] Object Management Group, "Decision Model And Notation (DMN), Version 1.0," 2015. [Online]. Available: http://www.omg.org/spec/DMN/1.0/.

[5] Object Management Group, "Decision Model And Notation (DMN), Version 1.1," 2016.

[6] A. F. Ackerman, L. S. Buchwald, and F. H. Lewski, "Software inspections: an effective verification process," *IEEE Softw.*, vol. 6, no. 3, pp. 31–36, 1989.

[7] A. Vermesan and F. Coenen, *Validation and Verification of Knowledge Based Systems: Theory, Tools and Practice*. Springer Science & Business Media, 2013.

[8] W. M. Van der Aalst, "Formalization and verification of event-driven process chains," *Inf. Softw. Technol.*, vol. 41, no. 10, pp. 639–650, 1999.

[9] B. Buchanan and E. Shortcliffe, *Rule-Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison Wesley Publishing Company, 1984.

[10] A. Deutsch, R. Hull, F. Patrizi, and V. Vianu, "Automatic verification of data-centric business processes," in *Proceedings of the 12th International Conference on Database Theory*, 2009, pp. 252–267.

[11] F. Puhlmann, "Soundness verification of business processes specified in the pi-calculus," in *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems,"* 2007, pp. 6–23.

[12] R. Studer, V. R. Benjamins, and D. Fensel, "Knowledge engineering: principles and methods," *Data Knowl. Eng.*, vol. 25, no. 1, pp. 161–197, 1998.

[13] B. Von Halle and L. Goldberg, *The Decision Model: A Business Logic Framework Linking Business and Technology*. CRC Press, 2009.

[14] T. Morgan, *Business rules and information systems: aligning IT with business goals*. Addison-Wesley Professional, 2002.

[15] A. Kovacic, "Business renovation: business rules (still) the missing link," *Bus. Process Manag. J.*, vol. 10, no. 2, pp. 158–170, 2004.

[16] S. Schlosser, E. Baghi, B. Otto, and H. Oesterle, "Toward a functional reference model for business rules management," in *the 47th Hawaii International Conference on System Sciences (HICSS)*, 2014, pp. 3837–3846.

[17] M. Zoet and K. Smit, "An Economic Approach to Business Rules Normalization," in *Proceedings of the The Ninth International Conference on Information, Process, and Knowledge Management*, 2017, pp. 1–6.

[18] K. Smit and M. Zoet, "Management Control System for Business Rules Management," *Int. J. Adv. Syst. Meas.*, vol. 9, no. 3–4, pp. 210–219, 2016.

[19] K. Smit, M. Zoet, and M. Berkhout, "A Framework for Traceability of Legal Requirements in the Dutch Governmental Context," in *Proceedings of the 29th Bled eConference*, 2016, pp. 151–162.

[20] K. Smit and M. Zoet, "Utilizing Change Effort Prediction to Analyze Modifiability of Business Rule Architectures at the NHS," in *PACIS 2016 Proceedings*, 2016, p. paper 261.

[21] B. W. Boehm, "A spiral model of software development and enhancement," *Computer (Long. Beach. Calif).*, vol. 21, no. 5, pp. 61–72, 1988.

[22] G. J. Holzmann, "The model checker SPIN," *IEEE Trans. Softw. Eng.*, vol. 23, no. 5, pp. 279–295, 1997.

[23] W. Van der Aalst, H. De Beer, and B. Van Dongen, "Process mining and verification of properties: An approach based on temporal logic.," in *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*, 2005, pp. 130–147.

[24] K. Smit and M. Zoet, "Management Control System for Business Rules Management," *Int. J. Adv. Syst. Meas.*, vol. 9, no. 3–4, pp. 210–219, 2016.

[25] S. Liao, "Expert system methodologies and applications—a decade review from 1995 to 2004," *Expert Syst. Appl.*, vol. 28, no. 1, pp. 93–103, Jan. 2004.

[26] Camunda, "Camunda Github - Supported Languages," 2017. [Online]. Available: https://github.com/camunda/camunda-docs-

manual/blob/master/content/user-guide/dmn-engine/expressions-and-scripts.md. [Accessed: 30-Aug-2017].

[27] M. L. Nelson, J. Peterson, R. L. Rariden, and R. Sen, "Transitioning to a business rule management service model: Case studies from the property and casualty insurance industry," *Inf. Manag.*, vol. 47, no. 1, pp. 30–41, Jan. 2010.

[28] A. C. Edmondson and S. E. Mcmanus, "Methodological Fit in Management Field Research," *Proc. Acad. Manag.*, vol. 32, no. 4, pp. 1155–1179, 2007.

[29] A. L. Delbecq and A. H. Van de Ven, "A group process model for problem identification and program planning," *J. Appl. Behav. Sci.*, vol. 7, no. 4, pp. 466–492, 1971.

[30] A. Strauss and J. M. Corbin, *Basics of qualitative research: Grounded theory procedures and techniques*. Sage Publications, Inc, 1990.

[31] B. G. Glaser, *Theoretical sensitivity: Advances in the methodology of grounded theory*. Sociology Press, 1978.

[32] K. Smit, M. Zoet, and M. Berkhout, "Technical Report Case-2016-0001," Utrecht, 2016.

APPENDIX A: Determine malnutrition risk DRD

Decision Requirements Level

In the decision requirements level, seven decisions are modelled with their corresponding business knowledge, input data and knowledge sources, see Figure A-1.

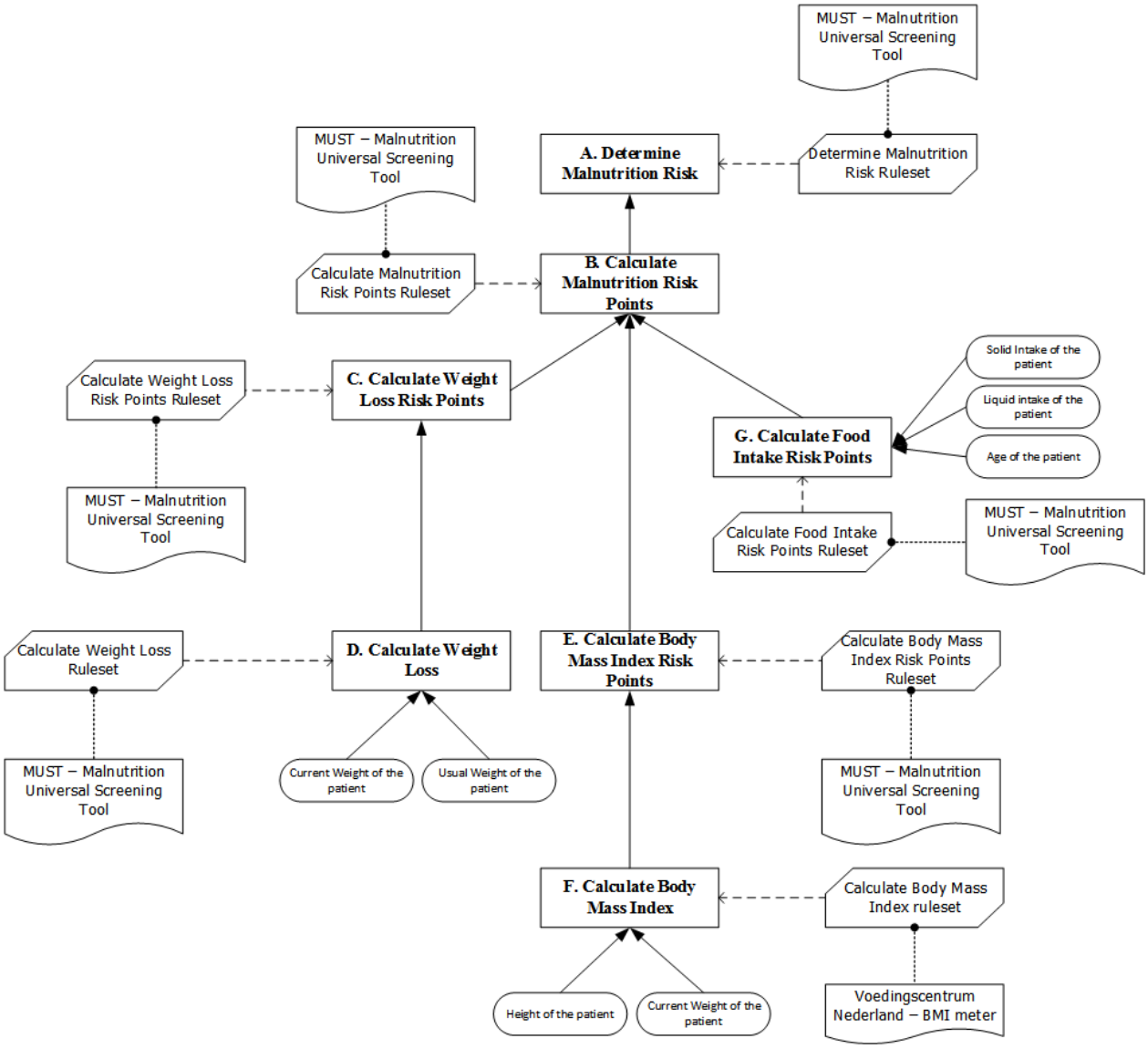


Figure A-1. Decision Requirements Level Diagram – Assess malnutrition risk [32]

An Analysis of the Extent to which Standard Management Models Encourage the Adoption of Green IT

William M. Campbell
School of Computing and Digital Technology
Birmingham City University
Birmingham, UK
B4 7XG
Email: william.campbell@bcu.ac.uk

Jagdev K. Bhogal
School of Computing and Digital Technology
Birmingham City University
Birmingham, UK
B4 7XG
Email: jagdev.bhogal@bcu.ac.uk

***Abstract*—This paper explores the extent to which senior managers using standard management models as tools for developing corporate strategy, structures and culture are likely to be encouraged to adopt green IT. A range of standard management models are considered: strategic, tactical and operational. Analysis reveals that many standard models, in particular older ones that rely heavily on numbers and take a narrow view of corporate responsibility, are not favourable to the adoption of green IT. Accordingly, managers need to avoid excessive reliance on such models and should consider using models which take account of softer issues, in particular those models which address sustainability directly. There is a need for the development of new management models, which more explicitly integrate traditional bottom line considerations with the wider ethical responsibilities of companies, including sustainability. Cameron and Quinn's "Competing Values Framework" is used as a tool to explore organizational culture. A statistical analysis of a survey of organizational culture and greenness is presented. It is concluded that organizational culture has a major impact on the adoption of green IT and consideration must be taken of it when introducing green IT initiatives.**

Keywords–Green; Sustainability; Green IT; Information Technology; Organizational Culture; Management Models; Statistics.

I. INTRODUCTION

This paper extends [1]. It also builds on [2] and [3], by undertaking a survey on organizational culture and green IT level, using the "Competing Values Framework" and seeking to identify significant statistical correlations.

The sustainable use of resources is a key issue facing the human race. It is widely accepted that the emission of Greenhouse gases has affected the climate. Other issues include pollution and the careless disposal of waste. Belkhir and Elmeligi [4] estimate that the carbon footprint of Information and Communication Technology (ICT) will rise from between 1 and 1.6% in 2007 to around 14% in 2040. The global greenhouse gas emissions (GHGE) resulting from ICT will exceed that of the agricultural sector. The sustainable use of resources in a way that does not damage the environment, the so-called "green" agenda, is one of the key issues facing the human race in the early 21st Century. The role of ICT is clearly significant.

However, while ICT significantly contributes to GHGE, it can also contribute to reducing pollution through technologies such as "intelligent buildings", cloud computing and smart logistics.

There has been pressure on individual companies to take note of environmental issues [5]. This has come not only from the need to comply with environmental legislation, but also from consumer pressure and concern about reputation. Many companies now accept that economic performance is not the only measure of success and have adopted a "Triple Bottom Line" of environment, society and economic performance [6] [7].

In determining corporate strategy and organizational structures senior managers often seek guidance from the standard management models taught in business schools. The extent to which these models encourage the adoption of green IT will, therefore, have an effect on the extent to which managers regard green IT as a serious, mainstream issue.

Organizational Culture has long been recognized as an issue of great importance within the business literature and, in recent years, substantial attention has been devoted to its impact on the adoption of green initiatives. It has been argued that for companies systematically to incorporate environmental concerns into their activities requires a major change of corporate culture [8][9]. However, there has been limited consideration of the impact of organizational culture on the adoption of green IT. A key theme of this paper is to explore the role of culture within IT, and the extent to which particular types of culture facilitate green initiatives. Cameron and Quinn's "Competing Values Framework" is used as a tool to explore organizational culture.

The remainder of the paper is structured as follows: Section II looks at the green agenda, focusing in particular on green IT. Section III explores management and organizational models, which specifically address green IT. Section IV investigates the extent to which standard management models focusing on strategy are favourable to green IT. Section V investigates the extent to which standard management models focusing on tactics are favourable to green IT. Section VI investigates the extent to which standard management models focusing on operational management are favourable to green IT. Section VII presents some general conclusions about management models.

The following sections present the survey on organizational culture and green IT level. Section VIII considers statistical applications of the Competing Values Framework. Section IX describes the survey and the research methodology. Section X presents the results of the study, along with analysis. Section XI has some discussion of the results, in particular reflecting on

issues which arise when statistically analysing culture surveys. Finally, the Conclusion summarises the key points of the paper, makes some recommendations and looks at possible future research directions.

II. The Green Agenda

The definition of sustainability provided by the Brundtland Commission has gained widespread acceptance: "Development that meets the needs of the present without compromising the ability of future generations to meet their needs"[10]. There has been a number of agreements, most recently the Paris Agreement in 2016. Its central aim was to strengthen the global response to the threat of climate change, by keeping the global temperature rise this century well below 2 degrees Celsius above pre-industrial levels and to pursue efforts to limit the temperature increase even further to 1.5 degrees Celsius [11].

The terms "IT" and "ICT" are not clearly distinguished in the literature or in general usage, although "ICT" more explicitly includes the use of communication networks. Jenkin et al. [12] distinguish between "Green IT" and "Green IS" . They define "Green IT" as the attempt to reduce energy consumption and waste associated with the use of both hardware and software. "Green IS" they define as the use of information systems to support environmental sustainability initiatives. In this paper we use "Green IT" as a generic term, covering all efforts to reduce the environmental damage caused by the use of IT (including networks), or to use IT in a positive way to improve the environment.

IT has played an increasingly important role in industry and commerce and makes a substantial contribution to the environmental footprints of companies, through both the use of IT and the construction and disposal of IT equipment. IT data centres make a major contribution to the carbon footprint of many corporations. Data centres worldwide are projected to produce around 495 million tonnes of carbon annually by 2020 [4]. The Internet of Things, smartphones and cryptocurrencies are growing sources of GHGE.

However, the application of IT can make a positive contribution to sustainability in various ways. Software as a Service (SAAS) and Cloud Computing offer ways for using IT resources more efficiently. Companies purchase data storage and rent software, as required, from external providers. These can be accessed using "thin client" computers. Server virtualization has provided the opportunity for servers to be used more efficiently; this allows several servers to be consolidated as virtual servers on one physical server, enabling sharing of resources and economies of scale.

Environmental information systems and "intelligent buildings" help to reduce energy wastage; supply chain information systems optimize routing and transportation [13]. Dao et al. [14] argue for combining IT resources with supply chain management and human resource management within an integrated sustainability framework.

Green IT must ultimately, in large part, be delivered by companies. In recent years, companies have started to recognise that having a "triple bottom line" of People, Profit and Planet is actually good for profitability. It enhances reputation and encourages the development of valuable capabilities relating to sustainability [7]. Deutsche Bank, for example, has an eight point Green IT policy [15].

III. Management and Organizational Models for Green IT

Bokolo et al. [16] provide a systematic and up-to-date review of literature on green IT. This illustrates that much effort, across a number of disciplines, has been put into developing models and frameworks for analysing green IT.

Murugesan and Gangadharan [17] divide enterprise green IT strategy into three approaches.

Tactical Incremental Approach. In this approach, the company retains the existing infrastructure and policies and introduces simple measures such as switching off computers when not in use.

Strategic Approach. In this approach, the company develops a comprehensive plan for making its deployment of IT more energy-efficient.

Companies following a *Deep Green Approach* go beyond the *Strategic Approach*, adopting additional measures such as a carbon offset policy to neutralize greenhouse gas emissions.

One of the mostly widely-cited models is Molla and Cooper's "Green IT Readiness" or "G-Readiness" framework [18]. It divides IT into IT Managerial Capability, IT Human Capability and IT Technical Capability. An organization's green IT maturity is assessed in terms of attitude, policy, practice, technology and governance. There is an accompanying G-Readiness Survey instrument.

Deng and Ji undertook a review of the literature, seeking to identify the motivating factors for companies to adopt green IT [19]. They noted that the literature has "scattered theoretical foundations", but identified the following key underlying theories.

The *Diffusion of Innovation Theory* investigates the process by which innovations spread.

Institutional Theory analyses the pressures which influence the development of organizations. A key institutional pressure is "mimetic isomorphism", the tendency of companies to follow leading companies in their field.

Organizational Culture views organizations as social structures and examines the way shared assumptions and norms emerge. This is discussed later in the section on Cameron and Quinn's Competing Values Framework .

The *Resource Based View* (RBV) [20] takes the view that a company's competitive advantage resides in its ownership of a set of resources that are not easily duplicated by a competitor. These resources can be physical, organizational or social.

Hart [21] extends this to the *Natural Resource Based View* (NRBV), by including resources and capabilities particularly relating to sustainability.

Deng and Ji introduce a theoretical framework for "Organizational Green IT Adoption" (OGITA). This has the external drivers of technological context and institutional pressures; and internal drivers of senior management attitudes, corporate strategy and organizational culture.

However, senior managers looking for guidance on changing company strategy, structures and culture are likely to refer to standard management models. Almost a third (31%) of the world's largest 500 companies have a chief executive with an MBA [22]. It is likely that the management models they

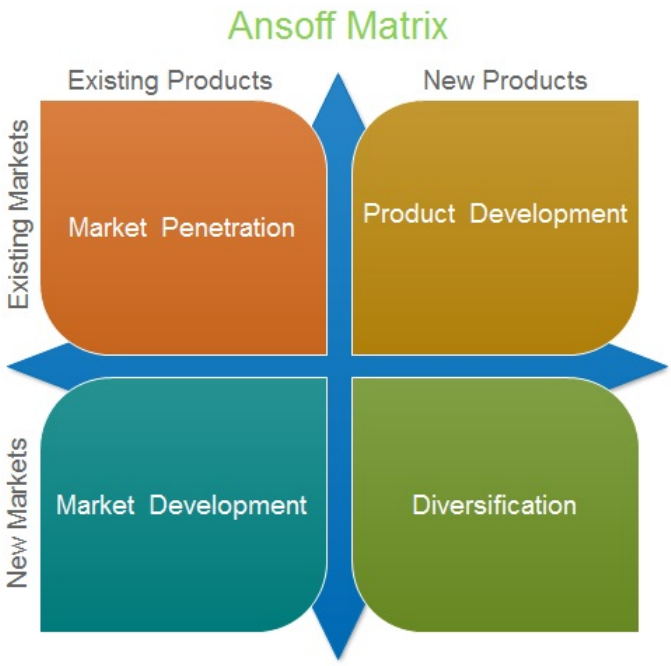


Figure 1. Ansoff Matrix

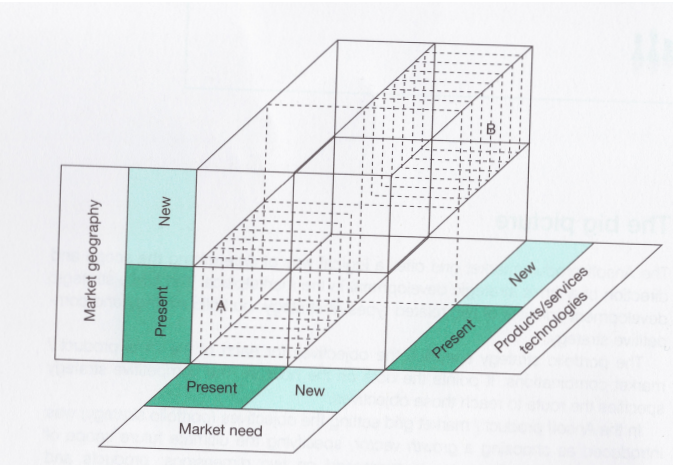


Figure 2. Ansoff Cube [23]

studied will have influenced them in their later careers. We discuss this in the next section.

We use a standard, widely used and influential book on management models [23]. We follow its separation of models into strategic, tactical and operational. In each case, we explore the extent to which managers employing these models are likely to be encouraged to adopt green IT. The extent to which these models "favour" green IT will, therefore, have a major impact on its adoption.

IV. Standard Management Models Focusing on Strategy

These models help a company to analyse its strategic position and develop strategic plans for the future.

A. Ansoff's Matrix

Ansoff's Matrix is a widely used model for helping companies determine their strategy for developing new products and entering new markets [24]. In terms of products, they would have a choice of retaining existing products or developing new products. In term of markets, they would have a choice of focusing on existing markets or developing new markets. This produces four top-level strategies, as illustrated in Figure 1. The top left quadrant is the "conservative" strategy of focusing on existing products and markets; the bottom right quadrant is the "aggressive" strategy of developing new products and seeking new markets.

The model has been extended to a cube, by introducing a geographical dimension, where companies consider expanding into new countries. This is illustrated in Figure 2.

We now consider each of the four quadrants from the perspective of green IT. Unless a company is already selling green products, only the two right hand quadrants are relevant. The *Product Development* quadrant would require the promotion of new green products to existing customers. Managers are likely to regard this as challenging, unless customers are dynamic, entrepreneurial and open to change. It would probably be easier to sell products which have undergone modest modification to be greener, rather than radical new green information systems using technology with which customers may not be familiar, such as the Internet of Things or "smart homes". Selling radical new green products such as environmental monitoring systems would probably require the development of new markets and therefore belongs in the *Diversification* quadrant.

The Ansoff model advises companies to consider four issues: competitive advantage, potential synergies across the company's core competencies, strategic flexibility (the ability easily to modify strategy to cope with unpredicted events), and the potential for geographical growth.

We now use the OGITA model discussed above to evaluate the extent to which use of the Ansoff Matrix would be likely to encourage companies to adopt green IT. We first consider external pressures.

From a technology perspective, the questions would be:

- Would going green give the company a relative technological advantage?
- Would it be technically challenging?
- Would it make use of core technical competencies within the company?

In considering these issues, technological experts within the company would be considering the challenges of developing new products, against the backdrop of the possibility of just going for greater market penetration in existing markets or developing new markets. Unless there was a compelling reason to suppose that the greener product would provide a competitive technical advantage or the existing product would become obsolete because of its poor green credentials, technology experts would be likely to favour avoiding radical changes to the existing product portfolio.

We noted above that there are essentially two types of green IT: those which try to avoid negative environmental impacts of IT-related products and those which use information systems to promote sustainability in applications such as environmental monitoring and smart cities. The latter are likely

to involve developing radically new products and be much more challenging in technical terms. They are likely, therefore, to be deemed unattractive.

From the perspective of external institutional pressures, the questions would be:

- Will the company be breaking the law, if it does not make its products greener?
- Will the company become out of step with the market if it does not become greener?
- Does the company face a risk of reputational damage?

Unless the company is driven by a powerful "mimetic isomorphism" pressure, external pressures for greenness are unlikely to be stronger than economic pressures.

Finally, we consider the internal motivations of the OSITA framework. Senior managers tend to be driven by numbers and verifiable evidence. It is likely to be easier to provide clear evidence for the benefits of taking existing products into new markets than to demonstrate that a market will exist for radical new green products. Many green products are "disruptive technologies", for which there is currently no market. As Christensen argues in his influential work "The Innovator's Dilemma", company culture is frequently hostile to such technologies [25]. Unless there are a number of green champions within the company at a senior level, top managers are likely to favour developing existing markets and making only incremental changes to existing products.

Bilgeri et al. [26] applied the Ansoff Framework to ten companies which were working in the Internet of Things. They found that the companies were most keen on the Product Development strategy. Only one company had a strong focus on Diversification.

In summary, the Ansoff model is likely to discourage companies from developing new greener products, because it juxtaposes the challenge of developing radical new products with the easier option of expanding the market for existing products. Insofar as the use of Ansoff's Matrix encourages the adoption of green IT, it is likely to be of a "Tactical Incremental" nature, within Murugesan's taxonomy of green initiatives discussed above.

Another strand in Ansoff's research on strategic management is presented in [24] and discussed in a modern context in [27]. Ansoff argued that strategic planning was only reasonable when historical trends were developing incrementally, and was not useful when dealing with surprises in an unpredictable environment. In such cases, managers had to respond to "weak signals" in a context of limited information. Ansoff argues that the rational response in such circumstances is to have a flexible strategy and determine which actions will be appropriate when more information becomes available. The signals coming from governments about the need for a more green approach by companies are vague and changeable and are probably perceived by senior managers as "weak". In such circumstances they are likely to postpone radical action.

B. Porter's Five Forces

Porter's Five Forces is one of the most established management models, and has been used for around forty years. It is used by companies contemplating entering a new industry. It identifies five things that need to be considered:

- New entrants
- Substitutes (will it be easy to replace the proposed product with something else?)
- Buyers
- Suppliers (companies which will be below you in the supply chain)
- Existing Competitors

The employment of Porter's Five Forces is likely to discourage companies from developing radical new green products and services, for the same reason as Ansoff's Matrix. As Christensen (discussed above) notes, you cannot analyse a market that does not exist. In particular, companies are likely to worry about finding buyers, where currently there are none. They will also be worried about the difficulty of constructing an efficient supply chain.

C. The BCG Matrix

The Boston Consulting Group Matrix goes back to the 1970s [28] [23]. It is used by companies for planning their product portfolio. It is similar to the Ansoff Matrix, having two dimensions; in this case, the dimensions are the projected Market Share and Market Growth. This again creates four main types of market:

1) high market share, high growth (best)
2) high market share, low growth
3) low market share, high growth
4) low market share, low growth (worst)

What "advice" will this model give? The market for a new green Cloud service is likely to be of the third type. The Cloud market is highly competitive but is likely to grow. The market for a new environmental monitoring system for reservoirs is likely to be of the second type. The market is small and unlikely to grow substantially, but a successful product could have reasonable expectations of dominating it. Markets for new IT products which are incrementally more energy efficient are likely to be of the fourth type. Few green markets are likely to be of the first type. It seems probable that senior decision makers using the BCG will favour potential new markets of the first type rather than green markets.

D. The Blue Ocean Strategy

This model makes a distinction between a Red Ocean Strategy, where a company seeks to beat the competition in an existing market; and a Blue Ocean Strategy, where a company seeks to develop a brand new market. It encourages companies to focus on the big picture rather than the numbers [29] [23]. It provides an antidote to the problems identified by Christensen, discussed above, where managers tend to avoid disruptive technologies. Two types of blue oceans can be created, either by inventing a new industry or by expanding the strategic boundaries of an existing industry.

Employment of the Blue Ocean model is likely to be positive for the development of new Green IT applications, which potential users were unlikely to have imagined as a possibility, such as the use of the Internet of Things in Western Africa to forecast air quality [30] or the application of blockchain technology in peer-to-peer transactions in photovoltaic power generation [31].

E. Kay's Distinctive Capabilities

The Kay's Distinctive Capabilities (KDC) model originates from the Resource Based View, discussed above, which regards a company as a collection of skills and capacities, many intangible, which cannot easily be imitated [32] [23]. KDC separates these into three categories:

- Architecture (features intrinsic to the company and its relationships with customers and suppliers)
- Reputation
- Capacity to innovate

To some extent this model encourages green innovation. It acknowledges the value of a company having a reputation for being ethical. Furthermore, the extension of the RBV discussed above, the Natural Resource Based View, explicitly recognizes that green capabilities are likely to be important in the future. But the model emphasises that it is very difficult to convert innovation into competitive advantage. The success of a radical new and efficient Cloud Computing model will be greatly affected by whether competitors are developing a similar product.

V. Standard Management Models Focusing on Tactics

These models help a company to organize its process, resources and people. They address "how to" questions.

A. Cameron and Quinn's Competing Values Framework

Anthropology takes the view that organizations are cultures; sociology takes the view that organizations have cultures [33]. Most organizational theory adopts the sociological perspective, regarding culture as an attribute of an organization that can be measured and analysed. Schein [34] defined organizational culture as: "A pattern of shared basic assumptions that the group learned as it solved its problems of external adaptation and internal integration that has worked well enough to be considered valid and hence to be taught to new members as the correct way to perceive, think and feel in relation to those problems."

Schein identified three levels of culture:

- Artifacts, those aspects which are on the surface such as dress and can be easily identified;
- Espoused Values, that is conscious goals, strategies and philosophies;
- Basic Assumptions and Values. These exist at a largely unconscious level, form the inner core of culture and are hard to identify.

Basic Assumptions and Values have the deepest influence and are the most difficult to change. Many attempts at organizational change fail because of a failure to change the underlying culture [35].

Many dimensions of organizational culture have been proposed, for example, Hofstede [36]: power distance, uncertainty avoidance, individualism, and masculinity. Cameron and Quinn's "Competing Values Framework" (CVF) originated from a cluster analysis of these dimension schemes. It identifies two key dimensions: Internal Focus and Integration versus External Focus and Differentiation; and Stability and Control versus Flexibility and Discretion [37] [38]. The CVF has been used in many research studies and has been shown to have a high degree of validity [39].

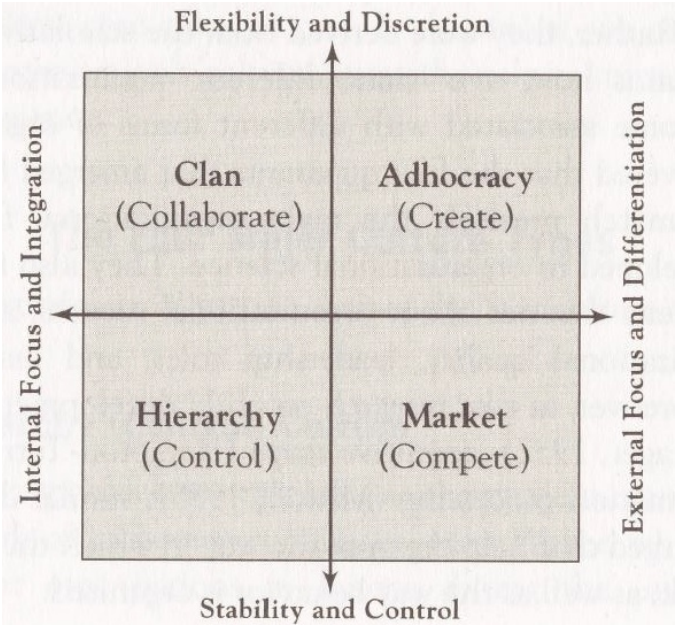


Figure 3. Cameron and Quinn [38]

The four key culture types identified by the CVF are illustrated in Figure 3 and may be summarized as follows (Adapted from [38]):

- Hierarchy. Such organizations tend to be bureaucratic. Formal rules and policies hold the organization together. The long-term goals of the organization are stability, predictability and efficiency. Government agencies and the military are typical hierarchical cultures.
- Market. The workplace is results-oriented. Leaders tend to be aggressive and demanding. The glue that holds the organization together is an emphasis on winning. Success is defined in terms of beating the competition and market share.
- Clan. The organization is held together by loyalty, tradition, and collaboration. It is a friendly place to work, where people share a lot of themselves. Leaders are thought of as mentors and coaches. Success is defined in terms of internal climate and concern for people. The organization places a premium on teamwork, participation, and consensus.
- Adhocracy The workplace is dynamic, creative, entrepreneurial and risk-oriented. The emphasis is on being at the leading edge of new knowledge, products, and/or services. The glue that holds the organization together is commitment to experimentation and innovation. Success is defined as the production of innovative and original products and services.

The CVF has an accompanying Organizational Culture Assessment Tool (OCAI). It consists of a questionnaire requiring employees to assess their organization, using an ipsative scale, on six characteristics: Dominant Characteristics, Organizational Leadership, Management of Employees, Organization

Glue, Strategic Emphases and Criteria for Success. A culture profile diagram can then be produced.

There has been a considerable amount of research on the relationship between types of organizational culture and effectiveness. Richard et al. [40] conducted a survey of US firms. They found that clan cultures resulted in higher earnings and employee satisfaction.

In the US health industry, Gregory et al. [41] found a positive link between group (clan) culture and patient and physician satisfaction and also a slight link between balanced cultures and satisfaction.

Linnenluecke and Griffiths [42] used the CVF as a framework for investigating the likely emphases which will be adopted by companies with different types of culture, in pursuing corporate sustainability. They argued that companies would favour initiatives that were congruent with their dominant culture.

The successful adoption and diffusion of green IT systems will also be affected by the organizational culture of companies. Green IT systems are likely to be "disruptive technologies", which are regarded as risky. For example, attempts to reduce energy use associated with data storage through the employment of "cloud computing" may raise fears about security. Green IT systems are, therefore, more likely to be favoured by companies with clan or adhocracy cultures, which are non-hierarchical, entrepreneurial and can embrace change.

The use of the Cameron and Quinn model as a framework for discussing the impact of organizational culture on the adoption of green IT is discussed in detail in [3] [2]. A statistical analysis of a survey is presented in Sections VIII to XI.

B. Beer and Nohria E and O Theories

Beer and Nohria is a modern management model, which explicitly emphasises the value of soft skills and the importance of companies behaving ethically and taking account of their corporate social responsibility [43] [23].

They have two main theories of change:

- Theory E. This focuses on the creation of economic value for shareholders. It involves formal systems and structures. The decision making process is top-down. Changes are carefully planned.
- Theory O. This focuses on a culture that develops employee commitment and takes note of a company's ethical responsibilities. Employees are encouraged to change and evolve. Change is emergent.

To be successful, a company must embrace both Theory E and Theory O and confront the tension between them.

The "Theory O" culture combines elements of the adhocracy and clan cultures of the Competing Values Framework which, it is argued above, are conducive to the adoption of green IT.

The Beer and Nohria model is favourable to the adoption of green IT, because it encourages managers and employees to think of the bigger picture and not just focus on narrow financial considerations. In particular, it asks companies to take account of their ethical responsibilities. But the model does not ignore the practical exigencies of operating a successful business. For companies successfully to adopt green IT they must both have a vision and have the operational capability to realise it in the real world of business. The Beer and Nohria model provides a framework for constructively reconciling the conflicting pressures this creates.

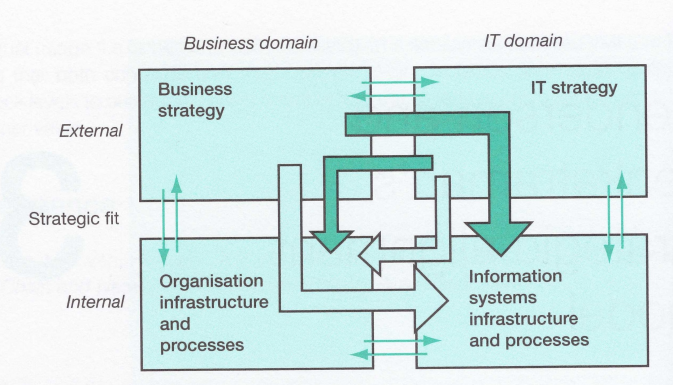


Figure 4. Strategic Alignment Model

C. Henderson and Venkatram's Strategic Alignment Model

This model addresses IT strategy directly. It seeks to promote alignment between business strategy and IT strategy and also between the IT infrastructure and business operations. Research on alignment between business strategy and IT goes back to the late 1980s, when Venkatram and Henderson developed the classic and still highly influential, "Strategic Alignment Model". It is presented in [44], [45] and, slightly amended, in [46]. Coltman et al. [47] describe it as being "on the list of seminal and transformative IS publications". A key feature of this model is that it provides for IT strategy influencing business strategy. Venkatram and Henderson noted that IT had shifted from a traditional back-office, support role towards being an integral part of the strategy of organizations [44].

The model recognises four domains: *Business strategy*, *Organization infrastructure and processes*, *IT strategy*, and *Information Systems infrastructure and processes*. The model is illustrated in Figure 4.

Strategic alignment has two elements: *strategic fit* and *functional integration*. *Strategic fit* refers to alignment between strategy and internal structure and processes, with regard both to business and IT. *Functional integration* refers to alignment between general business matters and IT matters, with regard both to strategy and internal structures and processes.

An important part of the underlying rationale of the model is that there should be cross-domain relationships between the business and IT domains. Four key alignment perspectives are identified. We consider each of them and their relevance to encouraging the adoption of Green IT/Green IS. The alignment perspectives correspond to the arrows in Figure 4.

Strategy Execution

This is visualized by the anti-clockwise arrow from top-left to bottom right in Figure 4. This corresponds to the traditional, hierarchical view of organizations, with business strategy driving organizational infrastructure and information

systems infrastructure and processes. This leaves IT managers in a subordinate role of Business Strategy Implementor. This is unlikely to be favourable to the adoption of green IT, unless the organizational strategy has a strong focus on sustainability.

Technology Transformation

This is visualized by the clockwise arrow from top-left to bottom right. This involves implementing business strategy through IT strategy and then the development of appropriate IT infrastructure and processes. This puts IT managers in the role of Technology Architects. They are in a more influential role than in *Strategy Execution*, because they are not constrained by the current organizational structure. This perspective will be conducive to the adoption of green IT, if there is a green organization strategy.

Competitive Potential

This is called *Technology Exploitation* in [44]. It is represented by the anti-clockwise arrow from top-right to bottom left. This perspective provides for IT strategy influencing organization strategy. A green IT strategy could then drive a change in the organization strategy, which led to changes in the organization infrastructure to reduce the carbon footprint of the organization's use of IT. It could also lead to changes, such as the use of cloud computing services or the development of environmental monitoring systems.

Service Level

This is called *Technology Implementation* in [44]. It is represented by the clockwise arrow from top-right to bottom left. Here the organizational infrastructure follows the IT infrastructure, which is determined by IT strategy. This could be conducive to limited changes to the organizational processes, which reduced the energy consumption of IT within the organization; but not to more fundamental changes such as server virtualisation or the use of the Internet of Things to support environmental sensors, which would need to be driven by business managers at board level.

There is no "right" perspective. Effective strategy development would use all the perspectives as lenses through which to view strategy. Coltman et al. [47] observe that in recent years digital business strategy has effectively become the strategy of many companies, which makes the concept of IT-business strategy alignment less meaningful.

Overall, Henderson and Venkatram's Strategic Alignment Model provides a framework which, if all the perspectives are analysed, should be conducive to the adoption of Green IT.

Loeser et al. [48] extend Henderson and Venkatraman's Strategic Alignment Model to a Strategic Green IT Alignment Framework (SGITAF), which explicitly incorporates Environmental sustainability and Green IT domains. They argue that SGITAF provides a framework which supports the leveraging of Green IT's full potential.

VI. Standard Management Models Focusing on Operational Management

These models help a company to optimize operational process and activities.

The *Change Quadrants* model is a tool to assist companies to effect a particular change [49] [23]. It analyses companies on two dimensions: whether they are "warm" or "cold"; and whether the key motivation for the proposed change is "warm" or "cold".

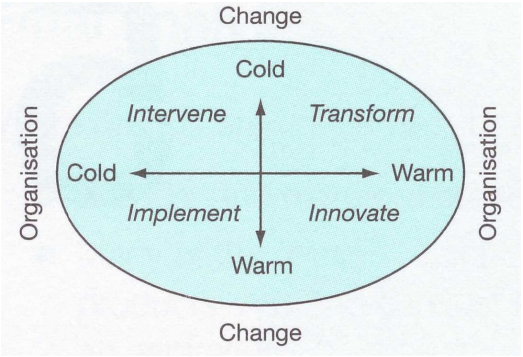


Figure 5. Change Quadrants

A warm organization is one where there is a shared sense of values and employees do not have a merely transactional relationship with the organization. It is rather like the "Clan Culture" in the Cameron and Quinn Competing Values model. A cold organization is one which is hierarchical and governed by rules, systems and procedures.

A warm motivation for a proposed change is driven by a shared sense of values across the company. A cold motivation is a response to a crisis such as the emergence of a dangerous competitor.

This produces the four quadrants in Figure 5. The change strategy should be tailored to the quadrant. A "warm organization that is willing" (the bottom right quadrant) will be open to change. It will be possible to develop a long-term vision bottom-up. It would be possible to adopt a Deep Green approach to Green IT, in Murugesan and Gangadharan's taxononmy. A "cold organization that is obligated" (the top left quadrant) will have to drive change top-down; employees will only have a say in the implementational details. It would only be possible to adopt a Tactical Incremental approach. The key message of the model is that real transformation, such as is involved in the systematic adoption of green IT, requires a warm organization and a warm motivation for change.

VII. General Conclusions about Management Models

Most of the older models are driven by relatively short-term bottom line considerations. These are likely to be unfavourable to green IT. More recent models, such as Beer and Nohria and Change Quadrants, tend to adopt a wider perspective on the responsibilities of companies and also take more note of "softer" people and ethical issues. They are more likely to be favourable towards green IT.

Managers need to be cautious about over-reliance on standard models, especially those which take a narrow view of corporate responsibilities. They should consider employing models which take account of wider issues, in particular those models which incorporate consideration of sustainability.

VIII. Statistical Applications of the "Competing Values Framework"

Sections VIII to XI present a survey on organizational culture and green IT level, using the "Competing Values

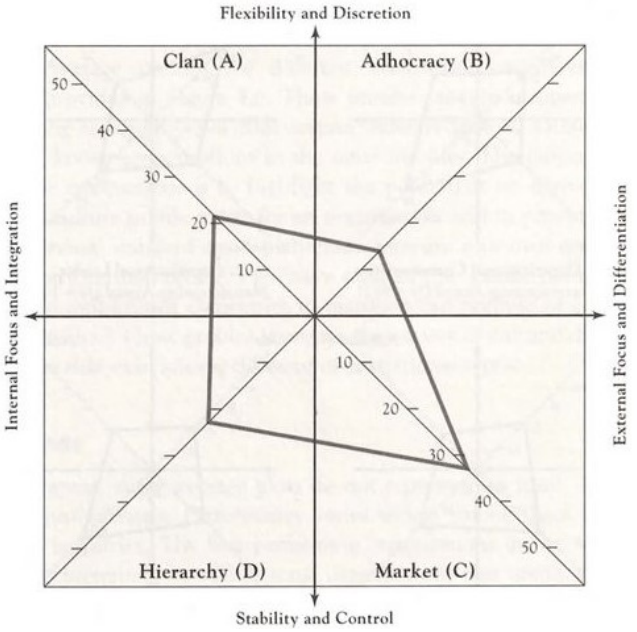


Figure 6. Average Culture Profile [38]

Framework" and seeking to identify significant statistical correlations. It used the OCAI, mentioned above. For each characteristic, people are required to distribute 100 points between statements about company culture, corresponding to the four culture types.

The results can be used for various purposes, e.g, to calculate the average profile of an organization and identify the main culture types(s); to identify discrepancies between current and preferred culture; and to ascertain the degree of congruence between results produced by different groups of employees. Cameron and Quinn averaged the results for over one thousand companies; this resulted in the average profile in Figure 6.

This indicates that companies tend, on average, to be dominated by market and hierarchy cultures. Cameron and Quinn also found a tendency for companies to drift towards a greater emphasis on market and hierarchy as they matured. Both of these factors suggest a tendency for corporate culture to be inimical to sustainability initiatives.

Some researchers use a Likert version of the OCAI. The implications of this are discussed later.

All of the example uses of the CVF mentioned in earlier sections use statistical analysis, with the exception of Linnenluecke and Griffiths [42]. There have been many other statistical applications of the OCAI to investigate the relationship between culture and various factors. Business Process Management is considered in [50]; Knowledge Management is considered in [51].

IX. Research Methodology

In this section, we state the research questions and outline the design and execution of the survey which was undertaken.

A. Research Questions

The key research questions are:

- Does organizational culture have an impact on the adoption of green IT within companies?
- Is the impact, if any, which organizational culture has on the adoption of green IT, affected by the country in which the company is located?

The following then are the research hypotheses:

H1Null The organizational culture type has no effect on the adoption of green IT.

H1Positive The organizational culture type has a significant effect on the adoption of green IT.

H2Null The impact of organizational culture on the adoption of green IT is not affected by the country in which a company is located.

H2Positive The impact of organizational culture on the adoption of green IT is significantly affected by the country in which a company is located.

B. Survey Design

A survey was used to evaluate the organizational culture of companies and the extent to which they had adopted measures to support the adoption of green IT.

Cameron and Quinn's "Competing Values Framework" provided the theoretical underpinnings of the culture test. The associated OCAI tool was employed.

The level of greenness in the deployment of IT in the respondent's company was measured by eight questions, covering policy, strategy and practical issues, such as whether the company purchased computers with silver or gold EPEAT ratings. One question, on whether the company had a Green IT Policy in place, had possible answers yes/no/don't know. The remaining questions required responses to statements about greenness on a 5 point Likert scale, from "Strongly Agree" to "Strongly Disagree".

In addition, there were questions about demographics, asking the respondent to identify the country in which they were located, the industry sector of their company, the size of the company and their primary role in the company.

C. Survey Execution

The survey was created using Qualtrics. It was distributed electronically to contacts of the authors for onward distribution and placed on a number of forums.

The preamble to the survey included the statement: "This survey is intended to be completed by people, for whom IT (widely defined) forms a substantial part of their job function."

X. Results and Analysis

A. Demographics

There were 29 usable replies, from a range of countries as shown in Figure 7.

The highest categories for Industry Sector were Education (16) and IT. The remaining respondents came from a range of areas, including government, banking and transport.

7 respondents identified their primary roles as IT Managers; 3 as Chief Information Officers. Over one third of respondents chose the "Other" box, giving a wide range of roles.

Country	Freq	%
India	3	10.3
UK	9	31.0
US	1	3.4
France	12	41.4
China	3	10.3
Russia	1	3.4
Total	29	100.0

Figure 7. Location

Culture	Cronbach's Alpha Based on Standardized Items
Clan	0.852
Market	0.817
Adhocracy	0.773
Hierarchy	0.829
Green IT Measure	0.888

Figure 8. Cronbach's Alpha

	Minimum	Maximum	Mean	Std. Deviation
ClanTotal	0.50	80.00	25.6609	17.28918
MarketTotal	0.00	55.00	26.8333	15.58948
AdhocTotal	0.00	33.33	17.5115	9.16125
HierTotal	6.67	75.00	29.9943	17.40630

Figure 9. Culture Statistics

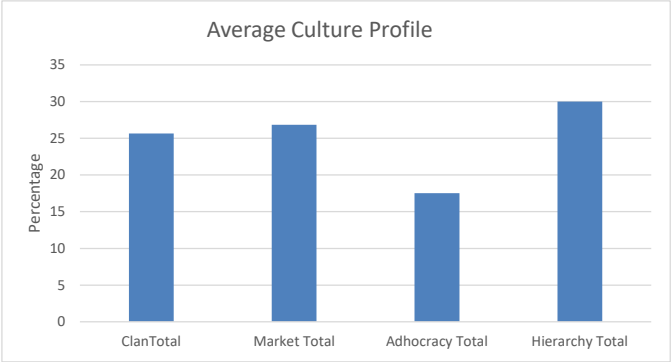


Figure 10. Culture Profile

Pearson Correlations		Level of Green IT
ClanTotal	Pearson Correlation	-.369*
	Sig. (2-tailed)	0.049
	N	29
MarketTotal	Pearson Correlation	0.151
	Sig. (2-tailed)	0.433
	N	29
AdhocTotal	Pearson Correlation	-.431*
	Sig. (2-tailed)	0.020
	N	29
HierTotal	Pearson Correlation	.458*
	Sig. (2-tailed)	0.012
	N	29

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 11

B. Reliability

We used Cronbach's Alpha measure, based on standardized items, to test reliability. Each of the six questions in the OCAI represents a test of culture, with the first option of the questions corresponding to clan culture, the second to adhocracy culture, the third to market culture and the fourth to hierarchy culture. We measured for consistency across these six questions.

Four of the questions used to measure the level of adoption of green IT required respondents to score on a five point scale. We tested for consistency across these questions.

The results are given in Figure 8. All variables met the generally recognised benchmark of being at least 0.75. It is noticeable that the score for adhocracy is significantly lower than the other culture types. This may be because this culture type is more nebulous and harder to recognise and characterize.

C. The Impact of Organizational Culture

Figures 9 and 10 give the average culture profile of the sample. It is consistent with the research of Cameron and Quinn mentioned above, with market and hierarchy being the dominant cultures. However, in Cameron and Quinn's research, market culture takes first place, whereas in the present study, hierarchy takes first place. The lower market value may reflect the relatively high proportion of people working in education.

Figure 11 applies Pearson Correlation to the four culture types with the level of Green IT in the organization. The Green IT measure has 1 as the highest level of Green IT and 5 as the lowest. Hence a negative value of the Pearson correlation coefficient indicates that a culture type is positively associated with greenness. We do not know in advance whether particular culture types will have a positive or negative influence, so we used 2-tailed correlation.

Clan and adhocracy cultures are positively associated with greenness and hierarchy culture is negatively associated. These correlations are at the 0.05 level. The smallness of the sample made it unlikely that an association at the 0.01 level would be found. No significant assocation was found for market culture.

We can therefore reject the H1Null hypothesis above and conclude that culture has an effect on the adoption of green IT.

D. The Impact of Location

Figure 12 gives the culture profile for the UK and France, which were the only countries to have a significant number of

		Minimum	Maximum	Mean	Std. Deviation
UK	ClanTotal	4.2	38.3	22.1	11.7
	MarketTotal	0.0	48.3	23.9	14.3
	AdhocTotal	4.2	33.3	19.5	10.3
	HierTotal	18.3	75.0	34.4	23.1
France	ClanTotal	10.8	80.0	34.0	21.7
	MarketTotal	5.8	50.0	26.3	17.8
	AdhocTotal	4.0	25.0	15.1	6.2
	HierTotal	6.7	48.3	24.7	14.0

Figure 12. Culture Profile by Country

			Green IT Level
UK	ClanTotal	Pearson Correlation	-0.526
		Sig. (2-tailed)	0.146
		N	9
	MarketTotal	Pearson Correlation	-0.299
		Sig. (2-tailed)	0.435
		N	9
	AdhocTotal	Pearson Correlation	-0.650
		Sig. (2-tailed)	0.058
		N	9
	HierTotal	Pearson Correlation	.742*
		Sig. (2-tailed)	0.022
		N	9
France	ClanTotal	Pearson Correlation	-0.094
		Sig. (2-tailed)	0.771
		N	12
	MarketTotal	Pearson Correlation	0.251
		Sig. (2-tailed)	0.432
		N	12
	AdhocTotal	Pearson Correlation	0.135
		Sig. (2-tailed)	0.676
		N	12
	HierTotal	Pearson Correlation	-0.233
		Sig. (2-tailed)	0.465
		N	12

Figure 13. Pearson Correlation for France and the UK

respondents. It is noticeable that the UK has a substantially lower clan and higher hierarchy score than France. This may reflect the collegiate nature of French universities, but the small number of the sample makes it impossible to draw firm conclusions.

Figure 13 gives Pearson correlations for France and the UK. The only statistically significant correlation (at the 0.05 level), between culture and greenness, for either the UK or France, was that for the UK, hierarchy culture has a negative impact on greenness. We therefore cannot reach a conclusion on the validity of the H2 hypotheses.

XI. Discussion

The literature generally indicates that clan and adhocracy cultures are positively correlated and hierarchy cultures negatively correlated, with characteristics that require openness to new practices, such as concern for green issues. Examples include [52], [53] and [50] (with regard to clan and hierarchy culture). The present study is consistent with this.

The CVF has been used in many research studies and has been shown to have a high degree of validity. Cameron and Freeman [54] and Zammuuto and Krakower [55] applied it in the field of higher education and found that culture type was a strong predictor of organizational effectiveness. These studies found evidence for concurrent validity. Quinn and Spreitzer [39] further found evidence for convergent validity and discriminant validity. However, the findings of the meta-analytic investigation of the CVF by Hartnell et al. [56], while supporting the contention that culture has a major impact on organizational effectiveness, provided only mixed support for the CVF's underlying suppositions.

There is a Likert version of the OCAI, which was used in some of the studies using the CVF, for example [51], [53] and [57]. The ipsative version forces people to make choices: giving a high score to one culture type reduces the points available for other types. With the Likert version by contrast, it is possible to give a high score to all culture types. In [57], it was found that all four culture types were positively correlated with successful use of knowledge management practices. This result would have been unlikely if the ipsative version of the OCAI had been used.

Eijnatten et al. [58] provide an interesting and extremely detailed analysis of the difference between ipsative and Likert surveys, with particular reference to the CVF. They argue that Likert surveys are norm-referenced, whereas ipsative studies are criterion referenced. For example, in an ipsative survey, a respondent might give a low score for adhocracy because they had given a high score to other culture types. Another respondent in a different company might allocate a higher score to a company with the same adhocracy characteristics, making it invalid to compare scores. Eijnatten et al. further contend that with ipsative studies only "only intra-individual, not inter-individual, comparisons are possible." This would make traditional correlation techniques invalid. Eijnatten et al. propose alternative statistical analysis techniques, involving the use of Fisher's permutation test and Aitchison distances.

Eijnatten et al. note that completing ipsative questions is more difficult that completing Likert ones, resulting in lower response rates.

It is noticeable that all the correlation coefficients for France are low, suggesting that the results of the survey are somewhat random. This may reflect the fact that the OCAI survey is linguistically quite challenging, especially for speakers of English as a second language. Or it may reflect the fact that the survey is rather rooted in the Anglo Saxon, especially US world, with its references to concepts such as "nurturing leadership".

The key weakness of this study is the small sample size. It is challenging to ensure that a survey is completed by a sample that is sufficiently large and representative to permit statistical analysis to be undertaken with a high degree of confidence. It may be worth exploring the use of web crawling to look for indications of culture and assess level of consideration of green issues.

There needs to be greater consideration of the international dimension. However, using Cameron and Quinn's Organizational Culture Assessment Instrument in an international set-

ting raises complex issues. The Competing Values Framework is located within the Anglo-American business tradition and there is a deep link between culture and language. Outside the English-speaking world, the OCAI either has to be translated or be completed by respondents using a second language. Both approaches make transnational comparison difficult.

XII. Conclusion

This paper has considered the extent to which standard management models are likely to support the adoption of green IT. It explored strategic, tactical and operational management models. It was concluded that many management models are not favourable to the adoption of green IT, in particular many of the older standard management models which do not take a holistic view of corporate responsibilities. It is, therefore, incumbent upon managers not to place excessive reliance on such models.

There is a need for the development of new management models, which more explicitly integrate traditional bottom line considerations with the wider ethical responsibilities of companies, in particular those relating to sustainability.

This paper has investigated the impact of culture, in particular organizational culture, on the success of green IT initiatives.

An international survey was undertaken on organizational culture and green IT level. The key finding is that culture has a major impact on the success of sustainability initiatives within ICT, with clan and adhocracy cultures being positively associated and hierarchy cultures negatively associated with greenness. Managers introducing sustainability initiatives must seek to understand the culture within their organization. They could seek to encourage a clan culture, for example, by using 360 degree performance evaluation where, in addition to being assessed by their managers, staff assess their managers and peers.

An investigation was also undertaken into whether the impact, if any, which organizational culture has on the adoption of green IT, is affected by the country in which the company is located. The results were inconclusive.

There was discussion of the complex technical issues which arise from the statistical analysis of culture surveys.

Future research directions include further empirical analysis of the impact of the use of management models on a sustainability culture within IT and consideration of the effect of operating within different cultures. There is also a need for development of more rigorous metrics for green IT.

References

[1] W. M. Campbell, "An exploration of the use of standard management models on the adoption of Green IT," in Proceedings of Green 2017: The second International Conference on Green Communications Computing and Technologies.

[2] W. M. Campbell, M. Ratcliffe, and P. Moore, "An exploration of the impact of organizational culture on the adoption of green IT," in Proceedings of the Green Computing and Communications Conference (Green-com), 2013, pp. 68–76.

[3] W. M. Campbell, M. Ratcliffe, P. Moore, and M. Sharma, "The influence of culture on the adoption of green IT," in Green Services Engineering,Optimization and Modeling in the Information Age, X. Liu and Y. Li, Eds. Earthscan, London, 2015, pp. 25–60.

[4] L. Belkhir and A. Elmeligi, "Assessing ICT global emissions footprint:trends to 2040 and recommendations," Journal of Cleaner Production, 2018.

[5] M. Menguc and L. Ozanne, "Challenges of the green imperative: a natural resource-based approach to the environmental-business performance relationship," Journal of Business Research, vol. 58, 2005, pp. 430–438.

[6] J. Elkington, "Towards the sustainable corporation." California Management Review, vol. Winter, 1994, pp. 90–100.

[7] ——, "Enter the triple bottom line," in The Triple Bottom Line: Does It All Add up?, A. Henriques and J. Richardson, Eds. Earthscan, London, 2004, pp. 1–16.

[8] W. E. Stead and J. G. Stead, Management for a small planet: Strategic decision making and the environment. Newberry Park, CA: Sage, 1992.

[9] J. E. Post and B. W. Altman, "Managing the environmental change process: Barriers and opportunities," Journal of Organizational Change Management, vol. 7, no. 4, 1994, pp. 64–81.

[10] G. H. Brundtland, "Our common future," in Report of the World Commission on Environment and Development. Oxford University Press, 1987.

[11] U. Nations, "Paris agreement," in Framework Convention on Climate Change. United Nations, 2016.

[12] T. Jenkin, J. Webster, and L. McShane, "An agenda for 'green' information technology and systems research," Information and Organization, vol. 21, 2011, pp. 17–140.

[13] R. T. Watson, M. C. Boudreau, A. Chen, and M. H. Huber, "Green IS: Building sustainable business practices," in Information Systems. Athens, GA, USA: Global Text Project, 2008.

[14] V. Dao, I. Langella, and J. Carbo, "From green to sustainability: Information technology and an integrated sustainability framework," Journal of Strategic Information Systems, vol. 20, 2011, pp. 63–79.

[15] DeutscheBank, "Green IT: Energy efficiency at work," https://www.db.com/cr/en/concrete-green-it.htm, Accessed 14 August 2018.

[16] A. Bokolo, M. Majid, and A. Romli, "Organizational culture and leadership: Preconditions for the development of a sustainable corporation," Sustainable and Applied Information Technology, vol. 95, no. 9, 2017, pp. 1875–1915.

[17] S. Murugesan and G. R. Gagadharan, Eds., Harnessing Green IT: Principles and Practices. Wiley, 2012.

[18] A. Molla and V. Cooper, "Enterprise Green IT readiness," in Harnessing Green IT: Principles and Practices, S. Murugesan and G. R. Gagadharan, Eds. Wiley, 2012.

[19] Q. Deng and S. Ji, "Organizational Green IT adoption: Concept and evidence," Sustainability, vol. 7, 2015.

[20] B. Wernenfelt, "A resource based view of the firm," Strategic Management Journal, vol. 5, 1984, pp. 171–180.

[21] S. L. Hart, "A natural-resource based view of the firm," Academy of Management Review, vol. 20, no. 4, 1995.

[22] FinancialTimes, "Where did FT500 chief executives go to business school?" https://www.ft.com/content/3a63c054-b885-11e5-b151-8e15c9a029fb, Accessed 3 June 2019, Published January 2016.

[23] M. VanAssen, Key Management Models: The 60+ models every manager needs to know. Prentice Hall, 2009.

[24] H. I. Ansoff, Corporate Strategy. Penguin Books, 1987.

[25] C. Christensen, The Innovator's Dilemma: Whe New Technologies Cause Great Firms to Fail. Harvard Business Review Press, 1997.

[26] D. Bilgeri, E. Fleisch, and F. Wortmann, "How the IoT affects multibusiness industrial companies: IoT organizational archetypes," in Proceedings of the Thirty Ninth Conference on Information Systems, San Francisco 2018.

[27] M. Holopainen and M. Toivonen, "Weak signals: Ansoff today," Future, Elsevier, vol. 44, 2012, pp. 198–205.

[28] D. C. Hambrick, I. C. Macmillan, and D. L. Day, "Strategic attributes and performance in the BCG matrix," Academy of Management Journal, vol. 25, 1982, pp. 510–531.

[29] W. C. Kim and R. Mauborgne, Blue Ocean Strategy. Harvard Business School Press, 2005.

[30] C. Dupont, M. Vecchio, C. Pham, B. Diop, C. Dupont, and S. Koffi, "An open IOT platform to promote eco-sustainable innovation in Western Africa: Real urban and rural testbeds," Wireless Communications and Mobile Computing, 2018.

[31] C. Gao, Y. Ji, J. Wang, and X. Sai, "Application of blockchain technology in peer-to-peer transaction of photovotaic power generation," in 2nd IEEE Advanced Information Management, Communications, Electronic and Automation Control Conference, 2018.

[32] J. Kay, Foundations of Corporate Success: How business strategies add value. Oxford University Press, 1993.

[33] K. Cameron and D. R. Ettington, "The conceptual foundations of organizational culture," in Higher Education: Handbook of Theory and Research, J. C. Smart, Ed. Norwell, Mass.: Kluwer, 1988, vol. 4.

[34] E. Schein, Organizational Culture and Leadership. Jossey-Bass, 1992.

[35] K. Cameron, D. Bright, and A. Caza, "Exploring the relationship between organizational virtuousness and performance," American Behavioral Scientist, vol. 47, 2004, pp. 766–790.

[36] G. Hofstede, Culture's Consequences. SAGE, 1980.

[37] R. Quinn and J. Rohrbaugh, "A spatial model of effectiveness criteria: Toward a competing values approach to organizational analysis," Management Science, vol. 29, 1983.

[38] K. Cameron and R. Quinn, Diagnosing and Changing Organizational Culture Based on the Competing Values Framework. Wiley, 2011.

[39] R. Quinn and G. Spreitzer, "The psychometrics of the competing values culture instrument and an analysis of the impact of organizational culture on quality of life," in Research in Organizational Change and Development, R. W. Woodman and W. A. Passmore, Eds. Greenwich Conn.: JAI Press, 1991, vol. 5.

[40] O. Richard, A. McMillan-Capehart, S. N. Bhuian, and E. C. Taylor, "Antecedents and consequences of psychological contracts: Does organizational culture really matter?" Journal of Business Research, vol. 62, 2009, pp. 818–825.

[41] B. T. Gregory, S. G. Harris, A. A. Armenakis, and C. L. Shook, "Organizational culture and effectiveness: A study of values, attitudes and organizational outcomes," Journal of Business Research, vol. 62, 2009, pp. 673–679.

[42] M. Linnenluecke and A. Griffiths, "Corporate sustainability and organizational culture," Journal of World Business, vol. 45, 2010, pp. 357–366.

[43] M. Beer and N. Nohria, Breaking the Code of Change. Harvard Business School Press, 2000.

[44] J. C. Henderson and N. Venkatram, "Strategic alignment: A model for organizational transformation via Information Technology," Publication of Centre for Information Research, Sloan School of Management Massuchusetts Institute of Technology, 1990.

[45] ——, "Understanding strategic alignment," Business Quarterly, vol. 55, no. 3, 1991, pp. 12–89.

[46] ——, "Strategic alignment: Leveraging information technology for transforming organizations," IBM Systems Journal, vol. 32, no. 1, 1993, pp. 4–16.

[47] T. Coltman, P. Tallon, R. Sharma, and M. Queiroz, "Strategic alignment: twenty-five years on," Journal of Information Technology, vol. 30, 2015, pp. 91–100.

[48] F. Loeser, K. Erek, N.-H. Schmidt, R. Zarenkow, and L. M. Kolbe, "Aligning Green IT with environmental strategies: Development of a conceptual framework that leverages sustainability and firm competitiveness," in AMCIS 2011 Proceedings.

[49] J. P. Cotter, Breaking the Code of Change. A Force for Change: How Leadership differs from Management, 1990.

[50] B. Hribar and J. Mendling, "The correlation of organizational culture and and success of BPM adoption," in Proceedings of the Twenty Second European Conference on Information Systems, Tel Aviv 2014.

[51] C. Chin-Loy and B. G. Mujtaba, "The influence of organizational culture on the success of knowledge management practice with North American companies," International Business and Economics Research Journal, 2007.

[52] Ülle Übius and R. Alas, "Organizational culture types as predictors of corporate social responsibilty," ISSN 1392-2785 Engineering Economics, vol. 1, no. 61, 2009.

[53] A. Akano and W. M. Campbell, "A Cross-cultural Survey of the Impact of Organizational Culture on Adoption of Green IT," in Proceedings of the Eighth International Conference on Complex, Intelligent and Software Intensive Systems 2014 (CISIS 2014). Birmingham UK, 2014.

[54] K. Cameron and S. Freeman, "Cultural congruence, strength and type: Relationships to effectiveness," in Research in Organizational Change and Development, R. W. Woodman and W. A. Passmore, Eds. Greenwich Conn.: JAI Press, 1991, vol. 5.

[55] R. F. Zammuto and J. Y. Krakower, "Quantitative and qualitative studies of organizational culture," in Research in Organizational Change and Development, R. W. Woodman and W. A. Passmore, Eds. Greenwich Conn.: JAI Press, 1991, vol. 5.

[56] C. A. Hartnell, A. Y. Ou, and A. Kinicki, "Organizational culture and organizational effectiveness: A meta-analytic investigation of the competing values framework's theoretical suppositions," Journal of Applied Psychology, vol. 96, no. 4, 2011, pp. 677–694.

[57] K. Chidambaranathan and S. BS, "Analysing the relationship between organizational culture and knowledge management dimensions in higher education libraries," Journal of Librarianship and Information Science, 2017.

[58] F. M. Eijnatten, L. van der Ark, and S. S. Holloway, "Ipsative measurement and the analysis of organizational values: an alternative approach for data analysis," Quality and Quantity, vol. 49, 2015.

Management Guidelines for Better Application of Business Process Management in SAP ERP Projects

Markus Grube

VOQUZ IT Solutions GmbH
Hamburg, Germany
e-mail: markus.grube@voquz.com

Martin Wynn
School of Business and Technology
University of Gloucestershire
Gloucester, UK
e-mail: mwynn@glos.ac.uk

***Abstract* — The SAP Enterprise Resource Planning (ERP) system is a leading software solution for corporate business functions and processes. Business Process Management (BPM) is a management approach designed to create and manage organizations' business processes. Both promise an improvement of business processes in companies and can be used together in organizations. In conjunction with the SAP ERP system and BPM approach, BPM maturity models can be used as diagnostic tools that allow an organization to assess and monitor the maturity of its business processes. The aim is to investigate and analyse the interaction between the use of the SAP ERP software package and the utilization of BPM. Findings derive from an analysis of eleven semi-structured expert interviews and a validation of the guidelines via an online survey with 151 participants that use SAP, BPM and/or BPM maturity models. This research analyses the complex relationships between SAP, BPM and BPM maturity models and develops management guidelines for improved use of BPM in SAP ERP projects.**

Keywords* — *SAP; ERP; BPM; Business Process Management; Maturity Models; BPM Maturity Models; Management Guidelines.

I. INTRODUCTION

This article builds upon previously published research [1] to provide a more detailed presentation and assessment of a set of management guidelines for the utilization of BPM in SAP ERP projects. SAP (Systeme, Anwendungen und Produkte) is a German company created in 1972 and the world's largest provider of enterprise software that, in 2019, had more than 437,000 customers in over 180 countries [2]. The SAP ERP package provides software solutions for the full range of business functions in companies – from human resource management, back office finance processes, the full sales order processing cycle and manufacturing, supply chain and distribution functions [3]. SAP ERP is usually installed on a database platform that handles several different business functions supported by the range of software modules that make up the SAP product suite. The implementation of an ERP system is often seen as a major strategic investment that may encompass innovative change, which may enhance a company's core capabilities [4].

BPM is a methodology for the definition and operation of company business processes, and can be used without any information technology (IT) systems or infrastructure [5]. In practice, however, companies often use IT software tools to administer the BPM of an organisation. Additionally, software such as SAP ERP can assist a company in standardizing and automating processes to make them as efficient as possible [6]. BPM will usually start with a process analysis of the actual business processes [7] by the application of specific methods, techniques and tools [8]. The following definition is assumed in this research: BPM is a process-oriented management approach to create, support and analyse an organisation's business processes. It is also worth noting that, for an optimum BPM outcome, an IT tool - to analyse and support BPM behaviour – may be usefully deployed.

BPM maturity models can be used to support the application of BPM. A maturity model is described by Saco [9] as a diagnostic tool for an organization, which provides a framework to test, analyse and improve business quality [10]. A BPM maturity model analyses the quality of a company process and classifies business quality into different levels such as 'Initial', 'Repeatable', 'Defined', 'Managed' and 'Optimizing' [11]. Each maturity model has formal descriptions to guide the use regarding how to reach the next level of a maturity. The purpose of such models is to reach the highest maturity level for all organisation processes [12].

Fig. 1 illustrates some of the concepts that are shared by the SAP ERP system, BPM and BPM maturity models. The SAP ERP system includes its own business process models; BPM has the objective of improving business processes in an organisation; both have the same aim of optimizing an organisation's processes.

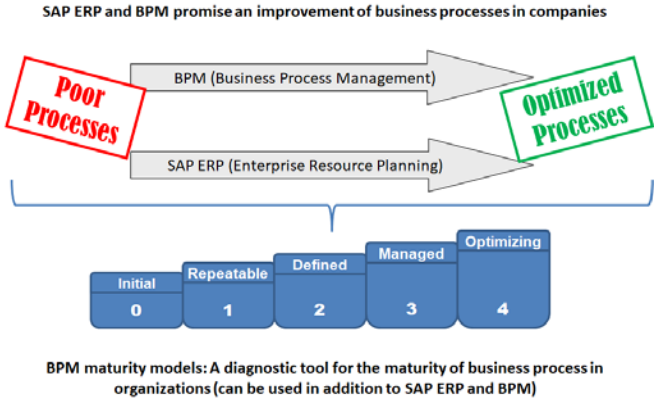


Figure 1. SAP ERP, BPM and BPM maturity models: overview

Diagnostic tools like BPM maturity models can be used alongside SAP ERP and BPM to measure the effectiveness of processes in organisations. Van Looy [13] states that most BPM maturity models favour the use of an IT system (such as SAP ERP) to improve the BPM approach of an organisation [14]. This study examines the relationship between these concepts and derives management guidelines for their use. These guidelines are based on views expressed in one-to-one interviews combined with feedback from online surveys and are geared to supporting to business managers.

Following this introductory Section, Section II discusses relevant literature. In Section III, the research methodology is outlined, and this is followed in Section IV by a summary of findings. Section V presents a further analysis of findings leading to the development of a number of management guidelines to support the use of BPM maturity models within an SAP systems environment. Section VI then details how these guidelines have been validated through an online survey and finally, in Section VII, the main themes of the paper are drawn together to provide overall conclusions regarding the research project.

II. Literature Review

BPM and process integration have been discussed for over 25 years [15], but existing literature is largely confined to general findings about the relationship between IT and the use of business processes, or about the relationship between ERP systems and business processes. For example, vom Brocke et al. [16] explain that the selection, acceptance and use of IT are a fundamental part of BPM, and Wynn [17] highlights the importance of a range of process issues in achieving successful ERP project implementation. Business and IT need to connect with each other in order to realize better business value. Neubauer [18] also notes that ERP systems generally influence a company's business processes.

Saco [9] explains that a maturity model is a diagnostic tool for an organisation to improve its processes. Such assessment can be used in conjunction with SAP ERP and BPM. Most authors view the use of an ERP system as a means of integrating business processes within one system, which is, used company-wide [18]. For example, an ERP system can hold all documents in relation to an invoice number or purchase order, and can show the document flow or action log for data changes that directly belong to a business transaction. Through the use of ERP systems, companies are expected to reduce costs by improving efficiencies and widening the availability of accurate and up to date business information, thereby enhancing overall company performance [19]. Antonucci et al. [20] indicate that ERP systems produce the data and information that are the basis for business decisions and strategies.

The extant literature demonstrates that an IT application like SAP ERP can enable higher process maturity [13]; but these studies focus on the general company level and IT systems as a whole. Van Looy [13] suggests the deployment of IT for business process maturity, concluding that most maturity models recommend IT to improve process modelling and optimization. She emphasises that, in general, IT deployment enables higher process maturity. Other literature illustrates that BPM maturity models measure and aggregate capabilities that can culminate in a road map for better business process management [21].

The literature research confirms that companies often simply apply maturity models blindly instead of addressing organizational needs [22]. Additionally, maturity models do not usually consider any link between, for example, an SAP system and the BPM approach. This research develops a set of guidelines embodied in management guidelines for the better collaboration of BPM in SAP ERP projects – this being in addition to, or instead of, the use of a BPM maturity model.

III. Research Methodology

Fig. 2 depicts the main elements of the research methodology used in this project, selected from a body of methods that can be used to gather and process data [23] [24].

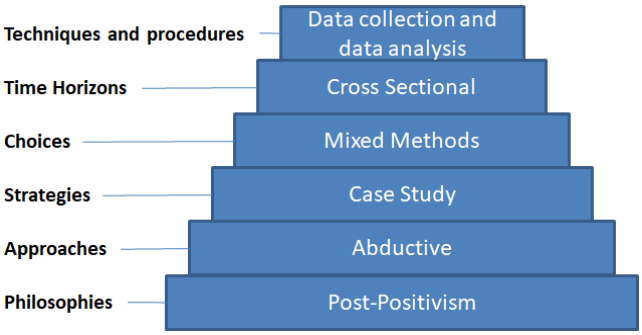


Figure 2. Research Methodology Layers [22] [23]

The research philosophy adopted here is post-positivist, based on the perspectives of Ryan [25] and Guba [26]. The goal of post-positivist research is the generation of "new knowledge that other people can learn from and even base decisions on" [27]. The development of management guidelines emanates from this research and aims to support a company for informed decisions in the context of SAP and the BPM approach. The post-positivist position takes the view that the world is much more complex than when the project was embarked upon, and that it is quite possible that, for example, the inclusion of other experts in the interview process would have led to different results. This research starts with some cases on organisations that use BPM maturity models, and concludes with generalised guidelines about the application of BPM and SAP ERP. As described by Thomas [28], this exploration uses the abductive approach to collect facts from the interviewees, followed by judgment about the best explanation of these, before producing a generalized output in the form of guidelines and management guidelines.

As an explanatory study [29], this research investigates the relationship between SAP, BPM, and BPM maturity models in ERP projects. Based on semi-structured expert interviews, a small number of organisations are examined in depth. In line with the use of documentation of BPM maturity models, a qualitative research approach was pursued. According to Saunders et al. [24], the use of different data sources in combination with secondary data collection techniques allows a form of triangulation to confirm the obtained data from the

interviews. The aim of this work is to evaluate whether SAP can affect the use of BPM in companies.

Through the use of semi-structured interviews with experts in their field, and the analysis of secondary literature such as user manuals of BPM maturity models, this research used mixed-methods to address the research objectives, thereby providing greater depth in a complex environment [30]. The time horizon for this research was a cross-sectional snapshot study [24]. The research analyses the current SAP impact on BPM and BPM maturity models in practice, and evaluates the picture at the time of the study [31].

The interviewees were selected with the objective of gleaning the greatest amount of expert knowledge possible from practice. The semi-structured interviews allowed a degree of flexibility that engendered an understanding and explanation of the experts' opinions regarding important issues, events and patterns in the complex interaction of SAP, BPM, and BPM maturity models in ERP projects [32]. The software tool MAXQDA was used for the qualitative data analysis and comparison of the interviews, and in arranging, organising and analysing all transcribed interviews, and also for analysing secondary literature sources. This allowed a special type of methodological triangulation through the use of more than one method to collect and analyse the data. A thematic analysis was used for the identification of topics. For this purpose, statements from the interviews were manually coded in order to recognize and interpret connections [32].

The final step of the data triangulation of this study examines, through an online survey, the guidelines that were previously developed. The online survey was created to consider questions about the developed guidelines and to confirm their acceptance. Potential participants for this online survey were users, process managers, researchers and consultants with a business background encompassing SAP, BPM and/or BPM maturity models in different industry sectors.

Fig. 3 depicts the research strategy and the different data sources for this research.

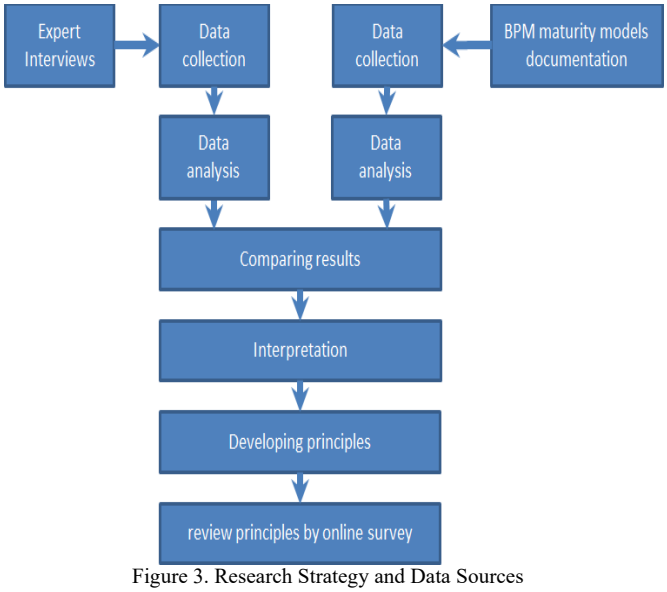


Figure 3. Research Strategy and Data Sources

IV. FINDINGS

The search for potential interview partners was a difficult process and resulted in many rejections. The search utilised existing networks of business and personal contacts, resulting in 64 people in Germany, Austria and Switzerland being identified as potential experts, who were then invited for an interview. From this initial pool, eleven people confirmed that they were willing to be interviewed for this research project. Most refusals were based on the fact that the experts did not have the necessary practical experience of the use of a BPM maturity model. Nevertheless, three of the experts interviewed are currently using no BPM maturity model, but perform some form of quality process assessment already at their company, or would like to apply a BPM maturity model in the future.

Baker and Edwards [33] explain that, within qualitative research, the attainment of a sufficient quantity of interviews cannot be set at a certain number. It is crucial to achieve saturation and try to gain new knowledge through additional interviews. In this research, there appeared to be a degree of saturation with the tenth interview, as no new knowledge surfaced. Interview findings are discussed in more detail below.

A. *Management*

Almost every interviewee recommends that senior management should be involved in driving forward the introduction of BPM, BPM maturity models and even SAP ERP. The introduction of these technologies and techniques can only be successful if the management proactively supports their introduction and operation. The motivation to increase the degree of process maturity has more influence if the top management is involved. It is important for employees as a whole to support their implementation, but they must also know that the management supports these initiatives and will provide the budget, time and resources for these projects. Management must actively support initiatives such as the introduction of BPM and SAP software implementation if they are to be successful.

B. *Standard SAP processes*

The introduction of an SAP system can be used strategically to remodel business processes. Each company should consider whether a standardisation of business processes is an important thing for their organisation. The application of standard processes and especially SAP standard processes could be helpful. The use of SAP does not generally lead to improved process maturity levels, but this depends on how intensively and professionally SAP is deployed as a corporate-wide business system. The practical experience of the interviewees shows that the use of standard processes can be very beneficial, reducing time, costs, resources and other expenses. Many processes are similar across different industries and therefore many experts recommend the use of the standard processes which are provided by an SAP ERP system.

However, the use of standard processes never releases a company from the obligation to precisely analyse these processes and test them against their own requirements. Each company must agree a procedure and determine how far a

company should adapt to a standard process or whether the process must modified to accommodate the company's requirements. With sufficient time and money, an SAP system can be customized to meet the process requirements of a company. In practice, however, the budget is often quite limited and a company must determine whether standard processes should be adopted or not.

C. Process analysis and documentation

The introduction of any significant process change should always be analysed, documented and understood, whether it is an SAP standard solution or a customer driven solution. In addition, a well-documented interface description (of both the software modules and the processes themselves) can provide a solid basis for the development of process improvements. Frequently used processes should be analysed more thoroughly, and standard processes should be analysed carefully because these processes can be used in an incorrect manner.

Some interviewees suggested that some of the more traditional and well-established business processes, such as those supporting finance or human resource management, can be represented very well by using standard SAP software. This is because they are core company functions, and the software has been implemented in many different industry sectors.

D. Interfaces

The interviewees were of the view that it is useful to consider the interfaces to and from an SAP system in more detail. They suggested that a process that is supported by one system is likely to have a higher maturity than processes that run over several systems boundaries. Frequently, in this case, not only different systems but also different employees are affected by process deficiencies, in part caused by the need to understand and handle several systems that may not be well integrated.

Processes should always be considered as a set of end-to-end activities. The more IT systems that are affected, the more employees are usually involved. These employees must then find a way to identify and analyse which data are necessary to support process improvements. All systems interfaces need to be assessed, and a well-documented interface description can provide a solid basis for the development of process optimisation. Even with standard interfaces or online services which are provided by SAP, a successful BPM project will still need to analyse and document these interfaces.

E. Process maps

All significant business processes should be described and depicted within an overall top-level company process map. Regardless of whether processes are supported by an IT system or not, they should be recorded and drawn in order to better understand the process flow and to store knowledge about the process. The modeling of a process map should be a basic requirement to introduce efficient BPM in a company. Of course, this is alone not enough and more activities will be required which deal with the establishment of process management in the company.

A significant risk - that must be avoided - is that a BPM team design a process map, which differs from the process map owned by the SAP project team operating within the same company. This can sometimes happen when different initiatives are introduced in different management areas. This may result, for example, in the purchasing process being designed both within the BPM project team and in the SAP team. This means that two teams will have independently developed this process with their own experts, with a knock-on effect on subsequent remodeling and overall costs. The BPM team should know what knowledge is already present in the company and be ready to build on it.

F. Measurement of key performance indicators (KPIs)

It is generally worth considering whether it is more effective in the long term to measure results generated directly from SAP reporting tools rather than via specialist standalone business intelligence software. It is even quite possible to effectively use an Excel spreadsheet or Access database for data analysis. In many cases, such a solution is only intended for the short term. But if data has to be specially prepared and processed for export to an Excel file, for example, it is likely to be more effective to analyze it immediately in the source system (such as SAP). This raises the issue of end-user access to SAP data and its reporting tools, which needs to be proactively and sensitively managed.

The experts reported that financial transactions usually constitute key information that is well supported within a company deploying the SAP system. This means that, for many companies, SAP is an important strategic IT component. Consistent business intelligence reporting and associated analytics is another benefit from using SAP as the core company IT system.

V. Analysis

For this research, the BPM approach encompassed a strong focus on IT systems and infrastructure. Even though, in theory, IT should play only a subordinate role, the expert interviews conveyed a contradictory opinion. All the interviews, which involved practitioners, confirmed that the SAP and BPM concepts are closely related. Theoretically, there is often no such link found in the documentation, but in practice the SAP system is the leading ERP system in many companies and therefore there is a practical connection.

Maturity models are already very complex, but companies are often interested in guidelines that are less complex and require a smaller budget. The guidelines developed in this research project will help companies to analyse and understand the interconnection of the SAP ERP system with the BPM approach. The goal is not to develop a more complex and comprehensive maturity model; indeed the success of easy to use maturity models is due to the fact that those models has less criteria and are easy to handle. The experts explained within the interviews that many companies prefer a checklist instead of a complex maturity model. For these reasons, it is not necessary to develop a separate and new BPM maturity model to understand and show possible dependencies.

On the basis of the interview analysis, the following SAP-specific guidelines have been developed to enable company

management and all relevant stakeholders to determine and understand possible connections between an SAP system and a BPM approach. These guidelines can be used to maximise the benefits of both the SAP system and BPM methods and techniques. They are not meant to be comprehensive, but are rather intended to allow the practitioner to start to think about the connections and to develop them later on for specific environments as appropriate.

Guideline 1: Ensure that management fully support the optimal deployment of SAP in the organisation.

The use of SAP ERP as the central IT software system within a company is usually a strategic decision. In this case, the company should decide how to integrate the system with the adopted BPM approach of the company. What does the SAP specification imply? Does that mean that only key figures have to be generated from the SAP system? Could there be other systems besides the SAP system? Should a company use as many standard SAP processes as possible? The company must determine who decides possible solutions or any adaptations of the SAP system. The successful implementation of an SAP system is only possible if the senior management are aware of and confront these issues.

Guideline 2: Establish as many SAP ERP standard processes as possible at the company in order to minimize the complexity of system upgrades or enhancements.

If the company wants to use SAP, and the management supports this, then companies should also decide whether, and to what extent, standard SAP processes should be used. The use of standard SAP processes reduces the time, cost, resources and other operational constraints, and supports the introduction of new SAP enhancement packages or release changes. Each change or upgrade makes it necessary to test customised solutions and adjust the customer-specific programming to the new version of the SAP system. It is important to prioritize when the standard SAP processes should be used, and when it is better to use self-defined solutions. A BPM team should not accept processes as given and must analyse which approach is best suited to a specific company environment. Not all standard processes are the optimal solutions for every company, and a company should not necessarily submit to the dictates of a rule-based IT system. However, the use of standard process solutions could also be very helpful and reduce the budget required to operate an IT system. Regular consideration should be given to whether IT innovations in the system could lead to process improvements. For example, mobile device applications can now operate in conjunction with SAP modules, and such mobile functionality is now integrated into the standard SAP system.

Guideline 3: Ensure that all processes have been documented, analysed and understood, even if they are pre-defined by the SAP system.

The use of SAP standard processes does not absolve a company from the duty to document, analyse and understand each process. It can be the case that standard processes, which run in a single system like SAP, run with an optimised composition, and are better coordinated than other processes; but each process should be analysed. Unfortunately, it is not always obvious which data is being stored and used within an SAP process. Technically, it is currently not possible to get a fast and actual process flowchart from an existing SAP system, and see how customizing settings within an SAP system may change a process flow. Therefore, it is very important to understand and analyse these SAP processes in detail. This is the only way to avoid incorrect or error-prone process operations. A company should know exactly how its processes are running, and therefore a company should not be dictated to by an IT system or by the opinion of an ERP system provider. An analysis of each pre-defined process should enable a company to decide whether these standard processes are usable, or whether an individual process should be developed for their specific company environment.

Guideline 4: Establish a procedure that ensures that all interfaces are analysed for their BPM relevance, including those between non-SAP systems and interfaces between those systems and the SAP system itself.

Interfaces between different systems often offer opportunities for systems optimization and process improvement. Many experts recommend considering the processes from an end-to-end perspective. They have learned from their practical experience that, especially when there are different systems with bespoke interfaces, data are often transmitted in formats that differ from that which is required. It is important to analyse the standard interfaces provided by the software provider, which may not be the best and optimal for the user organisation.

Guideline 5: Ensure that all teams within a company, especially the BPM team and the SAP team, contribute to the development of the same processes and process maps, and that only one process map exists within the organisation.

SAP is a very powerful tool that communicates with many different sub-modules and other systems. The early versions of SAP had a functional structure but with the application of BPM, the package is now more process-oriented in design. It is important to avoid different teams working in isolation and developing different process configurations within a company. The BPM team should consist of a variety of different stakeholders, to represent different requirements and knowledge inputs.

Guideline 6: Ensure that all key information is generated directly from the SAP system.

SAP provides many instruments for the generation and monitoring of KPIs and most BPM maturity models encompass the analysis of KPIs. For many experts in this study, the SAP system was often the leading financial system

in their company contexts. This offers many advantages for the analysis of KPIs. Much financial information is already stored in the SAP system, which can be used to support the BPM approach. Some companies, when trying to implement quick solutions or consultancy generated analysis, may turn to creating Excel spreadsheets rather than using the "one view of the truth" available in the SAP system. SAP provides many predefined reports, and can employ business intelligence tools to provide customised reports from the SAP database. It may take longer to determine the required fields for an analysis within the SAP system, but for frequent use, it is much faster to retrieve the numbers directly from the SAP system.

VI. VALIDATION

The guidelines discussed above provide a basis on which companies can make some judgements about the use of their SAP system when combined with a BPM approach. For further practical confirmation of the guidelines, an online survey was used to validate the general applicability of the developed guidelines and the general feasibility of the findings of the interviews. A questionnaire was presented as an online survey on the internet, allowing the collection of a larger amount of data from more participants in a shorter time and in a more flexible manner than the personal face-to-face interviews [24].

Potential participants for this online survey were those such as users, process managers, researchers and consultants with a business background in SAP process management or BPM in general. Respondents had to have several years of practical experience in at least two of the three investigated subject areas of SAP, BPM and BPM maturity models. In general, the online survey was created in a way that allowed each participant to answer the questions themselves in the form of a self-completion survey. Every question about the evaluation of the guidelines was followed by an open question in which the participant had the opportunity to make a comment about the established guideline or to list possible improvements about that guideline.

To facilitate a general classification of the participants, the first questions in the web survey asked about the personal background of the participant. This revealed that the participants have many years of experience in the field of SAP and BPM and come from different industries.

Table I shows the average experience of all participants in the three areas of expertise, and Table II illustrates the different industry sectors in which the participants work. This indicates that the participants represent the opinions of a variety of different industry sectors.

.

Table I. Years of experience

Area	*Average years of experience*
SAP	12.41 years
BPM	9.37 years
BPM maturity models	4.35 years

Table II. Industry sector

Industry sector	No. of Persons
Automation	1
Automotive	3
Aviation	18
Construction	1
Engineering	9
Finance	7
Food	1
Health Service	2
Insurance	1
IT	42
Logistics	2
Management Consulting	42
Medical engineering	2
Pharma	2
Power	2
Production	9
Telecommunications	2
University	5

The online survey was intended to be answered by business professionals, and so participants were asked to assign themselves to a given professional position. Table III reveals that over 82 percent of all respondents claimed that they work as users, (process) managers, researchers or consultants within an organisation.

Table III. Current position of respondents

Current Position	*No. of Persons*
Consultants	68
(Process-) Managers	44
(System-) Users	8
Researchers	5
Others	26

The main purpose of the online survey was to achieve a degree of validation of the guidelines, and to ascertain whether these guidelines were considered practicable by the business community. All guidelines were evaluated using a Likert Scale approach [34], establishing whether the participant agreed to the guideline or not by using the following classification.

Table IV. Classification

Classification
Agree Strongly
Agree
Disagree
Disagree Strongly
Don't know

This form of classification provides a simple validation method that has been applied to each of the six guidelines. The questions regarding the evaluations of the guideline also offered the answer 'don't know' to show that an answer did not have to be given. The literature illustrates that closed questions exhibit certain disadvantages if, for example, the question irritates the respondent or they want to explain their answer [32]. For this reason, every question about the evaluation of the guideline was followed by an open question in which the participant had the opportunity to make a comment about the established guideline or to list possible improvements regarding that guideline.

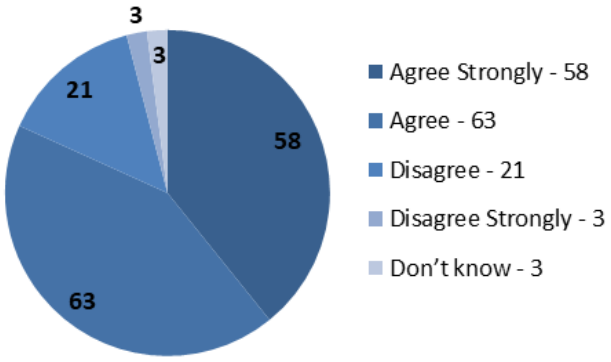


Figure 4. Guideline 1: Ensure that management fully support the optimal deployment of SAP in the organization - online responses

More than 81 percent agreed with the first guideline by selecting 'Agree Strongly' or 'Agree' and confirmed that the management should support the use of SAP in the organisation to the full extent. (Fig. 4) Some comments suggested that management could more effectively set out 'framework conditions', e.g., a basic systems or requirements specification. However, when it comes to the financial parameters of a system implementation, respondents were clear that the management must determine this scope. Similarly, over 85 percent agreed with the second guideline: establish as many SAP ERP standard processes as possible at the company in order to minimise the complexity of system upgrades or enhancements (Fig. 5). Participants noted that it is not easy, in some circumstances, to decide between a standard process or a process designed and customised to suite a company's specific requirements. The guideline does not indicate that standard processes must be used, but rather that as many as makes overall sense should be used. The decision on what to use must still be made by the user. Respondents noted that experience and knowledge play an important role in the implementation of processes.

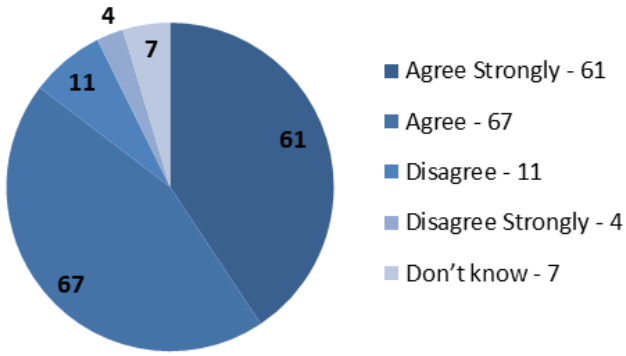


Figure 5. Guideline 2: Establish as many SAP ERP standard processes as possible to minimize the complexity of system upgrades - online responses.

The third guideline had one of the highest approval ratings in the survey (Fig. 6). Participants considered that it is important to analyse processes and that there exist different methods to support process descriptions or process depictions.

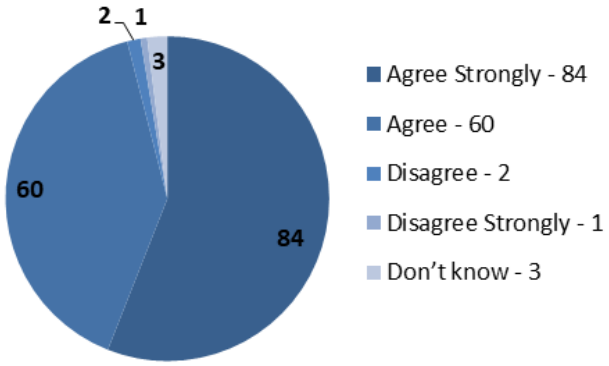


Figure 6. Guideline 3: Ensure that all processes have been documented, analysed and understood - online responses

It is very difficult to find the right level of detail for the process documentation and every company has to find its own way. It was very important for the respondents that the processes were regularly analysed.

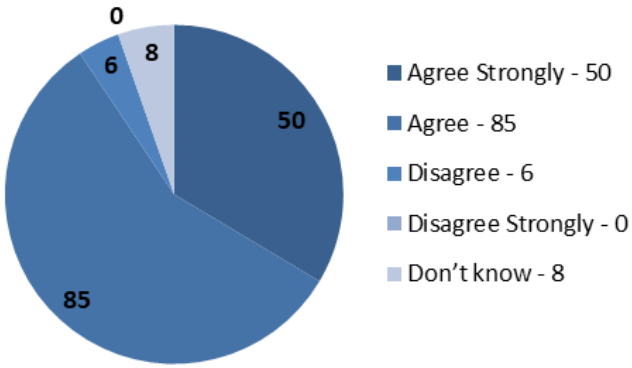


Figure 7. Guideline 4: Establish a procedure that ensures that all interfaces are analysed for their BPM relevance - online responses

In the context of the fourth guideline, the survey responses confirmed that additional systems and interfaces are likely to be necessary and need to be analysed for their BPM relevance (Fig. 7). An interface can be very maintenance intensive and should therefore be regularly analysed and tested. The participants referred to the use of standard processes in connection with this guideline, but this is already evidenced in guideline 2: a regular assessment and analysis of systems interfaces is important.

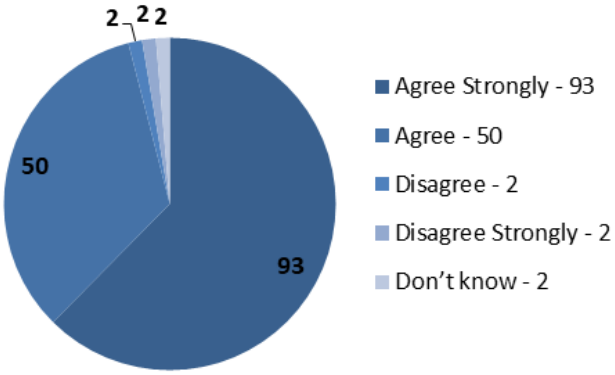


Figure 8. Guideline 5: Ensure that all teams within a company contribute to the development of the same processes and process maps - online responses

The fifth guideline also received a high approval of over 95% (Fig. 8). Participants agreed that an attempt should be made to adhere to this guideline, although this is not always possible in practice. For example, there may be more than one BPM team in a company, and inter-team communication and interaction are of the utmost importance. Political factors may also impede the pursuit of this guideline, and the survey results suggest that, contrary to the interview evidence, it can - in certain circumstances - make sense to have more than one process map for a short time. In in the end, however, these process maps should be combined and any contradictions resolved.

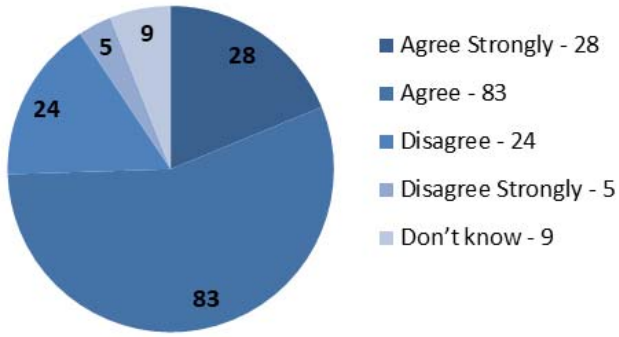


Figure 9. Guideline 6: Ensure that all key information is generated directly from the SAP system - online responses

The sixth guideline had marginally the lowest overall support, but 75% were still in agreement (Fig. 9). This guideline assumes that SAP is the main system in the company, but some survey respondents pointed out that SAP is sometimes just a financial management system and that key data is often created in other systems, and they thus did not share this assumption. They also suggest that only the KPIs that are really needed should be reported on, and that this is often fewer than initially thought.

VII. Conclusion

One view evident in the existing literature is that no specific IT system should determine the design and deployment of BPM tools and techniques, but should simply support the business transactions of a company [35]. However, the practical experience of the experts interviewed in this research provides a different perspective. Neubauer [18] asserts that ERP systems can influence a company's business processes and this is confirmed by this research as regards the SAP ERP system. All the interviews, which involved many practitioners, confirmed that the SAP and BPM concepts are closely related. Theoretically, there is often no such link found in the documentation, but in practice the SAP system is the leading ERP system in many companies, and therefore there is a practical connection. An IT system such as SAP ERP can influence a company and its processes. In many companies, SAP is the dominant system, and a BPM maturity model needs to accommodate this reality.

The interviewees confirmed that, from their practical experience, a deployed SAP system is usually well in-bedded in the culture and operations of a company. On the one hand, employees often think about their personal SAP process experience and how SAP handles processes in general. One reason for this is that companies have used an SAP ERP system for many years. On the other hand, financial information is also stored within an SAP system, which provides a sound basis for the analysis of key metrics within business processes.

Some BPM maturity models are very complex and not easy to use. To help address this issue, this research has developed a set of guidelines, which can be used by practising managers and other relevant stakeholders. They provide practical guidance for companies using SAP and BPM, and can lead to an improvement in business performance to the benefit of many stakeholders. The online survey demonstrated considerable support for the validity of these, but every organisation is different and guidelines should always be evaluated and applied, according to the particular requirements of the organisation concerned. In a wider context, this research underscores the value of a process management approach to analysing and assessing technology issues, and suggests that this may usefully be adapted and applied to look at other key change issues in the business-IT environment.

REFERENCES

[1] M. Grube and M. G. Wynn, "The Impact of SAP on the Utilisation of Business Process Management (BPM) Maturity Models in ERP projects," The Tenth International Conference on Information, Process, and Knowledge Management (eKNOW2018) IARIA, Rome, Italy 2018.

[2] SAP SE: SAP Corporate Fact Sheet [retrieved: May. 2019]. [Online]. Available from: https://www.sap.com/documents/2017/04/4666ecdd-b67c-0010-82c7-eda71af511fa.html

[3] SAP, "ERP System | Enterprise Resource Planning | SAP," [retrieved: May. 2019]. [Online]. Available from: https://www.sap.com/products/enterprise-management-erp.html

[4] M. Rezaeian and M. Wynn, "The implementation of ERP systems in Iranian manufacturing SMEs", International Journal On Advances in Intelligent Systems, vol. 9, nos. 3&4, pp.600 – 614, 2016. Available from: http://eprints.glos.ac.uk/4264/

[5] H. Tscherwitschke, "What you need to know about business process management," Computerwoche. from FAQ zu BPM: Was Sie über Business-Process-Management wissen müssen. [retrieved: May. 2019]. [Online]. Available from: http://www.computerwoche.de/software/soa-bpm/1906638/, 2011.

[6] B. Heilig and M. Möller, "Business Process Management with SAP NetWeaver BPM," Bonn: SAP PRESS, 2014.

[7] P. Busch and P. Fettke, "Business Process Management under the Microscope: The Potential of Social Network Analysis," 44th Hawaii International Conference on System Sciences (HICSS), 2011.

[8] T. M. Bekele and Z. Weihua, "Towards collaborative business process management development current and future approaches," IEEE 3rd International Conference on Communication Software and Networks (ICCSN), 2011.

[9] R. M. Saco, "Maturity Models," Industrial Management, vol. 50, no. 4, pp. 11-15, 2008.

[10] E. Ericsson, P. Gustafsson, D. Höök, L. Marcks von Würtemberg, and W. Rocha Flores, "Process improvement framework evaluation," International Conference on Management Science and Engineering (ICMSE), 2010.

[11] W. S. Humphrey, "Characterizing the Software Process: A Maturity Framework," IEEE Software, vol. 5, no. 2, pp. 73-79, 1988.

[12] N. Markovic, Business Performance Measurement and Information Evolution Model. Proceedings of the European Conference on Information Management & Evaluation, pp. 257-264, 2010.

[13] A. Van Looy, "Does IT matter for business process maturity? A comparative study on business process maturity models," Proceedings of the 2010 international conference on the move to meaningful internet systems, Hersonissos, Crete, Greece, 2010.

[14] S. Aeppli, "When SAP BPM calculates," Computerwoche. from Selbsttest für CW-Leser: Wann sich SAP BPM rechnet. [retrieved: May. 2019]. [Online]. Available from: http://www.computerwoche.de/2503123, Jan 2012.

[15] A. Corallo, A. Margherita, M. Scalvenzi, and D. Storelli, "Building a process-based organization: The design roadmap at Superjet," International. Knowledge & Process Management, vol. 17, no. 2, pp. 49-61, 2010, doi: 10.1002/kpm.340.

[16] J. vom Brocke, J. et al., "Ten guidelines of good business process management," Business Process Management Journal, vol. 20, no. 4, pp. 530-548, 2014, doi: 10.1108/bpmj-06-2013-0074.

[17] M. Wynn, University-Industry Technology Transfer in the UK: Emerging Research and Opportunities. Chapter 6 Manufacturing Companies. IGI-Global, 2018. Available from: http://eprints.glos.ac.uk/5915/2/5915%20Wynn%20Chapter%206.pdf

[18] T. Neubauer, "An empirical study about the status of business process management," Business Process Management Journal, vol. 15, no. 2, pp. 166 – 183, 2009, doi: 10.1108/14637150910949434.

[19] R. Poston and S. Grabski, "Financial impacts of enterprise resource planning implementations," International Journal of Accounting Information Systems, vol. 2, no. 4, pp. 271-294, 2001, doi: 10.1016/S1467-0895(01)00024-0.

[20] Y. L. Antonucci, G. Corbitt, G. Stewart, and A. L. Harris, "Enterprise Systems Education: Where Are We? Where Are We Going?" Journal of Information Systems Education, vol. 15, no. 3, pp. 227-234, 2004.

[21] A. Van Looy, M. De Backer, G. Poels, and M. Snoeck, "Choosing the right business process maturity model," Information & Management, vol. 50, no. 7, pp. 466-488, 2013, doi: 10.1016/j.im.2013.06.002.

[22] A. Shaikh, A. Ahmed, N. Memon, and M. Memon, "Strengths and Weaknesses of Maturity Driven Process Improvement Effort," in International Conference on Complex, Intelligent and Software Intensive Systems, 2009. CISIS '09, pp. 481–486, 2009. doi: 10.1109/CISIS.2009.182.

[23] P. Cryer, "The research student's guide to success," 3rd ed., Maidenhead: Open University Press, 2006.

[24] M. Saunders, P. Lewis, and A. Thornhill, "Research methods for business students," New York: Prentice Hall, 2009.

[25] A. B. Ryan, "Post-Positivist Approaches to Research Researching and Writing your thesis: a guide for postgraduate students," pp. 12-26, MACE: Maynooth Adult and Community Education, 2006.

[26] E. G. Guba, "The Paradigm dialog," Newbury Park, Calif.: Sage Publications, 1990.

[27] Z. O'Leary, "The Social Science Jargon Buster," London, UK: SAGE Publications Ltd, 2007.

[28] G. Thomas, "How to Do Your Case Study: A Guide for Students and Researchers," SAGE Publications, 2011.

[29] J. Collis and R. Hussey, "Business research: a practical guide for undergraduate & postgraduate students," Basingstoke: Palgrave Macmillan, 2009.

[30] R. K. Yin, "Case Study Research: Design and Methods," SAGE Publications, 2009.

[31] R. Kumar, "Research Methodology: A Step-by-Step Guide for Beginners," SAGE Publications, 2011.

[32] A. Bryman and E. Bell, "Business research methods," Oxford University Press, 2007.

[33] S. E. Baker and R. Edwards, "How many qualitative interviews is enough? Expert voices and early career reflections on sampling and cases in qualitative research," Southampton: ESRC National Centre for Research Methods, University of Southampton, 2012, [retrieved: May. 2019]. [Online]. Available from: http://eprints.ncrm.ac.uk/2273/4/how_many_interviews.pdf

[34] R., L. Armstrong, "The Midpoint on a Five-Point Likert-Type Scale. Perceptual and Motor Skills," vol. 64 no. 2, pp. 359-362, 1987.

[35] C. Li, "Improving Business - IT Alignment through Business Architecture," (Doctorate Doctoral Dissertations), Lawrence Technological University, Southfield, MI, 2010, [retrieved: May, 2019], [Online]. Available from: https://search.proquest.com/docview/858204513

Scalable Distributed Simulation for Evolutionary Optimization of Swarms of Cyber-Physical Systems

Micha Sende,
Melanie Schranz

Lakeside Labs GmbH
Klagenfurt, Austria
lastname@lakeside-labs.com

Davide Conzon,
Enrico Ferrera,
Claudio Pastrone

Pervasive Technologies
LINKS Foundation
Turin, Italy
firstname.lastname
@linksfoundation.com

Arthur Pitman,
Midhat Jdeed,
Wilfried Elmenreich

Institute of Networked and Embedded Systems
University of Klagenfurt
Klagenfurt, Austria
firstname.lastname@aau.at

***Abstract*—Swarms of cyber-physical systems can be used to tackle many challenges that traditional multi-robot systems fail to address. In particular, the self-organizing nature of swarms ensures they are both scalable and adaptable. Such benefits come at the cost of having a highly complex system that is extremely hard to design manually. Therefore, an automated process is required for designing the local interactions between the cyber-physical systems that lead to the desired swarm behavior. In this work, the authors employ evolutionary design methodologies to generate the local controllers of the cyber-physical systems. This requires many simulation runs, which can be parallelized. Two approaches are proposed for distributing simulations among multiple servers. First, an approach where the distributed simulators are controlled centrally and second, a distributed approach where the controllers are exported to the simulators running standalone. The authors show that the distributed approach is suited for most scenarios and propose a network-based architecture. To evaluate the performance, the authors provide an implementation that builds upon the eXtensible Messaging and Presence Protocol (XMPP) and supersedes a previous implementation based on the Message Queue Telemetry Transport (MQTT) protocol. Measurements of the total optimization time show that it outperforms the previous implementation in certain cases by a factor greater than three. A scalability analysis shows that it is inversely proportional to the number of simulation servers and scales very well. Finally, a proof of concept demonstrates the ability to deploy the resulting controller onto cyber-physical systems. The results demonstrate the flexibility of the architecture and its performance. Therefore, it is well suited for distributing the simulation workload among multiple servers.**

Keywords–Cyber-Physical System (CPS); Swarm; Evolutionary optimization; Distributed simulation; eXtensible Messaging and Presence Protocol (XMPP).

I. INTRODUCTION

Over the last decade, the paradigm of self-organization has gained significant traction in many research communities. Inspired by nature, swarm robotics is also seeing increased interest. This concept can be generalized from swarm robotics to swarms of Cyber-Physical Systems (CPSs) [1]. Applying self-organization to coordinate swarms of CPSs is a rather new approach, which aims at handling the highly complex systems, currently available. On the one hand, coordinating multiple CPSs using swarm approaches offers many opportunities, such as adaptability, scalability, and robustness [2]. On the other hand, it necessitates the difficult process of designing the individual CPSs to achieve the desired swarm behavior.

Designing swarms of CPSs poses two main challenges. First, selecting the hardware that best suits the requirements of the swarm (see [3], [4], [5], [6] for a further examination of this problem) and second, designing the control algorithm defining the behavior of the individual CPSs. This paper focuses on the latter problem because many platforms for swarm robotic research already exist, e.g., Kilobot [7], Spiderino [8], or Colias [9] and designing the hardware itself is out of scope of this work. Other platforms developed for traditional robotic applications such as the TurtleBot [10] are also suitable for executing swarm algorithms. Approaches for designing local controllers of individual CPSs in a swarm can be categorized into two types. First, hierarchical top-down design starting from the desired global behavior of the swarm and second, bottom-up design defining the individual CPS behaviors and observing the resulting global behavior [11]. Design using either one of the mentioned approaches is still a difficult process as neither can predict the resulting swarm behavior based on the complex interactions between the CPSs [12]. This is especially true in dynamic environments. One method to tackle such design challenges is evolutionary optimization [13].

In this paper, the authors employ the bottom-up design process based on evolutionary algorithms. Generally, evolutionary algorithms aim to mimic the process of natural selection by recombining the most successful solutions to a defined problem [14]. In the context of swarms of CPSs, a solution refers to a control algorithm of the individual CPSs that is gradually improved during the optimization process. As experiments with real CPSs require an extensive amount of time, such methods typically employ accurate and fast simulations to evaluate the performance of candidate solutions in the evolutionary process [15]. Using state-of-the-art simulators allows to build upon standard models and perform optimizations with varying level of detail. The evaluation of control algorithms in evolutionary optimization can be executed in parallel, which is for example supported in the FRamework for EVOlutionary design (FREVO) [16] by using multiple cores on the same ma-

chine. A further step would be the distribution of evolutionary optimization with a client-server-protocol, as exemplified by Kriesel [17].

To tackle this problem, the authors propose an architecture to distribute the simulations of an evolutionary optimization process onto multiple servers. Based on the work presented in [1], this paper describes an improved, extensible architecture that considers the lessons learned. This architecture builds on different generic tools for performing the optimization, evaluating the controller candidates through simulation, and managing the communication network. An implementation is provided, partially based on existing tools. The implementation is evaluated by demonstrating its usability among different test scenarios. First, the optimization process in heterogeneous network setups is demonstrated. Second, the deployment of the obtained controllers onto Robot Operating System (ROS) [18] based platforms is demonstrated, including simulations and hardware platforms. Finally, the scalability of the optimization performance is analyzed in terms of total time taken. There is a significant performance improvement compared to the previously presented architecture yielding an effective scalability with high numbers of simulators.

The proposed architecture is implemented as part of the *CPSwarm workbench* [19] which is a tool chain developed in the H2020 research project CPSwarm. This workbench aids in developing self-organizing swarm behavior for CPSs. It starts from modeling and design, goes over simulation and optimization, to deployment and monitoring. It is built around a central launcher application that allows to graphically access and configure the different tools. This work describes the simulation and optimization section of the architecture, known as the *simulation and optimization environment*.

The paper is organized as follows. In Section II, the evolutionary approach for designing swarms is briefly reviewed. Section III describes the two proposed approaches for distributing the simulations in an evolutionary optimization process by comparing them based on the results of the previous paper. Section IV introduces the newly proposed architecture that eliminates the problems previously experienced. Next, an implementation of this architecture is described in Section V. Section VI describes the testbeds that are used to evaluate the features provided and to measure the performance of the presented solution. The results of the performance analysis are detailed in Section VII, including a comparison to the previous approach. Finally, Section VIII provides a discussion, presents future work, and concludes the paper.

II. Designing Swarms by Evolution

As described in the previous section, design by evolution can be used to tackle challenges such as scalability and generality [20], as well as adaptive self-organization [21]. These issues are not easy to handle, especially in changing environments and with dynamic interactions among the individual CPSs in a swarm.

Designing a swarm using evolution is an automatic design method that creates an intended swarm behavior in a bottom-up process starting from very small interacting components. This process modifies potential solutions until a satisfactory result is achieved. Such an approach is based on evolutionary computation techniques [22],[23] and mimics the Darwinian principle [2]. It describes the process of natural selection by recombining the most proper solutions to a defined problem. Evolution can be applied either on the level of individuals or the swarm as a whole. Typically, the process of evolving a behavior starts with the generation of a random population of individual behavior control algorithms. Each member of the selected population is evaluated using simulations and graded by a fitness function that allows ranking the behaviors' performances on the swarm level. The higher a behavior is ranked, the more likely it is to survive to the next generation. This selection process makes sure that only the best performing behaviors survive to serve as input for the next iteration of the evolutionary process. They are reproduced using genetic operators like crossover or mutation. This process is iterated for a specific number of generations or until a CPS controller emerges that exhibits the desired global swarm behavior. Nevertheless, design by evolution poses several challenges, including no guaranteed or predictable convergence to the global optimum, complex data structures, and the high costs of evolutionary computation itself.

Design by evolution dictates several tasks that a designer must face during the design of a system model. According to Fehervari and Elmenreich [24], there are six tasks to consider:

1) The *problem description* gives a highly abstracted vision of the problem. This includes constraints and the desired objectives for such a problem.
2) The *simulation setup* transfers the problem description into an abstracted problem model. This model specifies the system components, i.e., details about the CPSs and the environment.
3) The *interaction interface* defines the interactions among CPSs and their interactions with the environment. For instance, the CPSs' sensors and actuators as well as the communication protocols should be specified here.
4) The *evolvable decision unit* represents the CPS controller and is responsible for achieving the desired objectives, i.e., the global behavior of the swarm to achieve a common goal. Such a decision unit must be evolvable to allow genetic operations as crossover and mutation. It is most commonly represented by an Artificial Neural Network (ANN). There are different types of ANNs, e.g., fully-meshed ANNs, feed-forward ANNs, or HebbNets [25].
5) The *search algorithm* performs the optimization using evolutionary algorithms by applying the results from the above steps. During this task, an iterative mathematical model is used to find the optimal solution. The optimization result is dependent on the fitness function of the problem.
6) The *fitness function* represents the quality of the optimization result in a numerical way. There is no specific way or rule to design such a function as it is highly dependent on the problem description. The main purpose of this function is to guide the search algorithm to find the best solution. In general, the applicability and performance of a fitness function depends on the employed Optimization Tool (OT), thus there are no universally suitable fitness functions [26]. However, many studies in the field of evolutionary optimization have considered generic methods for fitness function design [27], [25].

Recently, several software frameworks have been implemented to support the procedure of evolutionary design. AutoMoDe [28] is a software for automatic design, which generates modular control software in the form of a probabilistic finite state machine. JBotEvolver is a Java-based versatile open-source platform for education and research-driven experiments in evolutionary robotics [29], which has been used in many studies [30], [31], [32]. Gehlsen and Page [33] addressed the topic of parallel execution of experiments for heuristic optimum seeking procedures. Their approach, DIstributed SiMulation Optimization (DISMO), supports distributed execution of Java simulations for optimization projects. As simulation core, the framework for DEveloping object-oriented Simulation MOdels in Java (DESMO-J) is used. The SImulation-based Multi-objective Evolutionary OptimizatioN (SIMEON) framework [34] implements several components in Java for providing an evolutionary optimization of problems modeled via simulation. Examples involve supply chain optimization and flexible manufacturing scheduling. However, the SIMEON framework does not provide a strategy for distributed simulation across multiple servers. In addition, the previously mentioned tool FREVO supports creating and evaluating swarm behavior by evolution and has been used in several studies as an evolution tool including robotics [12] and pattern generation [35].

Designing swarm behaviors using evolutionary optimization requires a large number of simulation runs. The next section summarizes and compares two approaches proposed in [1] for distributing these simulations onto different machines.

III. Distributed Simulation

Depending on the level of realism in simulation of swarms of CPSs, a considerable amount of computational power is required [36]. This is especially true for the high number of simulations needed to complete an evolutionary optimization process. This can be accelerated by parallelizing the simulations that are carried out within one generation. In previous work [1], the authors proposed two architectural approaches on how to perform this parallelization and to distribute the workload. Both approaches have a common architecture composed by two core components, the OT and one or more Simulation Tools (STs). This concept is visualized in Figure 1. The OT is responsible for performing the evolutionary optimization process explained in Section II. The STs perform the required simulations in each step of the optimization process and evaluate the fitness value of a controller candidate. The interconnection between the OT and the STs allows to pass the simulations required during the evolutionary optimization from the OT to the STs where they are executed in parallel. The STs simulate a homogeneous swarm of CPSs, where each CPS is controlled by a controller generated by the OT. This controller translates the sensor inputs to actuator outputs. The two components are interconnected to each other through an interface that defines how they communicate during the optimization process. The definition of a generic interface gives the opportunity to build on well established STs that support accurate simulation of swarms of CPSs with different levels of detail.

The key difference between the two approaches is the location where the CPS controller resides during the simulation. The centralized approach lets the CPS controller reside

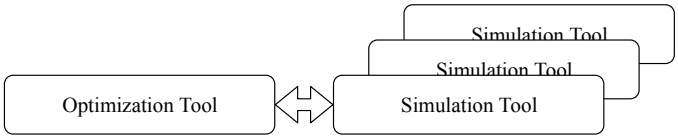


Figure 1. The concept of distributed simulation based on the components OT and ST.

centrally within the OT. The STs are merely executing the actions received from the controller in the OT and report back the sensor readings of the CPSs. With this approach the STs are hence centrally controlled by the OT. Instead, the distributed approach transfers the CPS controller from the OT to the STs, which can then independently run the simulation and only return back the resulting fitness value of the controller. This approach therefore distributes the control from the OT to the STs.

In [1] the two approaches have been implemented differently. The centralized approach has been implemented based on network socket inter-process communication. It allowed to distribute the simulations among multiple STs located on remote servers. The drawback of this implementation was that there was a high messaging overhead and a repeated polling for available STs. This inhibited the system to scale well with more than three STs. The distributed approach has been implemented based on file system inter-process communication. It allowed to fully pass over the control to the STs, but was limited to execution on a single machine. The authors now abstract away from the implementation specifics and thus refer to these approaches only as centralized and distributed.

Both cases require an interface between the OT and the STs, but there are some key differences. For the centralized approach, OT and STs must exchange the sensor readings and actuator controls. On the OT side, the interface must therefore receive the sensor inputs from the ST and feed them into the controller. The resulting actuator commands are determined by the controller and transmitted again to the corresponding ST. On the ST side, the interface must allow to control the CPS behavior using the actuator commands received from the OT. Once the commands are executed by the ST, the interface transmits back the sensor readings to the OT as they are perceived by the CPSs. The interface for the distributed approach requires to exchange the controller representation. On the OT side the interface must therefore export the controller representation and transmit it to the STs. Once the STs completed the simulations, it receives the result of the simulations to assess the performance of the controller. Depending on the implementation, this can be either raw log data or already processed information in form of a fitness value. The ST side of the interface receives the controller and integrates its representation into CPS behavior code. This allows the ST to translate the sensor readings to the actuator commands. Once the simulation is finished, it sends back the result of the simulation. Computing the fitness value of a simulation can be done on either side, in the OT or the ST.

Regardless of the approach, there are several messages that need to be exchanged. At first, there is a setup phase, which allows the OT to be aware of the available STs. This requires some kind of discovery process where the OT polls for STs, stating the requirements on the ST to be used (e.g., number

of dimensions or maximum number of agents supported). The STs satisfying these requirements then need to report back to the OT stating their availability. Then, the optimization can be performed, where the OT communicates with the selected STs. Depending on the chosen approach, a different number and different types of messages are exchanged between the OT and the STs. This communication takes place over several iterations until the OT has found a solution that it cannot further optimize. This optimal solution is represented by a CPS controller that can then be exported by the OT to be deployed on the CPSs.

The deployment of the optimized controller can take place in STs or on actual CPS hardware. The former can be used to further inspect the resulting controller. This allows to run further performance, scalability, or robustness analysis as well as visual inspection of the CPS behavior. The latter allows deploying the resulting controller to the CPSs and test it under real conditions. Whether the controller is further used in simulations or on actual CPS hardware, there is the requirement for an interface that allows connecting the CPSs' sensor inputs and actuator outputs to this controller. This interface must follow the same specification as the interface implemented in the STs used during the optimization process with the distributed approach. Therefore, this interface needs to be defined only once and can then be used for optimization and deployment.

When comparing the two approaches, both have their advantages and disadvantages. Looking at the centralized approach, the implementation of the ST is agnostic to the type of representation used for the controller. This has the advantage that new controller representations can be added to the OT easily, without the need to update the ST interface. The disadvantage is that there is a lot of message exchange between OT and the STs, throughout the simulation. When the number of STs is increased, the OT can become the bottleneck as it has to communicate constantly with each ST. Therefore, the distributed approach is less portable than the centralized approach but can reach a higher performance [1].

The performance of either architectural approach can be measured as the total time taken for the optimization process. As the authors shown in [1], this time can be expressed as

$$t_{\mathrm{opt}} = n_{\mathrm{gen}} \cdot \left(t_{\mathrm{evo}} + \frac{n_{\mathrm{pop}} \cdot n_{\mathrm{eval}}}{n_{\mathrm{sim}}} \cdot (t_{\mathrm{sim}} + t_{\mathrm{ohd}}) \right) \quad (1)$$

consisting of three time components. First, the time t_{evo}, which expresses the time required to perform the evolutionary calculations, such as selecting the best performing controllers and creating a new generation of controllers. Such tasks are executed for each generation of the optimization and hence are multiplied by the number of generations n_{gen}. Second, the time t_{sim}, which expresses the time taken by one simulation run. For simplification purposes, it is assumed that this time is measured in discrete steps and constant, regardless of the number of CPSs in the simulation. The simulation time can therefore be expressed as

$$t_{\mathrm{sim}} = n_{\mathrm{step}} \cdot t_{\mathrm{step}} \quad (2)$$

based on the number of discrete time steps n_{step} and the time t_{step} required to simulate one step. A simulation is performed for each controller in the population of n_{pop} controller candidates. For robustness and statistical significance, each controller can be evaluated n_{eval} times in a different variant of the same problem. This results in a number of $n_{\mathrm{pop}} \cdot n_{\mathrm{eval}}$ simulations that have to be performed during each of the n_{gen} generations. Depending on the number of available STs n_{sim}, the optimization process can be accelerated by distributing the simulations among these STs. The upper limit for the number of required STs is therefore $n_{\mathrm{pop}} \cdot n_{\mathrm{eval}}$. Finally, the third time component t_{ohd} states the amount of overhead time required during simulation. Where the other two time components are identical for both approaches, the overhead time varies between the centralized and the distributed approach. In [1] the authors differentiated between two implementations when calculating the overhead time. It is now generalized by calculating it for the centralized and distributed approach. This abstracts the implementation details and yields the overhead time as

$$t_{\mathrm{ohd}} = t_{\mathrm{setup}} + t_{\mathrm{run}} + t_{\mathrm{finalize}} \quad (3)$$

where the setup time t_{setup} is the time required to setup the STs, the run time t_{run} is the overhead time added while running the simulations, and the finalization time t_{finalize} the time to finalize the simulation and gather the results. For the centralized approach, the overhead time

$$\begin{aligned} t_{\mathrm{ohd,c}} = {} & n_{\mathrm{msg,setup}} \cdot t_{\mathrm{msg}} \\ & + n_{\mathrm{msg,run}} \cdot n_{\mathrm{step}} \cdot n_{\mathrm{cps}} \cdot t_{\mathrm{msg}} \\ & + n_{\mathrm{msg,finalize}} \cdot t_{\mathrm{msg}} + t_{\mathrm{fitness}} \end{aligned} \quad (4)$$

contains two time components. First, the message transmission time t_{msg} between the OT and the STs. During setup, there are $n_{\mathrm{msg,setup}}$ messages to be exchanged. During run time, each of the n_{cps} CPSs in the STs communicates $n_{\mathrm{msg,run}}$ messages for every simulation time step, where the simulation lasts for n_{step} steps. When finalizing a simulation, there are $n_{\mathrm{msg,finalize}}$ messages to be exchanged. Second, the time t_{fitness} to compute the fitness value of a controller adds to the finalization time. For the distributed approach, the total overhead time sums up to

$$\begin{aligned} t_{\mathrm{ohd,d}} = {} & t_{\mathrm{export}} + n_{\mathrm{msg,setup}} \cdot t_{\mathrm{msg}} + t_{\mathrm{import}} \\ & + n_{\mathrm{msg,finalize}} \cdot t_{\mathrm{msg}} + t_{\mathrm{fitness}} \end{aligned} \quad (5)$$

that contains two additional time components as compared to the centralized approach. First, the time t_{export} to export the controller representation from the OT into a format readable by the STs. Second, the time t_{import} to import the controller into a ST. To compare the performance of both approaches, it is possible to calculate the ratio

$$r = \frac{t_{\mathrm{opt,c}}}{t_{\mathrm{opt,d}}}, \quad (6)$$

which relates the total optimization run time of the centralized approach $t_{\mathrm{opt,c}}$ with the optimization time of the distributed approach $t_{\mathrm{opt,d}}$. This ratio expresses, which approach is more suitable for a specific setup of parameters. The authors determined the most relevant parameters for analysis to be the simulation length as number of simulated steps n_{step} and the number of CPSs n_{cps} that are being simulated. The number of parallel STs has a negligible influence as both approaches can use parallelization. To compare the performance scalability of both approaches, the authors set the other parameters to a fixed value that has been derived using measurements on the testbed described in [1]. They are summarized in Table I where the evolutionary parameters were chosen to yield good results.

The resulting ratio r using these values can be seen in Figure 2 for a varying number of CPSs. A value of $r > 1$ means that the distributed approach performs better whereas a value of $r < 1$ means that the centralized approach is favorable. As both approaches can use parallelization, the resultant ratio is independent of the number of parallel STs used. It can be seen that for most cases the distributed approach performs better, even though the time for importing the controller is dominating in Table I. This is because there is a lot of messaging overhead if all CPS controllers are executed in the OT. This creates a bottleneck where most work still is performed by a single tool. As seen in Figure 2, this is not so crucial for small swarm sizes, but already for a swarm size of eight CPSs, the optimization with central control takes longer when simulations last more than 18 steps.

To conclude the comparison between the two approaches, it can be stated that the distributed approach is favorable most of the time as it outperforms the centralized approach in terms of total time taken to run the optimization. If the communication between the OT and the STs is implemented using a network socket based interface, the simulation workload can be well distributed onto different machines. In this case, the OT needs to be aware of the available STs. In the network-based implementation, previously presented in [1], the OT was polling for new STs, before each simulation run. This created a considerable amount of overhead, rendering it impractical to use with more than three STs. Therefore, this paper now proposes to have two separate phases. First, the setup phase, where all available STs are discovered and second, the actual optimization phase, which uses the available STs. The new architecture realizing this proposal based on the previous experience is presented in the next section.

IV. ARCHITECTURE

This section presents a distributed architecture that uses a network socket based approach to allow distribution of the STs among different machines. The communication is managed by a central broker, which keeps track of the available STs. It builds on the experience of the previously proposed architecture described in [1].

The first problem addressed is the discovery protocol responsible for determining the available STs. The previous architecture required repeated discovery before each simulation, hence the performances did not scale well with the number of STs. The new architecture therefore introduces a setup phase, where the STs announce themselves and their capabilities. Any updates to the available STs are further communicated,

TABLE I. Optimization parameters measured through simulations.

parameter	value
n_{gen}	200
n_{pop}	50
n_{eval}	1
$n_{\text{msg,setup}}$	4
$n_{\text{msg,run}}$	2
$n_{\text{msg,finalize}}$	1
t_{evo}	12 ms
t_{msg}	30 ms
t_{export}	5.35 ms
t_{import}	8833 ms
t_{fitness}	0.69 ms
t_{step}	100 ms

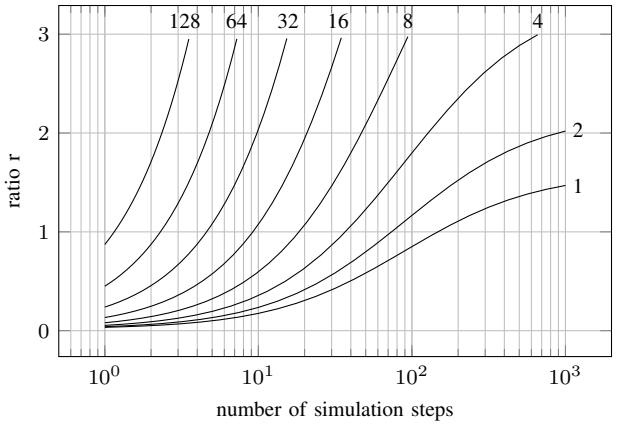


Figure 2. Ratio of optimization times between central and distributed control for different numbers of CPSs.

also during the optimization phase. These real time updates eliminate the need for repeated discovery by the OT. The second problem previously encountered is the high amount of messages that are exchanged during the optimization phase. The new architecture is therefore based on the distributed approach described in the previous section. The OT exports the controller to the STs to avoid exchanging the sensor readings and actuator commands. The STs can thus perform the simulation stand-alone without relying on the OT. This reduces the amount of exchanged messages considerably.

The architecture consists of four main components, as shown in Figure 3. First, it includes the previously mentioned OT, which is responsible for performing the evolutionary optimization. Second, there are several STs, distributed in different machines, called Simulation Servers (SSs), each one wrapped by a Simulation Manager (SM). This latter component is a software layer installed on the SS that implements the network interface and acts as a client that connects the ST to the broker. This allows the OT to communicate with the STs, without knowing the exact type of ST actually used. Third, there is a component called Simulation and Optimization Orchestrator (SOO), newly introduced in this architecture. The SOO is in charge of keeping track of all the SMs and coordinating the simulation tasks. It maintains a list of available SMs together with the capabilities of the STs that they wrap. When it is launched, the user can indicate the requirements on the STs, such as dimensionality or minimum CPS cardinality. In this

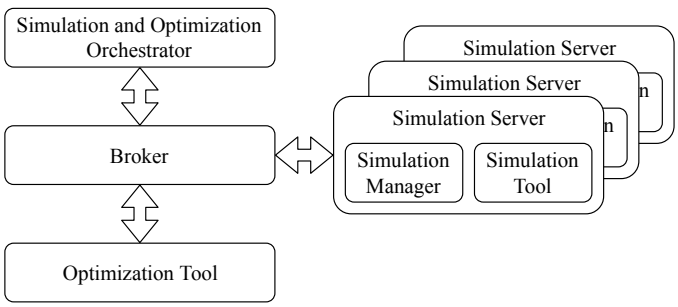


Figure 3. The network-based architecture consisting of the components SOO, broker, OT, and SS.

way, the SOO can select the SMs that fulfill the requirements. Finally, there is a broker that handles all communication between the other components.

The proposed architecture can perform two different kinds of workflows requiring simulations. First, the SOO can perform an optimization using the OT, where each controller candidate is simulated in a SS. Figure 4 illustrates the flow of messages between the SOO, the OT, and two exemplary SMs during the optimization process. In the initialization phase, all components announce their availability by broadcasting presence information. The SOO collects this information to create a list of available SMs and their capabilities to be used in the optimization phase. Similarly, the OT's presence informs the SOO that it is ready to perform optimization tasks. When the user starts the optimization, the SOO evaluates the available SMs, selects the ones that fulfill the indicated requirements, and transmits to them the configuration files needed to setup the ST. Once all the SMs have confirmed to be configured, the SOO sends a *StartOptimization* message to the OT, containing the list of SMs to be used for the optimization process. The OT replies with an *OptimizationStarted* message that includes a unique IDentifier (ID) of the optimization process. Then, the OT starts the optimization, using the STs that satisfy the requirements. It uses all the configured SMs in parallel by sending a sequence of *RunSimulation* messages to them. Each of these messages contains a candidate controller to be evaluated. The OT awaits the corresponding *SimulationResult* messages from the SMs. Throughout the optimization process, the SOO may request the progress of the optimization process intermittently or even cancel it by sending the OT a *GetProgress* or *CancelOptimization* message, respectively. Once the optimization process completed, the OT sends a final *OptimizationProgress* message to the SOO, containing the optimized controller.

Second, the SOO can send a specific controller candidate to a SM for more in depth analysis. This allows to evaluate a controller found by the OT more thoroughly, e.g., through visual replay using the ST Graphical User Interface (GUI). In case of visual replay, the selected ST must run on one machine directly accessible to the user, who needs to see the GUI of the ST. In this case, the SOO is responsible for sending the controller to the selected SM. This is visualized in Figure 5. In this much simpler scenario, the OT is not involved and SOO and SM communicate directly.

As this architecture introduces two separate phases for setup and optimization, the theoretical time required for optimization therefore changes from (1) to

$$t_{\text{opt}} = t_{\text{setup}} + n_{\text{gen}} \cdot \left(t_{\text{evo}} + \frac{n_{\text{pop}} \cdot n_{\text{eval}}}{n_{\text{sim}}} \cdot (t_{\text{sim}} + t_{\text{ohd}}) \right) \tag{7}$$

having the additional setup time

$$t_{\text{setup}} = (n_{\text{msg,presence}} + n_{\text{msg,config}} + 2) \cdot t_{\text{msg}} \tag{8}$$

being a multiple of the message transmission time t_{msg}. The total setup time is made up of the number of *Presence* messages $n_{\text{msg,presence}}$ transmitted from the SMs and the OT to the SOO, the number of *Configuration* messages $n_{\text{msg,config}}$ transmitted from the SOO to the SMs, and two messages to start the optimization and get the final result (*StartOptimization* and *OptimizationProgress*). As the OT requires only one *Presence* message and each of the n_{sim} SMs requires exactly one *Presence* message and one *Configuration* message, $n_{\text{msg,presence}} = n_{\text{sim}} + 1$ and $n_{\text{msg,config}} = n_{\text{sim}}$. Hence, the setup time can be rewritten as

$$t_{\text{setup}} = (2 \cdot n_{\text{sim}} + 3) \cdot t_{\text{msg}}. \tag{9}$$

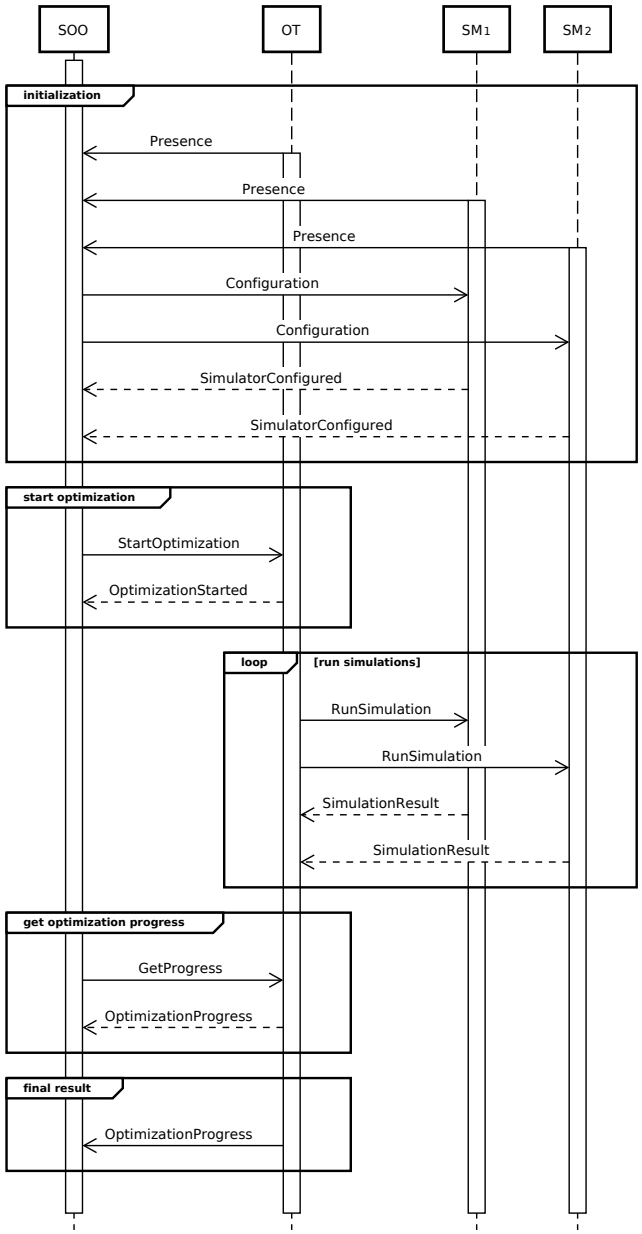


Figure 4. The messaging sequence during the optimization process.

Based on this architecture, the next section presents an implementation using the eXtensible Messaging and Presence Protocol (XMPP).

V. IMPLEMENTATION

The architecture presented in the previous section is implemented using existing tools wherever possible and developing new tools where necessary. The tools that where developed completely from scratch are the SOO as well as the SMs. The

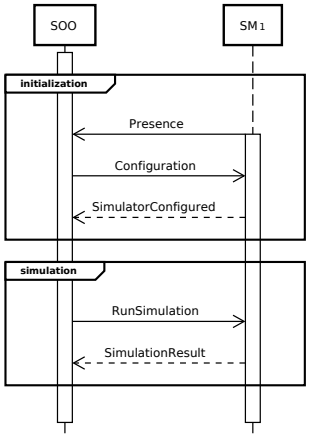


Figure 5. The messaging sequence for simulating a specific CPS controller.

existing tools are extended by a wrapper to enable integration with the proposed architecture. These are the OT and the STs. This section firstly describes the communication infrastructure that connects the different tools, followed by a description of their implementation.

The communication among the different tools happens according to the messaging sequence shown in Figures 4 and 5. To enable all tools to communicate with each other, the architecture relies on a central broker that manages the connections between them. The two messaging protocols Message Queue Telemetry Transport (MQTT) [37] and XMPP [38] have been tested extensively and are used for the implementations. MQTT is recognized as the de facto standard for event-driven architectures in the Internet of Things (IoT) domain. It has been chosen for the implementation previously presented in [1] because of its extreme simplicity. Its design principles attempt to minimize network bandwidth and device resource requirements whilst also ensuring reliability and some degree of assurance of delivery. Because XMPP owns features that allow to fulfill the requirements of the architecture described in Section IV it is selected as alternative communication protocol. It is favored over MQTT in the current implementation because MQTT does not offer all required features, such as one-to-one communication or a standardized presence protocol. Using XMPP, the proposed architecture can be implemented without having to rely on multiple different protocols.

In this work, the OT is implemented based on FREVO, a modular optimization system that applies the principles of genetic algorithms addressed in Section II [16]. FREVO divides the optimization task into several steps. First, it allows to model the problem as a simulation alongside an evaluating fitness function. Second, it allows to select different evolvable controller representations for the CPSs' controllers. Third, it allows to choose the evolution method used during the optimization process. Therefore, FREVO offers exceptional flexibility and allows many different setups to be explored. Currently, FREVO provides an implementation of the Neural Network Genetic Algorithm (NNGA) method [39]. It begins by creating n_{pop} controller candidates. In each of the n_{gen} generations, the controllers are evaluated and ranked according to their performance. Successful controllers, i.e., those with high fitness values, are carried to the next generation as elite, or are crossed or mutated to produce new controllers. In addition, a small proportion of entirely new random controllers is introduced with the intention of maintaining diversity in the population. For integrating FREVO in this architecture, it is extended with a layer supporting XMPP communication and hence called FREVO-XMPP.

The SOO is implemented as Java application embedding an XMPP client. The SOO can be configured with several parameters. They specify the requirements of the optimization task executed by the user. These parameters include requirements for the evolutionary process as well as requirements on the simulation.

The SM implementation is also done in Java and split into two parts. First, a common part is implemented as abstract class to provide a base module for all SMs implementations. Each SM specific for one ST is derived from this class and shared as a separate component. The specific part of the SMs defines how to handle the files received for the configuration and the messages with the controllers to be simulated.

When a SM starts, it adds the SOO to its roster, which is the list of "friend accounts". It signals its availability by sending a *Presence* message including a list of features provided by the wrapped ST, which is automatically received by all the "friends", i.e., the SOO. The current implementation features the number of dimensions and the maximum number of CPSs supported, but the list will be updated in future releases. In this way, the SOO receives and collects the availability of the SMs and it is able to choose the ones that fulfill the requirements for the execution of new tasks. After choosing the SMs to be used, the SOO sends to the SMs the files that are required for configuring the STs, using the XMPP file transfer. These include the models of CPSs and environment. The SMs confirm the reception of the configuration with an *SimulationConfigured* message. In case of simulation only, the SOO sends the controller to be replayed directly to the SM. In case of optimization, it sends a *StartOptimization* message to the OT, indicating the Jabber Identifiers (JIDs) of the SMs to be used.

Upon receiving a *StartOptimization* message, FREVO-XMPP creates an *OptimizationTask* to oversee the optimization process. As the evaluation of controllers is conducted by the SMs, FREVO-XMPP is largely input-output bound and can thus execute multiple *OptimizationTasks* in parallel without any significant Central Processing Unit (CPU) load. The *OptimizationTask* deserializes a FREVO-XMPP configuration, which specifies the type of evolution method, the controller representation, as well the operations to be performed on them throughout evolution. Furthermore, it receives a list of SMs, which may be used to evaluate controller candidates. Rather than evaluating a controller locally as is typically done in FREVO implementations, it sends a *RunSimulation* message to one of the associated SMs and blocks waiting for a *SimulationResult* message or a simulation timeout to occur.

As common basis for the STs, ROS [18] is chosen because it is an open-source solution, widely supported by several robotic platforms and many existing STs. It provides modularity and interoperability. In this way, the same CPS controllers can be tested on different STs and then deployed on actual ROS-based hardware platforms. Two specific SMs are implemented integrating the ROS simulations based on the

STs Stage [36] and Gazebo [40]. Several other STs have been considered as well, e.g., V-REP [41], ARGoS [42], jMAVSim [43], and STDR [44]. Resulting from an analysis of their controllability, configurability and support for standard models, the authors selected Stage and Gazebo for integration in the proposed architecture. Nevertheless, the modular approach of our architecture allows to integrate other ROS STs as well, with only little additional effort.

The communication between SM and ST is built on the work done in [1]. The ST is ROS-based and executed by the SM as ROS node using the standard ROS facilities such as launch files. The simulation to be run is installed beforehand on the SS as a ROS package. The CPS controller it is transmitted from FREVO-XMPP to the SM as C++ code, which allows them to be efficiently integrated into the simulation image. When the SM receives the controller, it forwards it to the ST, which is recompiled with the new controller. The ROS package contains a launch file that launches the required ST. The SM runs this launch file to launch the simulations, passing the parameters that the user indicates through the SOO. In case of optimization, the SMs are also in charge of calculating the fitness values of the tested controllers. This is achieved by parsing the ROS log files and computing the fitness value accordingly.

As a test case, a simple multi-CPS simulation called *EmergencyExit* is implemented in ROS. It realizes a problem where the CPSs have to escape from the environment while avoiding collisions with other CPSs running in discrete time and space. It is implemented as ROS package consisting of the controller representation and a wrapper class. The wrapper is adapted for different STs, while the controller can be reused seamlessly. During the optimization process, the wrapper stays fixed. The part that changes for each simulation is the controller code. Every CPS creates log files that are used by the SM to report to FREVO-XMPP the overall fitness value of the controller used in that simulation.

A third SM is implemented based solely on Java without the need for ROS to demonstrate the flexibility of the architecture. It is used by the centralized approach previously introduced in [1]. The ST is a very basic stand-alone Java simulator called *Minisim* [1]. It is a command-line, multi-CPS ST simulating a capture-the-flag game with multiple CPSs on a two-dimensional grid. *Minisim* has been specifically developed for testing the network communication between FREVO-MQTT and the SMs.

It is planned to release the code of this implementation in 2019 as open source on the CPSwarm Github repository (https://github.com/cpswarm). It will include the OT FREVO-XMPP, the SOO, and SM implementations for the STs Stage and Gazebo.

Several testbeds are setup to test the solution presented in this section and evaluate its performance. The description of the testbeds and the corresponding test cases are described in the next section.

VI. TESTBED

The solution that has been presented in the previous chapters is evaluated through three test cases. This section describes the testbed setup of these test cases. The first setup acts as a Proof of Concept (PoC) to demonstrate the provided features. The other two are used to evaluate the performance of the presented approaches.

For the first test case, three SSs running the Stage ST and the corresponding SM are used. Another computer is used both as SS, running the Gazebo ST with SM, and to run the SOO and the OT FREVO-XMPP. The XMPP server is installed in the cloud. Both Openfire and Tigase have been used. This setup is visualized in Figure 6.

To test the different components and workflows of the architecture, first an optimization is performed and the result is then replayed locally using a GUI. For this purpose, the authors implemented the *EmergencyExit* problem in simulation as ROS components, both for the Stage ST and the Gazebo ST. The implementations are based on a simple scenario with three CPSs and two exits. The scenario setup can be seen in Figures 7 and 8 for the Stage and Gazebo STs, respectively. This simple setup allows to effectively test the architecture without shifting the focus on the challenges related to performing complex and large-numbered multi-CPS simulations. The two implementations feature a different level of abstraction. First Stage, which implements the CPSs as simple squares and second Gazebo, which implements the CPSs as TurtleBot robots.

To perform the test, the authors launched the SOO selecting the optimization workflow with the requirement to perform simulations in two dimensions. This requirement was defined because a more abstract simulation yields better performance of the optimization, which includes a high number of simulations. As a result, the SOO successfully selected the three SSs running Stage and excluded the one running Gazebo. Then, the SOO launched FREVO-XMPP indicating to it the SSs to use and FREVO-XMPP distributed the simulation tasks onto them. Once the optimization finished, FREVO-XMPP returned the optimized controller to the SOO. To continue the test, the authors then launched the SOO again, this time selecting the simulation-only workflow with the requirement to perform the simulation in three dimensions with a GUI. As a result, the SOO successfully launched the simulations locally in Gazebo displaying the GUI with the 3D environment. The more detailed ST Gazebo allowed to replay and test the final optimization result under more realistic conditions. This test case showed the ability of the SOO to automatically choose the correct SS based on the requirements specified by the user and the capabilities exported by the SMs. It thus demonstrated

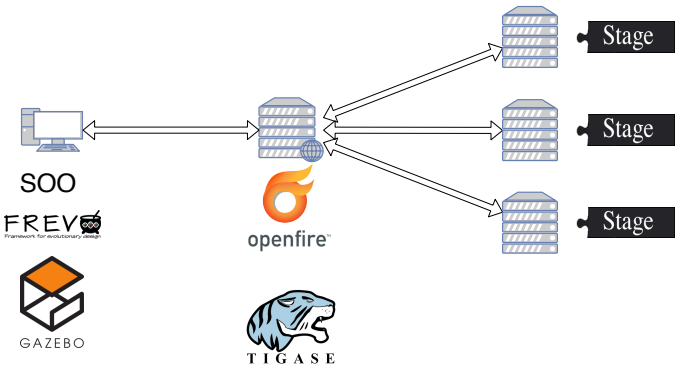


Figure 6. PoC testbed setup.

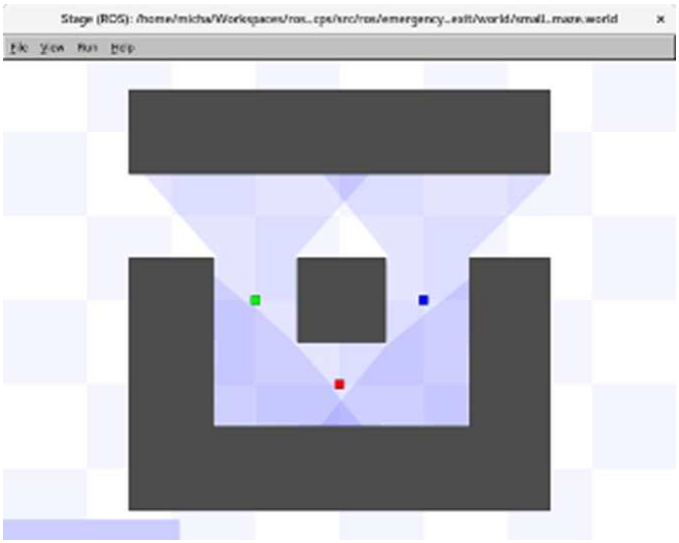


Figure 7. Stage ROS implementation of the *EmergencyExit* simulation.

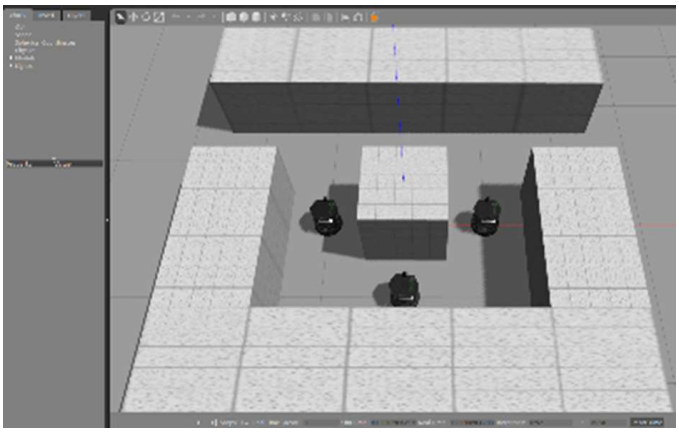


Figure 8. Gazebo ROS implementation of the *EmergencyExit* simulation.

how the STs are seamlessly integrated through the architecture described in this paper.

This test is complemented by using the optimized CPS controller for deployment on real hardware. By using ROS during the optimization and simulation process the authors have already demonstrated the portability of the controller among different STs. To take this even one step further, the same controller has been installed onto TurtleBot robots. The authors performed tests in an environment similar to the one used in simulation, see Figure 9. In this way, it has been demonstrated the complete chain from optimization, over simulation, up to deployment on a CPS hardware platform.

A second test case is constructed for a realistic distributed comparison between the centralized approach and the distributed approach. For this objective, the setup is the one shown in Figure 10 with four distributed SSs. For technical reasons, the centralized approach is implemented using the *Minisim* Java simulation, while the distributed approach is based on the *EmergencyExit* ROS simulation. Nevertheless, both approaches are comparable as both perform simulations lasting for the given number of steps. Specifically, for this test case, four

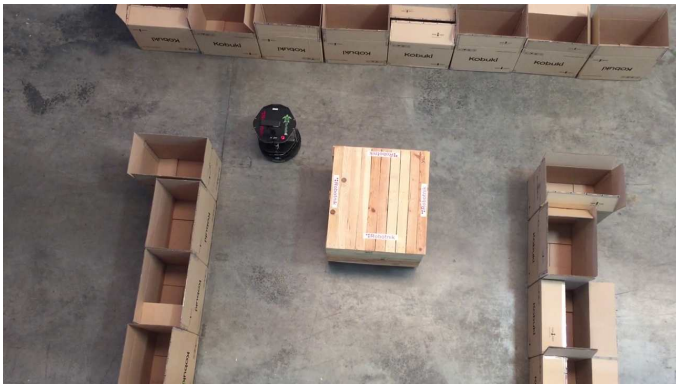


Figure 9. Real world experiment of the of the *EmergencyExit* problem using TurtleBot robots.

distributed computers are used: three SSs having installed the Stage ST with corresponding SM, the *EmergencyExit* ROS simulation and ROS Kinetic. The SOO and the OT FREVO-XMPP run on another computer that acts also as one SS, with the same setup as the others. All the components have been connected to a Tigase XMPP server running in the cloud. With this test case it is possible to test the complete optimization process, first using one SS and then parallelizing it on two, three, and four SSs.

Finally, a third test case is constructed to evaluate the scalability of the implementation with the number of SSs. Specifically, the objective is to show what degree of parallelization is possible with the current implementation, based on the distributed approach. To do this, all implemented tools (SOO, OT, and SMs) are executed on a single computer with 12 Intel Xeon X5675 processors running at $3.07\,\mathrm{GHz}$ and $16\,\mathrm{GB}$ of memory. Using hyper-threading, it supports 24 threads that can run in parallel. The operating system is Ubuntu 16.04 $64\,\mathrm{bit}$ running OpenJDK 9 Java. This setup is visualized in Figure 11. As before, the OT used is FREVO-XMPP and the Tigase XMPP server runs in the cloud. To rule out performance limitations of the test computer on FREVO-XMPP, the simulations used for the scalability analysis are just a sleep phase, which does not put any computational load on the computer. As it only serves to analyze the scalability of the network performance with the number of SSs it emulates the overhead time, which changes from (5) to

$$t_{\mathrm{ohd}} = (n_{\mathrm{msg,setup}} + n_{\mathrm{msg,finalize}}) \cdot t_{\mathrm{msg}} \qquad (10)$$

excluding import, export, and fitness computation times. The performance measurements are discussed in the next section.

VII. Performance Evaluation

This section presents the performance evaluation of the proposed architecture acquired using the testbed described in the previous section. The performance is measured in total time taken for a complete optimization run. This optimization time is measured for a varying number of simulation steps n_{step} and a varying number of SSs n_{sim}. All other parameters are fixed. To get reliable results, each measurement is repeated at least five times until the relative error of the sample is at most $10\,\%$ with a confidence of $99.9\,\%$. As a first step, the centralized approach is compared to the distributed approach introduced

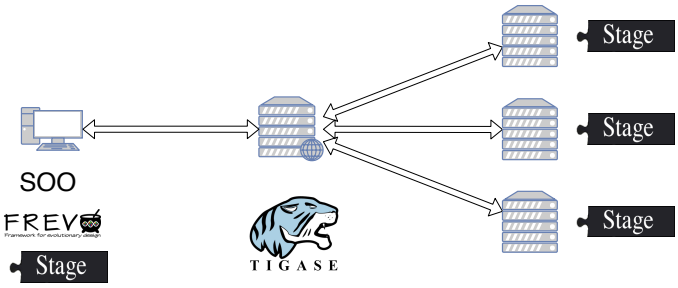


Figure 10. Performance comparison testbed setup.

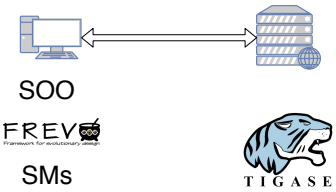


Figure 11. Scalability evaluation testbed setup.

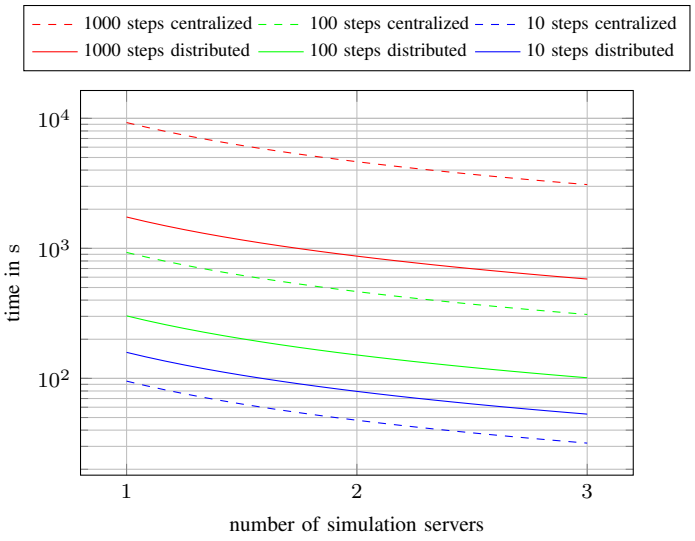


Figure 12. Theoretical comparison of the scalability with number of SSs of the optimization time between centralized and distributed approach for varying simulation lengths and eight CPSs.

in Section III. Then follows a more in depth analysis of the distributed approach that analyzes its scalability with the number of SSs.

A. Comparison of Centralized and Distributed Approach

To compare the centralized and the distributed approach, the authors first compute the total optimization time including setup time, based on (1) and (7), respectively. Then the authors perform measurements according to the testbed setup described as the second test case in the previous section. For calculating the optimization time, the measurements presented in Table I are used. The number of CPSs being simulated is $n_{\text{cps}} = 8$ and the evolutionary parameters are set to $n_{\text{gen}} = 4$ generations and $n_{\text{pop}} = 4$ controller candidates per generation. This yields the optimization times

$$t_{\text{opt,c}} = \frac{9.28\,\text{s} \cdot n_{\text{step}} + 2.41\,\text{s}}{n_{\text{sim}}} + 0.048\,\text{s} \qquad (11)$$

for the centralized approach and

$$t_{\text{opt,d}} = 0.06\,\text{s} \cdot n_{\text{sim}} + \frac{1.6\,\text{s} \cdot n_{\text{step}} + 142.39\,\text{s}}{n_{\text{sim}}} + 0.14\,\text{s} \qquad (12)$$

for the distributed approach with n_{sim} SSs. These optimizations times are plotted in Figure 12 as function of SSs and simulation steps. It shows the inverse proportionality between the simulation time and the number of SSs. Increasing the number of SSs is therefore well suited for reducing the total optimization time. The small term of direct proportionality of the distributed approach does not prevail for such low numbers of SSs. The major difference between the approaches lies within the dependency on the number of simulation steps. Here it becomes clear that the longer the simulation, the more favorable becomes the distributed approach. In this example with eight CPSs, the centralized approach is favorable only for short simulations in the order of ten steps. This is in line with the conclusions from the ratio shown in Figure 2.

Next, measurements using the testbed described in the previous section are performed. They are compared to the performance results of the centralized MQTT implementation presented in [1]. The results can be seen in Figure 13. They show the limitations of the implementation of the centralized approach. Because it performs SS discovery before each simulation it scales only up to three SSs. The implementation of the distributed approach mitigates this problem by introducing a different presence mechanism. It can be seen that the performance is mostly in line with the calculations presented above. For short simulations the centralized approach is preferable whereas for the other cases the distributed approach performs better.

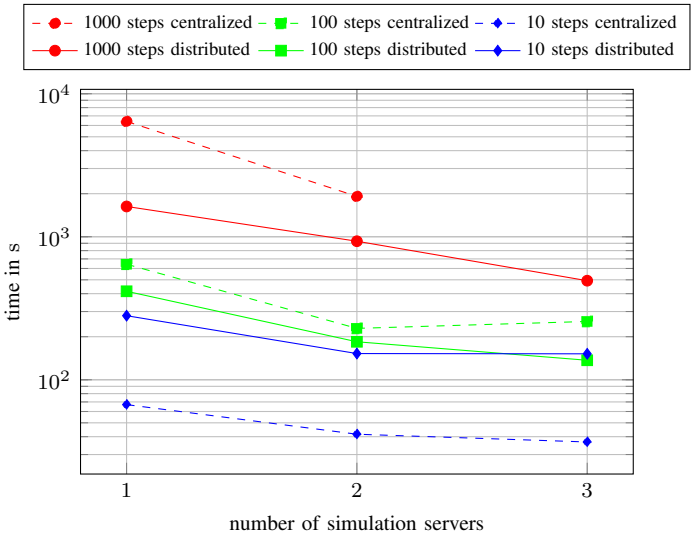


Figure 13. Measured comparison of the scalability with number of SSs of the optimization time between centralized and distributed approach for varying simulation lengths and eight CPSs.

B. Scalability Analysis

So far the authors showed the difference between the two approaches where the distributed approach excelled in most scenarios. To further investigate how well the performance scales with a larger range of SSs, measurements using the third test case described in Section VI are performed. To be able to assess the parallelization, the evolutionary parameters are set to $n_{\text{gen}} = 4$ and $n_{\text{pop}} = 32$. Because a typical optimization process includes only a single setup phase, only the optimization time is measured, which is the time between transmitting the *StartOptimization* message and receiving the final *OptimizationProgress* messages at the SOO. Figure 14 shows the resulting optimization time. The measurements are mostly in line with the theoretically calculated performance. The performance scales well with the number of SSs. There is only a small offset between measurements and theory, which is due to implementation details not captured in the model. When increasing the number of SSs to more than 16, it can be seen that the performance does not scale well anymore. This is due to the fact that the testbed reaches its limitations as the computer used for running the tests has only 24 cores.

VIII. Conclusion and Future Work

This paper presents a solution for the evolutionary design of swarms of CPSs based on remote simulation tools. The principal idea is to parallelize the simulations at each iteration of the evolutionary optimization process. The architecture designed for this solution builds upon the lessons learned from previous work [1], which introduced two different approaches with corresponding implementations. Starting from the evaluation of these approaches, the authors describe an XMPP based implementation of the architecture that combines the strengths of the two approaches previously presented in [1] and, at the same time, mitigates their weaknesses.

The new XMPP based implementation that uses the distributed approach is then compared to the previous implementation, which was based on the centralized approach. The

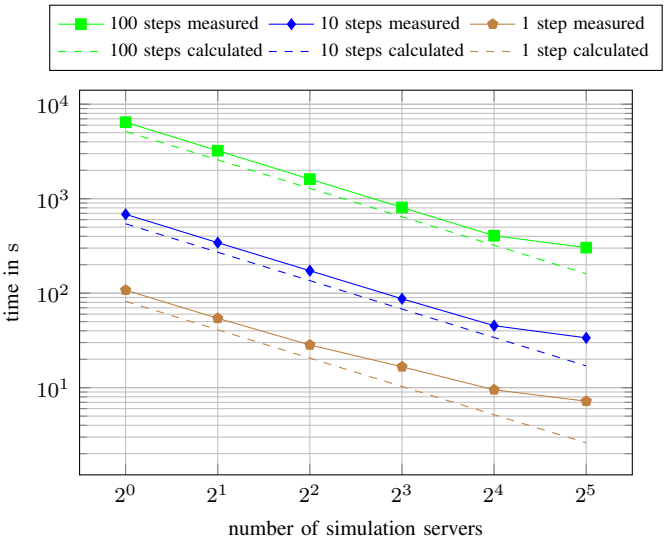


Figure 14. Scalability with number of SSs of the optimization time of the distributed approach for varying simulation lengths.

results in terms of total optimization time show that the new implementation is favorable in most cases. Only optimizations where the simulations have a short duration in the order of 1 s perform better using the previously implemented centralized approach. To take the performance analysis one step further, the new implementation is analyzed for its scalability with higher numbers of SSs. The results show that it scales well up to 32 SSs. The performance is expected to further scale well with even more SS, only the available hardware for creating such a large test setup was not available to the authors. Based on the new implementation, the authors demonstrate the ability of the architecture to successfully perform an optimization resulting in an optimized CPS controller. This controller is then replayed visually using the GUI of a locally installed ST connected to the OT. As a final step this controller is then exported and deployed on a TurtleBot robot to show that the resulting controller can bridge the gap from simulation to real world experiments.

Regardless of these achievements, the authors plan to extend the system in future to address several aspects. First, the compilation time within the STs could be reduced. Compiling the simulation code prior to testing a controller candidate is a major proportion of the overall time required by the optimization process. Incremental builds, i.e., recompiling only parts of the simulation that have changed, may reduce this significantly. In general, simulation code may be separated into four categories:

- Code specific to the simulation type. In the case of a ROS simulation, this includes the majority of ROS code. This code needs only be compiled once during the initial configuration of a SM.
- Code specific to the problem variant. This includes code such as randomly generated terrain and initial CPS positions. This code needs only be compiled during the initialization of a problem variant.
- Code specific to the controller representation type. This code implements a given type of representation, such as an ANN. This code needs only be compiled once for each representation type.
- Code specific to a given controller. This code is unique to a particular controller candidate, such as the weights and biases of an ANN, which can be updated without compilation.

A more optimized compilation strategy would separate code into these categories and would only recompile the modified pieces for a given simulation.

Second, the ST performance could be improved. In addition to compilation, the time required to start and stop ST instances adds significant overhead to simulation runs. Performance may also be improved by keeping the ST running and resetting its state for each individual run, e.g., returning CPSs to their start positions and resetting timer and counters. This would be particularly advantageous in situations where a controller may be changed purely by setting parameters, such as the weights and biases of an ANN. In any case, to implement this approach, improvements would need to be made to the simulation protocol to allow STs to distinguish the first simulation run from subsequent ones in a sequence, i.e., those from where compilation and simulation set up is required from those where it may be reused.

Third, the system robustness could be increased. One weakness of the current system is that it requires to restart from the beginning the (very long) optimization process, if it is disrupted. To address this, the authors propose saving the optimization state periodically by requesting the OT to send the last entire generation of controller candidates back to the SOO every m generations. Furthermore, the OT must be extended to allow it start optimization with such a set of controller candidates, rather than just random ones. Communication with the SMs could also be improved by continuously monitoring XMPP presence notifications. If a SM goes offline, the OT can immediately identify this and ask to the SOO if there is another SS available to use. Also additional policies must be defined for retrying controller candidates in the event of a timeout.

Fourth, the system scalability could be extended. Currently, the system requires to instantiate one dedicated machine for each ST. By using a container service such as Docker [45], multiple STs may be run by each SS. To do this, a set of docker images containing the ROS based STs and related SMs need to be created. Furthermore, this strategy combined with solutions like Docker Swarm [46] or Kubernetes [47] would open the possibility to allow the user to rapidly deploy and easily maintain large-scale sets of STs. This would allow deploying the system to the cloud and thus addressing the inherently resource-bound nature of simulation. Corresponding improvements could also be made to the OT to allow it to support a large number of SMs.

Fifth, the optimization algorithms could be improved. Currently, the system offers the NNGA as the optimization method. While this algorithm produces a predicable number of simulation runs, other algorithms may offer improved performance. Similarly, the authors also plan to implement more advanced controller representations.

Finally, the tools and protocols could be generalized. In the future, the authors have already planned to support a greater range of OTs and STs (e.g., V-REP and ARGoS) and thus improve the value of the system to the community as a whole.

ACKNOWLEDGMENT

The authors thank Robotnik Automation S.L.L. for porting the implementation to TurtleBot robots and the Gazebo ST. The research leading to these results has received funding from the European Union Horizon 2020 research and innovation program under grant agreement no. 731946.

REFERENCES

[1] M. Rappaport, D. Conzon, M. Jdeed, M. Schranz, E. Ferrera, and W. Elmenreich, "Distributed simulation for evolutionary design of swarms of cyber-physical systems," in Proc. Int. Conf. on Adaptive and Self-Adaptive Systems and Applications (ADAPTIVE). IARIA, Feb. 2018, pp. 60–65, ISBN: 978-1-61208-610-1.

[2] M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo, "Swarm robotics: a review from the swarm engineering perspective," Swarm Intelligence, vol. 7, no. 1, Mar. 2013, pp. 1–41, DOI: 10.1007/s11721-012-0075-2.

[3] I. Fehérvári, V. Trianni, and W. Elmenreich, "On the effects of the robot configuration on evolving coordinated motion behaviors," in Proc. Congress on Evolutionary Computation (CEC). IEEE, Jun. 2013, pp. 1209–1216, ISBN: 978-1-4799-0452-5.

[4] R. Goldsmith, "Real world hardware evolution: A mobile platform for sensor evolution," in Proc. Int. Conf. on Evolvable Systems: From Biology to Hardware (ICES). Springer, Mar. 2003, pp. 355–364, ISBN: 978-3-540-36553-2.

[5] J. Bongard, "Exploiting multiple robots to accelerate self-modeling," in Proc. Conf. on Genetic and Evolutionary Computation (GECCO). ACM, Jul. 2007, pp. 214–221, ISBN: 978-1-59593-697-4.

[6] D. Floreano and F. Mondada, "Hardware solutions for evolutionary robotics," in Proc. European Workshop on Evolutionary Robotics (EvoRobot). Springer, Apr. 1998, pp. 137–151, ISBN: 978-3-540-49902-2.

[7] M. Rubenstein, C. Ahler, and R. Nagpal, "Kilobot: A low cost scalable robot system for collective behaviors," in Int. Conf. on Robotics and Automation (ICRA). IEEE, May 2012, pp. 3293–3298, ISBN: 978-1-4673-1404-6.

[8] M. Jdeed, S. Zhevzhyk, F. Steinkellner, and W. Elmenreich, "Spiderino - a low-cost robot for swarm research and educational purposes," in Proc. Workshop on Intelligent Solutions in Embedded Systems (WISES). IEEE, Jun. 2017, pp. 35–39, ISBN: 978-1-5386-1157-9.

[9] F. Arvin, J. Murray, C. Zhang, and S. Yue, "Colias: An autonomous micro robot for swarm robotic applications," Int. J. Advanced Robotic Systems, vol. 11, no. 7, Jul. 2014, pp. 1–10, DOI: 10.5772/58730.

[10] Open Source Robotics Foundation, Inc. TurtleBot2. [Online]. Available: https://www.turtlebot.com/turtlebot2/ [retrieved: May, 2019]

[11] V. Crespi, A. Galstyan, and K. Lerman, "Top-down vs bottom-up methodologies in multi-agent system design," Autonomous Robots, vol. 24, no. 3, Apr. 2008, pp. 303–313, DOI: 10.1007/s10514-007-9080-5.

[12] I. Fehérvári and W. Elmenreich, "Evolving neural network controllers for a team of self-organizing robots," Journal of Robotics, vol. 2010, Mar. 2010, pp. 1–10, DOI: 10.1155/2010/841286.

[13] J. Xiao, Z. Michalewicz, L. Zhang, and K. Trojanowski, "Adaptive evolutionary planner/navigator for mobile robots," IEEE Transactions on Evolutionary Computation, vol. 1, no. 1, Apr. 1997, pp. 18–28, DOI: 10.1109/4235.585889.

[14] C. M. Fernandes and A. C. Rosa, "Evolutionary algorithms with dissortative mating on static and dynamic environments," in Advances in Evolutionary Algorithms. InTech, Nov. 2008, pp. 181–206, ISBN: 978-953-7619-11-4.

[15] L. Winkler and H. Wörn, "Symbricator3D – a distributed simulation environment for modular robots," in Proc. Int. Conf. on Intelligent Robotics and Applications (ICIRA). Springer, Dec. 2009, pp. 1266–1277, ISBN: 978-3-642-10817-4.

[16] A. Sobe, I. Fehérvári, and W. Elmenreich, "FREVO: A tool for evolving and evaluating self-organizing systems," in Proc. Int. Conf. on Self-Adaptive and Self-Organizing Systems Workshops (SASOW). IEEE, Sep. 2012, pp. 105–110, ISBN: 978-0-7695-4895-1.

[17] D. Kriesel, "Verteilte, evolutionäre Optimierung von Schwärmen [distributed evolution of swarms]," Master's thesis, Rheinische Friedrich-Wilhelm-Universität Bonn, Mar. 2009, URL: http://www.dkriesel.com/_media/science/diplomarbeit-de-1column-11pt.pdf.

[18] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "ROS: an open-source robot operating system," in Proc. ICRA Workshop on Open Source Software in Robotics, May 2009, URL: http://www.willowgarage.com/sites/default/files/icraoss09-ROS.pdf.

[19] A. Bagnato, R. K. Bíró, D. Bonino, C. Pastrone, W. Elmenreich, R. Reiners, M. Schranz, and E. Arnautovic, "Designing swarms of cyber-physical systems: The H2020 CPSwarm project: Invited paper," in Proc. Computing Frontiers Conf. ACM, May 2017, pp. 305–312, ISBN: 978-1-4503-4487-6.

[20] M. Dorigo, V. Trianni, E. Şahin, R. Groß, T. H. Labella, G. Baldassarre, S. Nolfi, J.-L. Deneubourg, F. Mondada, D. Floreano, and L. M. Gambardella, "Evolving self-organizing behaviors for a swarm-bot," Autonomous Robots, vol. 17, no. 2, Sep. 2004, pp. 223–245, DOI: 10.1023/B:AURO.0000033973.24945.f3.

[21] Y. Yao, K. Marchal, and Y. Van de Peer, "Improving the adaptability of simulated evolutionary swarm robots in dynamically changing environments," PLoS ONE, vol. 9, no. 3, Mar. 2014, pp. 1–9, DOI: 10.1371/journal.pone.0090695.

[22] J. H. Holland, Adaptation in Natural and Artificial Systems. MIT Press, Apr. 1992, ISBN: 9780262082136.

[23] I. Rechenberg, Evolutionsstrategie – Optimierung technischer Systeme nach Prinzipien der biologischen Evolution [Evolution strategy – Optimization of technical systems according to the principles of biological evolution]. Fromman-Holzboog, 1973, ISBN: 978-3-772-80373-4.

[24] I. Fehérvári and W. Elmenreich, "Evolution as a tool to design self-organizing systems," in Proc. Int. Workshop on Self-Organizing Systems (IWSOS). Springer, Jan. 2014, pp. 139–144, ISBN: 978-3-642-54140-7.

[25] I. Fehérvári, "On evolving self-organizing technical systems," Ph.D. dissertation, Alpen-Adria-Universität Klagenfurt, Nov. 2013.

[26] A. J. Lockett, "Insights from adversarial fitness functions," in Pro. Conf. on Foundations of Genetic Algorithms (FOGA). ACM, Jan. 2015, pp. 25–39, ISBN: 978-1-4503-3434-1.

[27] D. Floreano and J. Urzelai, "Evolutionary robots with on-line self-organization and behavioral fitness," Neural Networks, vol. 13, no. 4-5, Jun. 2000, pp. 431–443, DOI: 10.1016/S0893-6080(00)00032-0.

[28] G. Francesca, M. Brambilla, A. Brutschy, V. Trianni, and M. Birattari, "AutoMoDe: A novel approach to the automatic design of control software for robot swarms," Swarm Intelligence, vol. 8, no. 2, Jun. 2014, pp. 89–112, DOI: 10.1007/s11721-014-0092-4.

[29] M. Duarte, F. Silva, T. Rodrigues, S. M. Oliveira, and A. L. Christensen, "JBotEvolver: A versatile simulation platform for evolutionary robotics," in Proc. Int. Conf. on the Synthesis and Simulation of Living Systems (ALIFE). MIT Press, Jul. 2014, pp. 210–211, ISBN: 978-0-262-32621-6.

[30] M. Duarte, A. L. Christensen, and S. Oliveira, "Towards artificial evolution of complex behaviors observed in insect colonies," in Proc. Portuguese Conference on Artificial Intelligence (EPIA). Springer, Oct. 2011, pp. 153–167, ISBN: 978-3-642-24769-9.

[31] J. Gomes, P. Urbano, and A. L. Christensen, "Evolution of swarm robotics systems with novelty search," Swarm Intelligence, vol. 7, no. 2-3, Sep. 2013, pp. 115–144, DOI: 10.1007/s11721-013-0081-z.

[32] T. Rodrigues, M. Duarte, S. Oliveira, and A. L. Christensen, "What you choose to see is what you get: an experiment with learnt sensory modulation in a robotic foraging task," in European Conf. on the Applications of Evolutionary Computation (EvoApplications). Springer, Apr. 2014, pp. 789–801, ISBN: 978-3-662-45523-4.

[33] B. Gehlsen and B. Page, "A framework for distributed simulation optimization," in Proc. Winter Simulation Conf. (WSC). ACM, Dec. 2001, pp. 508–514, ISBN: 0-7803-7309-X.

[34] R. A. Halim and M. D. Seck, "The simulation-based multi-objective evolutionary optimization (SIMEON) framework," in Proc. Winter Simulation Conf. (WSC). IEEE, Dec. 2011, pp. 2834–2846, DOI: 10.1109/WSC.2011.6147987.

[35] W. Elmenreich and I. Fehérvári, "Evolving self-organizing cellular automata based on neural network genotypes," in Proc. Int. Workshop on Self-Organizing Systems (IWSOS). Springer, Feb. 2011, pp. 16–25, ISBN: 978-3-642-19167-1.

[36] R. Vaughan, "Massively multiple robot simulations in Stage," Swarm Intelligence, vol. 2, no. 2-4, Dec. 2008, pp. 189–208, DOI: 10.1007/s11721-008-0014-4.

[37] A. Banks and R. Gupta, "MQTT version 3.1.1 plus errata 01," Organization for the Advancement of Structured Information Standards (OASIS), Standard, Dec. 2015, URL: http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/mqtt-v3.1.1.html.

[38] P. Saint-Andre, "Extensible messaging and presence protocol (XMPP): Core," Internet Engineering Task Force (IETF), RFC 6120, Oct. 2004, DOI: 10.17487/RFC6120.

[39] A. J. F. Van Rooij, L. C. Jain, and R. P. Johnson, Neural Network Training Using Genetic Algorithms. World Scientific Publishing Co., Mar. 1997, ISBN: 978-9-810-22919-1.

[40] N. Koenig and A. Howard, "Design and use paradigms for Gazebo, an open-source multi-robot simulator," in Int. Conf. on Intelligent Robots and Systems (IROS). IEEE/RSJ, Sep. 2004, pp. 2149–2154, ISBN: 0-7803-8463-6.

[41] M. Freese, S. Singh, F. Ozaki, and N. Matsuhira, "Virtual robot experimentation platform V-REP: A versatile 3d robot simulator," in Proc. Int. Conf. on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR). Springer, Nov. 2010, pp. 51–62, ISBN: 978-3-642-17319-6.

[42] C. Pinciroli, V. Trianni, R. O'Grady, G. Pini, A. Brutschy, M. Brambilla, N. Mathews, E. Ferrante, G. Di Caro, F. Ducatelle, T. Stirling, A. Gutirrez, L. Maria Gambardella, and M. Dorigo, "ARGoS: A modular, multi-engine simulator for heterogeneous swarm robotics," in Proc. Int. Conf. on Intelligent Robots and Systems (IROS). IEEE, Sep. 2011, pp. 5027–5034, ISBN: 978-1-61284-456-5.

[43] A. Driss, L. Krichen, F. Mohamed, and L. Fourati, "Simulation tools, environments and frameworks for UAV systems performance analysis," in Proc. Int. Wireless Communications and Mobile Computing Conf. (IWCMC). IEEE, Jun. 2018, pp. 1495–1500, ISBN: 978-1-5386-2070-0.

[44] F. M. Noori, D. Portugal, R. P Rocha, and M. Couceiro, "On 3D simulators for multi-robot systems in ROS: MORSE or Gazebo?" in Proc. Int. Symposium on Safety, Security and Rescue Robotics (SSRR). IEEE, Oct. 2017, pp. 19–24, ISBN: 978-1-5386-3923-8.

[45] B. B. Rad, H. J. Bhatti, and M. Ahmadi, "An introduction to docker and analysis of its performance," Int. J. Computer Science and Network Security (IJCSNS), vol. 17, no. 3, Mar. 2017, pp. 228–235, ISSN: 1738-7906.

[46] F. Soppelsa and C. Kaewkasi, Native Docker Clustering with Swarm. Packt Publishing, Dec. 2016, ISBN: 978-1-786-46975-5.

[47] D. K. Rensin, Kubernetes - Scheduling the Future at Cloud Scale. O'Reilly, Sep. 2015, ISBN: 978-1-492-04871-8.

www.iariajournals.org

International Journal On Advances in Intelligent Systems
issn: 1942-2679

International Journal On Advances in Internet Technology
issn: 1942-2652

International Journal On Advances in Life Sciences
issn: 1942-2660

International Journal On Advances in Networks and Services
issn: 1942-2644

International Journal On Advances in Security
issn: 1942-2636

International Journal On Advances in Software
issn: 1942-2628

International Journal On Advances in Systems and Measurements
issn: 1942-261x

International Journal On Advances in Telecommunications
issn: 1942-2601