DAGSTUHL
REPORTS

Volume 11, Issue 10, November 2021

Adaptive Resource Management for HPC Systems (Dagstuhl Seminar 21441)
Michael Gerndt, Masaaki Kondo, Barton P. Miller, and Tapasya Patki 1

Ensuring the Reliability and Robustness of Database Management Systems (Dagstuhl Seminar 21442)
Alexander Böhm, Maria Christakis, Eric Lo, and Manuel Rigger 20

Managing Industrial Control Systems Security Risks for Cyber Insurance (Dagstuhl Seminar 21451)
Simon Dejung, Mingyan Liu, Arndt Lüder, and Edgar Weippl 36

Unambiguity in Automata Theory (Dagstuhl Seminar 21452)
Thomas Colcombet, Karin Quaas, and Michał Skrzypczak 57

Descriptive Set Theory and Computable Topology (Dagstuhl Seminar 21461)
Mathieu Hoyrup, Arno Pauly, Victor Selivanov, and Mariya I. Soskova 72

Foundations of Persistent Programming (Dagstuhl Seminar 21462)
Hans-J. Boehm, Ori Lahav, and Azalea Raad 94

Geometric Modeling: Interoperability and New Challenges (Dagstuhl Seminar 21471)
Falai Chen, Tor Dokken, and Géraldine Morin 111

Geometric Logic, Constructivisation, and Automated Theorem Proving (Dagstuhl Seminar 21472)
Thierry Coquand, Hajime Ishihara, Sara Negri, and Peter M. Schuster 151

Secure Compilation (Dagstuhl Seminar 21481)
David Chisnall, Deepak Garg, Catalin Hritcu, and Mathias Payer 173

ISSN 2192-5283

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at
http://www.dagstuhl.de/dagpub/2192-5283

Publication date
April, 2022

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at `http://dnb.d-nb.de`.

Digital Object Identifier: 10.4230/DagRep.11.10.i

Aims and Scope
The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.
In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e. g. summaries from panel discussions or open problem sessions.

Contact
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de
http://www.dagstuhl.de/dagrep

Report from Dagstuhl Seminar 21441

Adaptive Resource Management for HPC Systems

Edited by

Michael Gerndt[1], Masaaki Kondo[2], Barton P. Miller[3], and Tapasya Patki[4]

1 **TU München, DE,** gerndt@in.tum.de
2 **Keio University – Yokohama, JP,** kondo@acsl.ics.keio.ac.jp
3 **University of Wisconsin-Madison, US,** bart@cs.wisc.edu
4 **LLNL – Livermore, US,** patki1@llnl.gov

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 21441 "Adaptive Resource Management for HPC Systems". The seminar investigated the impact of adaptive resource management of malleable applications on the management of the HPC system, the programming of the applications, the tools for performance analysis and tuning, as well as the efficient usage of the HPC systems. The discussions led to a joint summary presenting the state-of-the-art, required techniques on the various levels of HPC systems, as well as the foreseen advantages of adaptive resource management.

Seminar November 1–5, 2021 – http://www.dagstuhl.de/21441

2012 ACM Subject Classification Computer systems organization → Distributed architectures; Software and its engineering → Parallel programming languages; Software and its engineering → Process management

Keywords and phrases High Performance Computing, dynamic resource management, malleable HPC applications, power management, computing continuum

Digital Object Identifier 10.4230/DagRep.11.10.1

1 Executive Summary

Michael Gerndt
Masaaki Kondo
Barton P. Miller
Tapasya Patki

Today's supercomputers have very static resource management. Jobs are submitted via batch scripts to the resource manager, then scheduled on the machine with a fixed set of nodes. Other resources, such as power, network bandwidth and storage are not actively managed and are provided only on a best-effort basis. This inflexible, node-focused and static resource management will have to change in the future due to many reasons, some of them listed below.

First, applications are becoming increasingly more dynamic. Techniques such as adaptive mesh refinement, e.g., as used in Tsunami simulations, lead to scalability changes over the application's execution. Furthermore, only some application phases might profit from specialized accelerators, and I/O phases might even run best with a limited number of compute resources.

DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Additionally, the execution environment of applications is also becoming dynamic. Modern processors change the clock frequency according to the instruction mix as well as power and thermal envelopes. Heavy use of the vector units can lead to a lower clock frequency to stay in the thermal power budget, for example.

As an independent concern, due to the sheer number of components, failure rates are expected to increase thus slowing down computation or even leading to an increased number of node failures.

Finally, the upcoming machines will be power constrained, which means that the power will have to be carefully distributed among all running applications. The resulting power capping will impact the application's performance due to adaptation of the clock frequency and due to manufacturing variability. These challenges in HPC will only be solvable by using a more adaptive resource management approach. For example, compute nodes need to be redistributed among running applications to adapt to changes in the application's resource requirements either due to a varying number of grid points or interspersed algorithmic phases that profit from certain accelerators; network and I/O bandwidth will have to be assigned to applications to avoid interference caused by contention of concurrent communication and I/O phases; power needs to be dynamically redistributed both within an application and across applications to enable increased efficiency. Dynamic redistribution of resources will also give more flexibility to the resource manager to schedule jobs on the available resources and thus reduce idle times and efficiency lowering contention scenarios, e.g., in the situation of big jobs waiting for execution.

This Dagstuhl Seminar investigated a holistic, layered approach for adaptive resource management. It started with the resource management layer being responsible for scheduling applications on the machine and dynamically allocating resources to the running applications. At the programming level, applications need to be programmed in a resource-aware style such that they can adapt to resource changes and can make most efficient usage of the resources. On top of the programming interfaces, programming tools have to be available that allow the application developers to analyze and tune the applications for the varying amount of available resources. At the application level, applications have to be redesigned to enable significant gains in efficiency and throughput, e.g., adaptive mesh refinement, approximate computing, and power-aware algorithms are a few aspects to mention here.

The discussions led to a joint summary presenting the state-of-the-art, required techniques on these layers of HPC systems, as well as the foreseen advantages of adaptive resource management.

2 Table of Contents

Executive Summary

Michael Gerndt, Masaaki Kondo, Barton P. Miller, and Tapasya Patki 1

Overview of Talks

Flux: Next-Generation Resource Management and Scheduling for Scientific Workflow Enablement
Dong Ahn 5

Ongoing Efforts on Co-scheduling and Holistic Power Management
Eishi Arima 5

A Scalable RISC-V Power Controller Platform for HPC Processors
Andrea Bartolini 5

The Malleability Problem Statement: Differences between Supercomputing and the Cloud
Isaías A. Comprés Ureña 6

Towards Machine Learning Generation of Parallel Applications Performance Models
Eduardo César and Anna Sikora 6

Invasive Computing in HPC
Michael Gerndt 7

Towards Dynamic Node Resource Management in Next-Generation HPC Environments
Balazs Gerofi 8

Challenges of Resource Management on the "Data Platform": mdx
Toshihiro Hanawa 8

InvasiC HPC Programming: iMPI and EPOP
Jophin John and Santiago Narvaez Rivas 8

From GEOPM to the OIEP Reference Model: Embedding Energy and Power Runtime Systems into the Big Picture of HPC
Matthias Maiterth 9

Converged Computing: Combining the Best of HPC and the Cloud
Daniel John Milroy 9

Overview of State-of-the-Art Parallel Performance Measurement and Analysis Tools for heterogeneous systems
Bernd Mohr 10

Dynamism in HPC Resource Control
Frank Mueller 10

Dynamic Tuning of HPC Applications – ESPRESO FEM Library
Lubomir Riha 11

Invasive Computing – A Systems Programming Perspective
Wolfgang Schröder-Preikschat 11

From COMM_WORLD to Sessions and Bubbles: How new MPI Features Need to Interact with Resource Managers
Martin Schulz . . . 11

Dynamic Tuning of HPC Applications – MERIC
Ondrej Vysocky . . . 12

The Price Performance of Performance Models
Felix Wolf . . . 12

Outlook: Techniques for Malleability-Enabled Machines

Resource Management . . . 12

Programming Models: MPI and Beyond . . . 13

Unified Monitoring . . . 14

Turning Applications into Malleable Applications . . . 15

Incentives to Help in Dynamic Resource Management . . . 16

What is the Potential of Malleable Applications? . . . 16

Converged Computing . . . 17

Participants . . . 19

Remote Participants . . . 19

3 Overview of Talks

3.1 Flux: Next-Generation Resource Management and Scheduling for Scientific Workflow Enablement

Dong Ahn (LLNL – Livermore, US)

Many emerging scientific workflows that target high-end HPC systems require complex interplay with the resource and job management software (RJMS). However, portable, efficient and easy-to-use scheduling and execution of these workflows is still an unsolved problem. In this talk, I will present Flux, a next-generation RJMS designed specifically to address the key scheduling challenges of modern workflows in a scalable, easy-to-use, and portable manner. At the heart of Flux lies its ability to be seamlessly nested within batch allocations created by itself as well as other system schedulers (e.g., SLURM, MOAB, LSF, etc), serving the target workflows as their "personalized RJMS instances". In particular, Flux's consistent and rich set of well-defined APIs portably and efficiently support those workflows that can often feature non-traditional execution patterns such as requirements for complex co-scheduling, massive ensembles of small jobs and coordination among jobs in an ensemble. As part of this talk, I will also discuss Flux's graph-based resource data model, Flux's response to needing to schedule increasingly diverse resources, and how this model is becoming the center of our industry co-design efforts: for example, multi-tiered storage scheduling co-design with HPE and Cloud resource co-design with IBM T.J. Watson and RedHat OpenShift.

3.2 Ongoing Efforts on Co-scheduling and Holistic Power Management

Eishi Arima (TU München, DE)

This presentation covers our ongoing efforts related to co-scheduling, i.e., colocating multiple jobs at the same time on a node, and also holistic power management for HPC systems. More specifically, the talk will include: (1) open software architecture for sophisticated resource management, in particular power stack, and integrating our software stack based on the architecture; (2) a variety of co-scheduling studies for sophisticating them; and (3) opportunities to co-ordinate with the malleability.

3.3 A Scalable RISC-V Power Controller Platform for HPC Processors

Andrea Bartolini (University of Bologna, IT)

Today's high-performance computing (HPC) workloads crave data bandwidth, capacity and floating-point performance. High-performance chips feature with many performance-capable cores, vector units, DDRs and HBMs memory controllers', high-bandwidth and low-latency coherent I/Os, as well as domain-specific accelerators with staggering (several hundreds of Watts) peak power requirements.The peak power exceeds the TDP, and the package cost constrains the maximum TDP and sustainable peak power. Motherboards' form-factor, layout, and cost constraint the power distribution design and demand effective and reliable on-chip thermal management. Power, temperature, and energy are critical aspects that must be controlled and optimized online with a low-latency feedback loop with the on-chip power management IPs and sensors, Operating System, Security Subsystem, off-chip Board Management Controller (BMC) and power converters. We propose ControlPULP, a fully-digital and highly capable RISC-V based parallel microcontroller IP optimized for power management of complex HPC processors. Its design supports a single-core manager core and peripherals paired with a cluster of 8 processors to accelerate the Power Control Firmware workload, Direct Memory Access (DMA) engine for accessing on-chip sensors, a uDMA engine for off-chip AVSBUS/PMBUS peripheral support and BMC-based communication through the Management Component Transport Protocol (MCTP). The controller implements basic System Control and Management Interface (SCMI) doorbell-based protocol hosting up to 144 external interrupt lines. On the SW side, it relies on an open-source Real-time Operating System (FreeRTOS) for agile scheduling of the underlying Control Policy.

3.4 The Malleability Problem Statement: Differences between Supercomputing and the Cloud

Isaías A. Comprés Ureña (TU München, DE)

Malleability can bring benefits to our distributed memory supercomputers that are not possible with current static allocations. Malleability has already been achieved in cloud computing systems. The reasons are related to the differences between the workloads that are typical across these systems. The workloads of supercomputing pose additional challenges that have caused delays in the deployment of malleability in this domain.

3.5 Towards Machine Learning Generation of Parallel Applications Performance Models

Eduardo César (Autonomus University of Barcelona, ES), Anna Sikora (Autonomus University of Barcelona, ES)

Joint work of Eduardo César, Jordi Alcaraz, Anna Sikora
Main reference Jordi Alcaraz, Steven Sleder, Ali Tehrani Jamsaz, Anna Sikora, Ali Jannesari, Joan Sorribes, Eduardo César: "Building representative and balanced datasets of OpenMP parallel regions", in Proc. of the 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, PDP 2021, Valladolid, Spain, March 10-12, 2021, pp. 67–74, IEEE, 2021.
URL http://dx.doi.org/10.1109/PDP52278.2021.00019

Malleable HPC computing will require, among other things, efficient methods for estimating the amount of resources that an application needs to be executed efficiently. However, due to the heterogeneity found in HPC systems, adequate analytical models for performance improvement can be very difficult to generate. An alternative to traditional analytical models could be the use of machine learning algorithms, which may help to automatically create performance models to predict the appropriate configuration for one or multiple application's parameters.

Incorporating machine learning for automatic performance analysis and tuning is a promising path, but it introduces the need for generating balanced and representative datasets of parallel applications' executions.

First, to be able to build performance models, measurements are needed to calculate or select the proper values for one or multiple parameters which can impact performance. The selection of the right measurements is important as information can be redundant or, in the worst case, insufficient to correctly describe the relationship between them and performance parameters.

Second, these measurements should be used for bulding datasets of representative parallel code regions patterns. Thus, a methodology is needed for determining whether a given region covers a unique part of the input space not yet covered by the patterns already included in the dataset.

Finally, when such a dataset is used for performance tuning, an imbalance problem appears as the targets are now performance parameters instead of representative code regions. This imbalance appears because some performance parameters' values generally provide better performance than others for most cases. Consequently, machine learning algorithms may under-perform due to underrepresented cases, making the use of techniques to counter this natural imbalance necessary.

3.6 Invasive Computing in HPC

Michael Gerndt (TU München, DE)

TUM started the research on malleable HPC application as part of the Transregional Research Center TRR89 "Invasive Computing". This is a collaboration between the FAU Erlangen, KIT Karlsruhe, and Technische Universität München. The major focus is to develop concepts for resource-aware programming of embedded application for highly parallel chip multiprocessors.

TUM is investigating the extension of this concept for HPC applications. The resource management of HPC systems is static. The system is partitioned for system services, interactive access, and batch jobs. Nodes are assigned to applications for the applications whole lifetime. The concept of dynamic resource management for HPC systems allows to distribute resources dynamically to system services and running applications, and thus allows for a more efficient sharing of the resources.

TUM developed extensions of OMP and MPI, known as iMPI, for writing malleable MPI applications. Furthermore, a scalable in-memory application-level checkpointing system iCheck is under development. The benefits demonstrated in the TRR are better system utilization, more efficient resource usage, and a dynamic power corridor management. The work of invasive computing is continued in the European HPC project Deep-Sea.

3.7 Towards Dynamic Node Resource Management in Next-Generation HPC Environments

Balazs Gerofi (RIKEN – Kobe, JP)

Workload diversity in high-performance computing (HPC) environments has experienced an explosion in recent years. The increasing prevalence of Big Data processing, in-situ analytics, artificial intelligence (AI) and machine learning (ML) workloads, as well as multi-component workflows is pushing the limits of supercomputing systems that have been primarily designed to serve parallel simulations. In addition, with the growing complexity of the hardware there is also a growing interest for multi-tenancy and for a more dynamic, cloud-like execution environment. All these trends bring together a large variety of runtime components that do not cooperate well with each other, which in turn can lead to suboptimal performance. This talk will enumerate a number of representative workloads that stress the limitations of the traditional HPC center. We then highlight some of the underlying forces which shape requirements of next generation systems and propose a cross-stack coordination layer that aims to resolve these conflicts. Finally, through some of our previous efforts in this space we demonstrate the benefits of the overall approach.

3.8 Challenges of Resource Management on the “Data Platform”: mdx

Toshihiro Hanawa (University of Tokyo, JP)

mdx is the infrastructure for leveraging data all over Japan that enables (1) a rapid PoC environment for R&D data leveraging activities including industry-academia-government collaboration projects (2) wide-area virtual private infrastructure with high performance computing and storage resources (3) realtime data processing with security.

mdx introduces a virtualization technology like multi-tenant cloud. On the other hand due to the resource limitation, we have to efficiently manage the resource and consider to offload heavy-load, large-capacity processing to Supercomputer.

3.9 InvasiC HPC Programming: iMPI and EPOP

Jophin John (TU München, DE) and Santiago Narvaez Rivas (TU München, DE)

As more and more emphasis is given to the malleable resources and adaptive resource management, it is necessary to facilitate programming models that enable the application programmers to exploit this dynamism. Our talk focussed on programming models for malleable application development, specifically the malleable MPI API (iMPI) provided by the invasive infrastructure developed as part of the InvasIC project (TCRC 89 “Invasive Computing”). A tsunami simulation code (eSamoa) was adapted using such extension, showing that, albeit the runtime of malleable applications might increase with respect to static ones, the resources are used more efficiently. During the development of the eSamoa, several challenges were discovered. In particular, both the redistribution of data after an adaptation occur, as well as keeping the logical flow of the application consistent, proved to be relevant. To tackle the latter, we proposed a new phase-oriented programming extension, namely the Elastic Phase Oriented Programming (EPOP) model on top of iMPI, that facilitates easier malleable application development. We also discussed a scenario of system-level power management using the EPOP based malleable model.

Along with application malleability, our discussion also extended to dynamic fault-tolerant systems for malleable and non-malleable distributed-memory applications. We talked about iCheck, an invasive checkpointing system that could dynamically increase and decrease the resources for checkpointing. Additionally, using such a system for data redistribution among malleable applications will benefit malleable application development. We showed results that emphasize dynamism and provides better and faster checkpointing abilities.

3.10 From GEOPM to the OIEP Reference Model: Embedding Energy and Power Runtime Systems into the Big Picture of HPC

Matthias Maiterth (TU München, DE)

The talk briefly introduces GEOPM, PowerStack and the OIEP Reference Model, and shows their connection.

The presentations initially gives an overview of the GEOPM runtime and shows the software infrastructure provided by the GEOPM framework.

Since tools do not exist in isolation, a sound setup of tools (such as GEOPM) have to exist in a software echo system. The definition of a software stack for energy and power was addressed by the PowerStack effort with still an open outcome.

The later part of the presentation shows the presenters Dissertation work, where the OIEP reference model is presented, giving vocabulary and a method for representing and arranging energy and power management setups in HPC. This is a missing foundation for efforts such as the PowerStack and other tools, which so far often only consider specialized setups or even lack integration in a holistic energy and power management setup.

3.11 Converged Computing: Combining the Best of HPC and the Cloud

Daniel John Milroy (LLNL – Livermore, US)

Since the early 2000s, computing has relied on increasing levels of parallelism together with Moore's law to drive performance improvement. As Moore's law now begins to taper, demand for increasing performance and new capabilities is spurring development of heterogeneous and dynamic systems and new software environments. Large-scale scientific applications are adapting to use the new tools and technologies and pushing computing boundaries through multi-component workflows.

This talk describes work on facilitating cutting-edge current and next-generation scientific workflows through integration of cloud computing with Flux, a novel graph-based Resource and Job Management Software (RJMS) developed at LLNL. The integration is aimed to advance converged computing, an environment that offers the best features of HPC (performance, efficiency, sophisticated scheduling) and the cloud (resiliency, elasticity, portability, and automation) to next-generation high-performance workflows. The talk will also detail work to build industry collaborations to make lasting, sustainable contributions to the broader computing community.

3.12 Overview of State-of-the-Art Parallel Performance Measurement and Analysis Tools for heterogeneous systems

Bernd Mohr (Jülich Supercomputing Centre, DE)

Current HPC systems consist of complex configurations of potentially heterogeneous components. In addition, the hard- and software configuration can change dynamically due to fault recovering processes or power saving efforts. Deep hierarchies of large, complex software components are needed to operate them. Developing efficient and high performance application software for these systems is challenging. Therefore, sophisticated performance measurement and analysis capabilities are required. The talk will present an overview of state-of-the-art parallel performance measurement and analysis tools, high-lightening the scalability of the tools and their support for current heterogeneous node architectures.

3.13 Dynamism in HPC Resource Control

Frank Mueller (North Carolina State University – Raleigh, US)

This talk provides an overview of our recent work on dynamic resource control in HPC environments. First, power controls are discussed for phase-changing applications. Second, performance and power for application execution over hybrid DRAM/non-volatile memory is characterized and used to provide guided allocation within both memory spaces to reduce energy while maintaining performance. Third, a method for co-scheduling of jobs on multiple heterogeneous accelerators is developed. Fourth, HPC resilience is extended to workflows and jobs are scheduled on heterogeneous nodes according to their resource needs to trade off response time, utilization, and cost for HPC and cloud allocations. Overall, HPC resource control is becoming increasingly dynamic. Future work needs to coordinate different control mechanisms to achieve higher-level objectives in terms of workload and center objectives.

3.14 Dynamic Tuning of HPC Applications – ESPRESO FEM Library

Lubomir Riha (VSB-Technical University of Ostrava, CZ)

This talk introduces the ESPRESO FEM library developed at IT4Innovations. The library was described from computer science point of view, and we highlighted its potential for dynamic resource management. The key component of ESPRESO that enables its elasticity is the I/O module which is capable of checkpoint / restart simulation on various number of MPI ranks. Finally, we have proposed changes needed in this module to fully support iMPI.

3.15 Invasive Computing – A Systems Programming Perspective

Wolfgang Schröder-Preikschat (Universität Erlangen-Nürnberg, DE)

Invasive Computing is a research program that aims at developing a new paradigm to address the hardware- and software challenges of managing and using massively-parallel MPSoCs of the years 2020 and beyond. The program follows the idea of a corresponding Transregional Collaborative Research Center funded by the DFG in its third four-year period (2018-2022). It currently comprises seventeen projects from the areas of computer architecture, system software, programming systems, algorithm engineering and applications. Basic concept is to let applications manage the available computing resources on a local scope and to provide means for a dynamic and fine-grained expansion and contraction of parallelism. This talk provides a brief overview of the program and presents thoughts and intermediate results on system software support for it.

3.16 From COMM_WORLD to Sessions and Bubbles: How new MPI Features Need to Interact with Resource Managers

Martin Schulz (TU München, DE)

MPI 4.0 introduced the new concept of MPI Sessions, which enables a new way of MPI initialization and resource management. While currently still defined as an interface for static resources, it provides the needed base mechanisms to develop a more dynamic view. In this talk I will introduce the MPI Sessions API and its implications, and will then discuss options for extensions, which could provide a truly dynamic and malleable MPI.

3.17 Dynamic Tuning of HPC Applications – MERIC

Ondrej Vysocky (VSB-Technical University of Ostrava, CZ)

This talk presents the MERIC tool for dynamic tuning of HPC hardware or runtime systems while running parallel application. The goal is to minimize the energy to solution with user-defined impact on application performance. Additionally, we also discuss the potential of tuning the hardware under power-cap which not only opens opportunity for energy savings but also for performance improvements. As the tool is continuously used to evaluate potential of energy savings for different applications under H2020 and EuroHPC projects it is being extended with new features. This includes support for new hardware, such as GPUs or new CPU architectures as presented. As the approach we use can perform dynamic tuning at relatively high rate, it is suitable for dynamic resource management.

3.18 The Price Performance of Performance Models

Felix Wolf (TU Darmstadt, DE)

To understand the scaling behavior of HPC applications, developers often use performance models. A performance model is a formula that expresses a key performance metric, such as runtime, as a function of one or more execution parameters, such as core count and input size. Performance models offer quick insights on a very high level of abstraction, including predictions of future behavior. In view of the complexity of today's applications, which often combine several sophisticated algorithms, creating performance models manually is extremely laborious. Empirical performance modeling, the process of learning such models from performance data, offers a convenient alternative, but comes with its own set of challenges. The two most prominent ones are noise and the cost of the experiments needed to generate the underlying data. In this talk, we will review the state of the art in empirical performance modeling and investigate how we can employ machine learning and other strategies to improve the quality and lower the cost of the resulting models.

4 Outlook: Techniques for Malleability-Enabled Machines

This section is the outcome of a joint effort of the seminar participants to provide a summary of the results of the fruitful discussions during the seminar. The section covers all the aspects of introducing dynamic resource management for malleable applications on HPC systems.

4.1 Resource Management

The dynamic behavior of malleable applications introduces new requirements to the Resource Management (RM) infrastructure. Current RM implementations are static, as a consequence of our traditional workloads being static themselves. This aspect of our workloads has been changing over the years. The amount of dynamism varies, and malleability support has become important for a subset of current highly dynamic workloads.

Many of these workloads are developed with MPI, with an increasing number of them using emerging programming models. Because of this, RM malleability support should be programming model agnostic. This can be achieved by the use of PMIx for RM, runtime system, tools and application interactions.

New features are more likely to be accepted by emerging RM infrastructures. Well established workload managers, such as Slurm, are static in design, and large scope changes to it may be undesirable; therefore, we expect malleability features to be more easily embraced by emerging workload managers, such as Flux.

Job and application level malleability specifications are necessary. For example, instead of specific resource specifications, valid ranges should be specified. These can be more than just compute resources, and can include, for example, memory and energy requirements.

The dynamic behavior of these workloads, together with the increasing parallelism of nodes (e.g., provided by GPUs), lowers the likelihood that hardware resources will be properly utilized by a single application. Co-scheduling is a way to improve node-local resource utilization, by allowing applications from multiple jobs to share nodes.

These changes will likely require updates to accounting mechanisms. The resources used by users need to be tracked in time, instead of being trivially determined on job starts. Instead of node-hours, if co-scheduling is enabled, slices of nodes will need to be tracked. Furthermore, since there may be detrimental effects of co-scheduling and malleability, accounting incentives need to be provided. Finally, node-slice counts will vary in time, if malleability is enabled.

Our systems will need to have configuration options that enable the definition of new system policies. For example, a system should be able to prioritize new job starts over resource allocation expansions via malleability, and vice versa. It should also be possible to prioritize based on a new job taxonomy that includes types of jobs that are malleable.

4.2 Programming Models: MPI and Beyond

To support malleability, applications themselves need to become malleable and for that need matching mechanisms in the parallel programming systems and APIs. This discussion can be split into four main categories: cross-node support based on MPI, on-node support based on shared memory models, coarse-grain support for workflows and models for heterogeneous systems. Additionally, combinations of these aspects will have to be considered and will add additional complexity and dependencies in the design of complete system-wide programming abstractions.

4.2.1 Support for on-node malleability and the benefits of tasking

Intra-node programming models require less effort to support malleability as there is no need for data redistribution in the event of a compute resources change. For example, the OpenMP language constructs and runtime system already allows the use of different numbers of threads for different (executions of) parallel regions. The same is true for shared-memory applications using a task-based approach (like OmpSS, StarPU, OpenMP tasks) where the computation is described as a set of tasks and the dependencies between them. The actual execution is then left to the runtime system. Enabling malleability can be supported for all applications (without the need to change them) by extending the runtime system to work together with the resource manager. However, this is only useful for HPC systems which would allow more than one application to use the same node (co-scheduling). This is rare on today's HPC production systems, which are optimized for high performance, due to the fear of too much interference between the different applications on the same node resulting in overall bad performance.

4.2.2 Support for coarse-grained malleability in Workflows

Aside from malleability of single applications, the malleability in workflows is a major topic. In this case, it may be sufficient for applications (in the sense of workflow components) to be simply moldable, but it must be possible to shift resources dynamically between different elements of the workflow. For this to be successful, though, it is critical for the resource manager to understand the workflow and its control flow and to be able to take this (in its entirety) into account when scheduling individual elements of the workflow. For this, the needed interfaces need to be designed and possibly standardized, across different workflow and management systems. These interfaces could then be made available via system resource managers like SLURM and Flux.

4.2.3 Malleability in heterogeneous systems

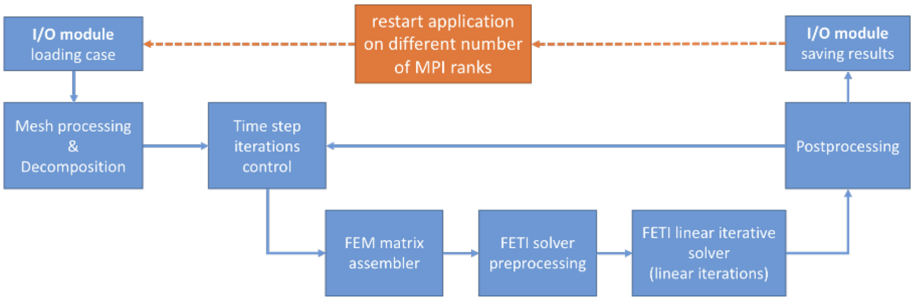
Heterogeneous systems (e.g., integrated systems with homogeneous nodes, but a range of accelerators in each, or modular systems with heterogeneous nodes, each hosting a different accelerator) pose additional challenges, but also offer additional opportunities for applying the concept of malleability. On one side, they add additional resource components and turn scheduling into a multi-objective problem, but on the other side this provides new flexibilities, especially when combined with more fine-grained workload and task descriptions that can be mapped to resources in a flexible manner. As also already discussed wrt. on-node malleability, it will likely lead to the need to support co-scheduling on nodes in order to allow the resource manager to tap the individual resource types on a node independently. Further, aspects like granularity, resource availability, contention and location have to be taken into account. A special aspect is added to the problem when also considering power and energy limitations, especially if power is limited to the point that not all accelerators (or more general, all resources) on a node can be powered at the same time and tradeoffs have to made to shift computational power between heterogeneous resources.

4.3 Unified Monitoring

Malleable and adaptive systems need to collect a variety of performance data to allow allocation and scheduling decisions to be made during runtime. Our premise is that this information could also be used by application performance and visualization tools to provide useful information to programmers as to what changes in their program might result in more efficient uses of these adaptive systems.

We see a research agenda that could proceed in the following steps:

1. Identify what application and system performance information is currently being collected by adaptive systems and runtimes to enable malleability.
2. Design an interface so that tools can access and monitor this information.
3. Develop abstractions and mechanisms to deliver this information in forms used by current performance tools, including tracing, sampling, statistical summarization.
4. Study how to combine this adaptive system performance data with data provided by traditional performance tools.


I/O module
loading case
restart application on different number of MPI ranks
I/O module
saving results
Mesh processing & Decomposition
Time step iterations control
Postprocessing
FEM matrix assembler
FETI solver preprocessing
FETI linear iterative solver (linear iterations)

5. Study how the techniques used by performance tools for more static applications need to be extended or changed to provide an understanding of programs running on an adaptive system.
6. Study the security and privacy issues associated with this new source of performance information and develop strategies for avoiding information leakage.

4.4 Turning Applications into Malleable Applications

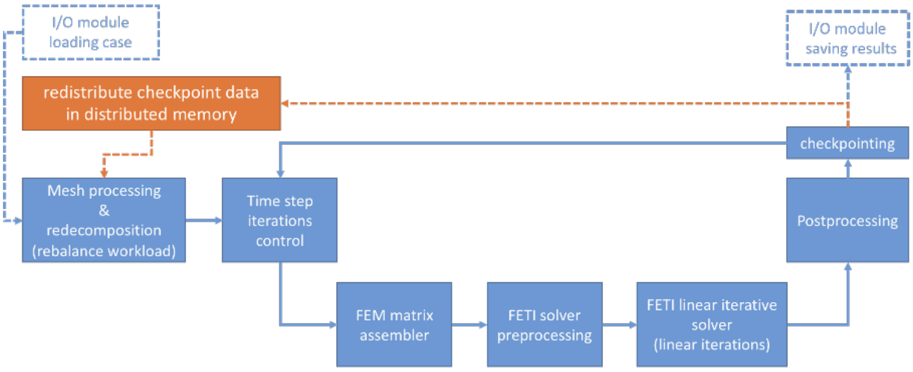
There might be more or less suitable applications for converting them into malleable. If we want to identify applications which can be in “reasonable” time and effort converted into malleable we should identify whether following capabilities are already implemented: (1) checkpoint and restart and (2) load balancing (in the best case the dynamic one). In particular, if an application is able to checkpoint at N MPI ranks and restart at M ranks it means that it already contains the functionality related to redistribution of its data structures. In the simplest scenario this can be supported through a shared filesystem that is used to store the checkpointed data.

An example of this simple scenario is shown in the figure below on an ESPRESO FEM application developed at IT4Innovations NSC in Czech Republic (http://espreso.it4i.cz). This application is based on a domain decomposition method and uses several variations of parallel FETI linear solvers.

The simple scenario then can be converted into a more advanced one which does not rely on a shared filesystem. This step contains extra work that must be done during the conversion into a malleable application. In case of the ESPRESO the I/O module creates the representation of a checkpoint file which contains the current state of a simulation in a distributed memory before it is saved. This file then can be directly redistributed using MPI communication.

The load balancing is then performed by calling the ParMETIS which repartition the mesh into a new number of domains. After that the mesh is redistributed and transferred into new MPI ranks. Based on mesh redistribution the results data are redistributed, accordingly.

There is also an overhead associated with mesh redistribution as all the FEM related objects generated for particular decomposition have to be rebuilt, i.e. all the FEM matrices have to be assembled again. This in general has to be considered for all applications.


I/O module loading case
redistribute checkpoint data in distributed memory
I/O module saving results
checkpointing
Mesh processing & redecomposition (rebalance workload)
Time step iterations control
Postprocessing
FEM matrix assembler
FETI solver preprocessing
FETI linear iterative solver (linear iterations)

4.5 Incentives to Help in Dynamic Resource Management

Dynamic resource management can be improved using application knowledge. Applications can, non-intrusively, provide information in their jobscript specifications. Moldable applications would specify a set of operation points and not ask for a single specific configuration. Malleable applications would specify their properties like "can shrink", "can grow", probably including limits and costs. In return applications that provide such information can get higher priorities in the scheduling queues (because they help to utilize the machine) and a chance to finish earlier.

Annotations about processing phases are a more intrusive way to help the resource manager. The knowledge about job phases can be used to learn and predict resource usage and provide just the right amount of resources to execute the current phase. The benefit for the application appears when the accounting excludes the unused resources, regardless of whether or not the resources are used elsewhere. With co-scheduling the phase annotations can be used to improve utilization of heterogeneous resources, e.g. to hand over the otherwise unused GPUs to a different application.

Malleable Applications are doing some internal resource management and might know about phases when less or more resources are best to achieve good efficiency. Those applications want to return some of their resources to the resource manager and also want to have a guarantee to get the resources back at a later point in time. This can be achieved by leasing the resources for a certain amount of time. If the resource manager agrees to take the resources, then it would guarantee their availability after the deadline has expired and would not charge the application for the leased resources. Such a mechanism can be used to establish a spot market where the central resource manager as well as the applications negotiate resource access and its costs.

4.6 What is the Potential of Malleable Applications?

The malleability support will obviously improve the total system throughput, however it is still not clear how much we can ideally gain. Assessing the potential improvement or the theoretical upper limit will motivate supercomputing centers, HPC research community, as well as the industry toward the malleability support. Throughout the discussion in the meeting, as a community, we concluded that we should quantify the potential gain for an idealistic scenario, by using a real system job trace collected at a supercomputing center, such as LRZ.

First, to conduct this estimation, we need a metric like scale-time product – here we define it as the integral of the number of nodes over time. For instance, if we can reduce the scale-time product by one Xth for all the jobs by supporting malleability, we can potentially gain X times system throughput improvement (of course, this is an ideal case, i.e., only if the job scheduler can fully utilize the extra resources brought by the malleability support).

As a next step, we should know how much we can potentially reduce the scale-time product for each application in the job trace, and then we can quantify the potential system-level throughput improvement by just integrating them (again, this is an ideal case). However, as it is extremely difficult to estimate it for all the jobs, we should introduce some assumptions, conditions, and so forth.

One option for this is focusing on few large-scale and time-consuming jobs, investigating the malleability opportunities for them, and estimating the potential throughput improvement at system level by the simple calculation stated above. Another option for this is a rather statistical approach, e.g., assuming a distribution function that describes the relationship between the scale-time product and the reduction rate of it by the malleability support, and calculates the potential gain based on it.

Then, the next step should be more realistic, such as estimating the gain for some different scheduling algorithms, which is going to be simulation-based experiments. In this evaluation, we will need the number of nodes as a function of time for each job. In the end, this can be extended to cover other sophisticated resource managements such as co-scheduling or power management.

4.7 Converged Computing

Malleability would enable the integration of interactive workloads with typical HPC jobs on the system. The different characteristics of interactive workloads have to be taken into account, i.e., they come in bursts, they have to be executed immediately, and the length and resource requirements are not specified in the form of a batch script.

The opportunities for the HPC center are that the interactive workloads are excellent candidates for increasing the overall system utilization. Furthermore, new application domains become available for the centers, while these domains would profit from computation on powerful HPC systems.

Three use cases were introduced in the seminar. Argonne is working on integrating HPC into the continuum of resources for edge based IoT applications. Another is to bring entire scientific workflows that combine microservice based components with HPC jobs. At LLNL the combination of Kubernetes resource management and Flux is investigated. TUM is researching automatic distribution of function invocations in serverless computing to the heterogeneous computing continuum of HPC, Cloud, Edge, and IoT devices. Due to the limitations on production HPC systems, the use of webassembly for isolation and optimization for heterogeneous nodes of an HPC system is investigated at TUM.

The malleability of HPC jobs and the dynamic resource management can be used to provide resources even on an HPC system on demand for the interactive tasks, without overprovisioning in low demand periods. The approach taken by different groups is to integrate the resource management of a Spark cluster or Kubernetes with the resource management on the machine, providing additional resources to the interactive cluster if required. These resources can be idle nodes or be taken from malleable applications. In case the interactive cluster shrinks, the nodes are given back to the HPC applications.

Malleability would also be beneficial in case the HPC system would be connected to HPC in the Cloud for overflow computation. Application scaling is a major advantage of Cloud systems, e.g., auto-scaling is used to adapt cloud based services to a changing workload. If malleable HPC applications are provided already on the HPC system these could significantly benefit from the practically unlimited resources in the cloud. On-demand allocation of resources depending on the status of the application matches the inherent strength of the cloud. In the cloud context, the incentive to make your applications malleable would even be much higher than on HPC systems as resource usage is paid in the pay-per-use model.

Participants

- Eishi Arima
TU München, DE
- Eduardo César
Autonomus University of Barcelona, ES
- Isaías Alberto Comprés Ureña
TU München, DE
- Michael Gerndt
TU München, DE
- Jophin John
TU München, DE
- Matthias Maiterth
TU München, DE
- Barton P. Miller
University of Wisconsin-Madison, US
- Bernd Mohr
Jülich Supercomputing Centre, DE
- Frank Mueller
North Carolina State University – Raleigh, US
- Santiago Narvaez Rivas
TU München, DE
- Mirko Rahn
Fraunhofer ITWM – Kaiserslautern, DE
- Lubomir Riha
VSB-Technical University of Ostrava, CZ
- Martin Schulz
TU München, DE
- Anna Sikora
Autonomus University of Barcelona, ES
- Ondrej Vysocky
VSB-Technical University of Ostrava, CZ
- Felix Wolf
TU Darmstadt, DE

Remote Participants

- Dong Ahn
LLNL – Livermore, US
- Andrea Bartolini
University of Bologna, IT
- Pete Beckman
Argonne National Laboratory – Lemont, US
- Mohak Chadha
TU München, DE
- Julita Corbalan
Barcelona Supercomputing Center, ES
- Balazs Gerofi
RIKEN – Kobe, JP
- Toshihiro Hanawa
University of Tokyo, JP
- Shantenu Jha
Rutgers University – Piscataway, US
- Rashawn Knapp
Intel – Hillsboro, US
- Masaaki Kondo
Keio University – Yokohama, JP
- Daniel John Milroy
LLNL – Livermore, US
- Tapasya Patki
LLNL – Livermore, US
- Barry L. Rountree
LLNL – Livermore, US
- Roxana Rusitoru
Arm – Cambridge, GB
- Sakamoto Ryuichi
Tokyo Institute of Technology, JP
- Wolfgang Schröder-Preikschat
Universität Erlangen-Nürnberg, DE

Report from Dagstuhl Seminar 21442

Ensuring the Reliability and Robustness of Database Management Systems

Edited by

Alexander Böhm[1], Maria Christakis[2], Eric Lo[3], and Manuel Rigger[4]

1 **SAP SE – Walldorf, DE,** `alexander@boehm.global`
2 **MPI-SWS – Kaiserslautern, DE,** `maria@mpi-sws.org`
3 **The Chinese University of Hong Kong, HK,** `ericlo@cse.cuhk.edu.hk`
4 **ETH Zürich, CH,** `manuel.rigger@inf.ethz.ch`

Abstract

The goal of this seminar was to bring together researchers and practitioners from various domains such as of databases, automatic testing, and formal methods to build a common ground and to explore possibilities for systematically improving the state of the art in database management system engineering. The outcome of the seminar was a joint understanding of the specific intricacies of building stateful system software, as well as the identification of several areas of future work. In particular, we believe that database system engineering can both be significantly improved by adopting additional verification techniques and testing tools, and can provide important feedback and additional challenges (e.g. related to state management) to neighboring domains.

Seminar November 1–4, 2021 – http://www.dagstuhl.de/21442
2012 ACM Subject Classification Information systems → Data management systems
Keywords and phrases Databases, Reliability, Robustness, Testing, Dagstuhl Seminar
Digital Object Identifier 10.4230/DagRep.11.10.20
Edited in cooperation with Kulahcioglu Ozkan, Burcu

1 Executive Summary

Maria Christakis (MPI-SWS – Kaiserslautern, DE)
Alexander Böhm (SAP SE – Walldorf, DE)
Eric Lo (The Chinese University of Hong Kong, HK)
Manuel Rigger (ETH Zürich, CH)

DataBase Management Systems (DBMSs) are used ubiquitously. Due to the ever-growing number and size of data sets, increasing performance demands, and the virtually unlimited hardware resources that are provided by public cloud infrastructure, sophisticated systems and optimizations are developed continuously. This dynamic and demanding environment is a major challenge for developers of DBMSs, which have to ensure that their systems are both correct and efficient.

Database management systems are a well-established field with several decades of research and engineering attention. These efforts have resulted in a multitude of both open-source and commercial systems that are widely deployed in production today and provide the backbone of a vast range of mission-critical applications. Still, surprisingly, recent work on automatic testing of DBMSs found a large number of bugs in widely-used DBMSs. This

Ensuring the Reliability and Robustness of Database Management Systems, *Dagstuhl Reports*, Vol. 11, Issue 10, pp. 20–35
Editors: Alexander Böhm, Maria Christakis, Eric Lo, and Manuel Rigger
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

clearly indicated that the topic of ensuring the reliability and robustness of DBMS deserves more attention, and that key insights from neighboring domains such as automatic testing and formal methods could potentially help to advance the state of the art in DBMS engineering.

Goals and Outcomes

One of the central goals and outcomes of the seminar was to build a common foundation and understanding for the key challenges of DBMS engineering, and how they can be potentially addressed. To this end, the seminar focused on

- Best practices and challenges in building open source and commercial database engines. Here, the key objectives include a high developer efficiency, mandating quick feedback by tests and verification tools already during feature development, as well as systematic (stress) testing of the software under high load and error conditions.
- The applicability of formal methods and verification tools to DBMS.
  Formal methods can be of great help to prove the correctness of key database system components such as query compilers, distributed consensus protocols, data replication components, or modules dealing with high availability. Still, an important question is how to systematically identify those components that can benefit from formal verification with reasonable implementation effort, and how to best integrate these methods into existing systems.
- Advanced testing techniques such as fuzzers, query synthesis, and workload generators.
  These methods allow to significantly increase the test coverage of a DBMS by systematically exploring uncovered code paths and putting stress on individual, important subsystems such as input verification and error handling that are a frequent source of software defects.
- Methods for the automatic generation of test data and testcase reduction.
  Occasionally, defects in database software are only found by customers running very complex queries operating on confidential data sets. Thus, to allow for problem reproduction, developers benefit from a minimal data set and a simplified query specification that does not disclose confidential data or exhibit unnecessary complexity.
- Security aspects such as ensuring confidentiality and data integrity in the presence of different classes of attackers.

Attendee Mix and Seminar Structure

The seminar lasted 2.5 days. Its format and attendee mix was significantly influenced by the ongoing pandemic. Of the 34 attendees, 13 attended in person and 21 remotely. All but one of the in-person attendees were based in Europe. Overall, we received the highest response rate from Europe (20 attendees), and a lower one from Asia (8 attendees) and the US (6 attendees). We are grateful to the two Video Conference Assistants (VCAs), Jack Clark and Mark Raasveldt, who managed the equipment to ensure a smooth experience for all attendees.

We started the seminar with an introduction round in which every attendee introduced themselves. We held another such session in the late afternoon, to accommodate the US attendees. Prior to the seminar, we contacted attendees to give overview talks to establish a common discussion basis, which was useful given that the attendees came from different scientific communities. We had such overview talks on the first and second day. On the second and third day, we had in-depth talks. While we had planned breakout sessions, many of the talks were followed by fruitful and unplanned discussions. On the last day, we had a group discussion on the takeaways and future plans.

Future Plans

One major result from the seminar was to identify open problems and areas of future work that the group wants to address in an interdisciplinary manner. Among others, this includes the creation of a reference manual for database engineering groups to avoid redundant work and re-inventing techniques already established (or discarded) by other teams, the identification of database modules (e.g. the query compiler and transaction processing system) that can benefit from formal verification, designing new test oracles to test various data-centric systems for different kind of bugs, as well as the establishment of a common testcase specification format and a test corpus that can be shared between DBMS engineering teams. We discussed proposing another instance of the Dagstuhl seminar to utilize the established discussion basis and work on addressing these specific challenges.

2 Table of Contents

Executive Summary

Maria Christakis, Alexander Böhm, Eric Lo, and Manuel Rigger 20

Overview of Talks

Dynamic Symbolic Execution: An Introduction
Cristian Cadar 25

How to Make Serializable Concurrency Control Protocol Executions Verifiable
Jack Clark 25

Three Ways to Get Test Case Reduction Almost for Free
Alastair F. Donaldson 26

Formal Verification of Databases
Stefania Dumbrava 27

Database Security: Formalization, Verification, and Testing – Challenges and Open Questions
Marco Guarnieri 28

Robust Query Execution with Guarantees
Jayant R. Haritsa 28

Isolation Levels and Isolation Level Testing
Kyle Kingsbury 28

Hyper's Reliability and Robustness Challenges
Marcel Kost 29

Testing Consensus Implementations
Burcu Kulahcioglu Ozkan 29

Metamorphic Testing of Datalog Engines
Muhammad Numair Mansur, Maria Christakis, and Valentin Wüstholz 30

DuckDB Testing – Present and Future
Mark Raasveldt and Hannes Mühleisen 30

A Whirlwind Tour of Automated Database Management System Testing
Manuel Rigger 31

Fuzz Testing
Abhik Roychoudhury 31

Big Data, Small Testing
Anupam Sanghi 31

Working groups

Towards Bug-Free DBMS Ecosystems
Mai Zheng, Jack Clark, and Miryung Kim 32

Findings

General Conclusions 33

Open Problems and Future Work 34

Participants 35

Remote Participants 35

3 Overview of Talks

3.1 Dynamic Symbolic Execution: An Introduction

Cristian Cadar (Imperial College London, GB)

In this overview talk, I give an introduction to dynamic symbolic execution, discussing its key strengths, as well as its main scalability challenges. I also give an overview of some of the many applications of dynamic symbolic execution, and discuss the opportunities it offers for analysing database management systems.

3.2 How to Make Serializable Concurrency Control Protocol Executions Verifiable

Jack Clark (ETH Zürich, CH)

Main reference Jack Clark: "Verifying Serializability Protocols With Version Order Recovery", ETH Zürich, 2021.
URL https://doi.org/10.3929/ethz-b-000507577

A core feature of many database systems is the ability to group operations into transactions. An isolation level defines the extent to which operations within a transaction interact with operations from other concurrent transactions. The serializable isolation level provides correctness guarantees that many programmers implicitly assume, since it provides the illusion of transactions running in some sequential order. However, implementing serializable transactions, particularly in a distributed setting, has proven to be a challenging task, with many systems failing to live up to their guarantees.

Unfortunately, verifying that an execution history is serializable is NP-complete. However, with access to a database system's internal version order, serializability can be efficiently checked. Existing tools for checking the serializability of histories either have exponential running time or require specially crafted operations to be able to recover version order information, which significantly limits the amount of functionality that can be tested. Furthermore, existing tools cannot handle predicate operations which are a key feature of most database systems.

This talk demonstrates that it is possible to recover the version order directly from real database systems and that it can be used to efficiently verify execution histories. Additionally, it is shown that recovering additional object visibility information enables verification of histories that contain predicate operations, a condition which distinguishes the serializable isolation level from weaker levels. Building on these foundations, I demonstrate how we can move towards verifying histories generated by real-world workloads, something not achievable with existing tools and techniques.

3.3 Three Ways to Get Test Case Reduction Almost for Free

Alastair F. Donaldson (Imperial College London, GB)

Joint work of Alastair F. Donaldson, Paul Thomson, Vasyl Teliman, Stefano Milizia, André Perez Maselco, Antoni Karpinski
Main reference Alastair F. Donaldson, Paul Thomson, Vasyl Teliman, Stefano Milizia, André Perez Maselco, Antoni Karpinski: "Test-case reduction and deduplication almost for free with transformation-based compiler testing", in Proc. of the PLDI '21: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, Virtual Event, Canada, June 20-25, 2021, pp. 1017–1032, ACM, 2021.
URL http://dx.doi.org/10.1145/3453483.3454092

Randomised testing techniques are good at finding bugs, but tend to produce large bug-inducing test cases that are difficult to understand. Test case reduction is an essential technology for increasing the utility of randomised testing. A test case reduction tool takes a large bug-inducing test case and shrinks it to a smaller test case that still triggers the bug of interest, typically using a variant of the delta debugging algorithm [5].

Test case reduction is challenging when test cases have associated validity constraints. For example, a test case that triggers a miscompilation bug in a C compiler must be free from undefined behaviour. If a test case reducer introduces undefined behaviour when reducing a test case, the result of test case reduction may end up being a nonsensical C program that yields different results across multiple compilers, but due to undefined behaviour in the program rather than due to a compiler bug. The C-Reduce test case reducer for C programs [4] relies on a plethora of external tools to ensure test case validity.

In this talk I give an overview of three ways that test case reduction can be obtained almost for free in a manner that preserves test case validity automatically.

The first is based on a technique called transformation-based testing, used in the spirv-fuzz tool [1]. In this approach a test case comprises an original input and a mutated input where the mutated input is obtained by applying a sequence of semantics-preserving transformations to the original input. A bug is identified when the system under tests treats these equivalent-by-construction inputs differently. Test case reduction then involves using delta debugging to search for a minimal sub-sequence of transformations such that the original and minimally-transformed inputs yield different results. Test case reduction is thus applied on the transformations, rather than on the input; as a result, the approach relies on fuzzing using relatively small seed inputs.

The second is the test case reduction approach employed by the Hypothesis property-based testing tool for Python [3]. This involves performing test case reduction on the sequence of bits that was used to generate a bug-inducing input, searching for a shorter, simpler sequence of bits that, when fed to the generator, yields a smaller, simpler input that still triggers the bug. Because reduced test cases are emitted by the same generator that emitted the original bug-inducing test case, reduced test cases automatically enjoy any validity guarantees that the generator provides.

The final approach, specific to compiler testing, involves using "program reconditioning". With this approach, undefined behaviours are eliminated from a program after the program is generated, rather than being avoided during generation. Treating the removal of undefined behaviour as a separate "reconditioning" step means that reconditioning can also be used during test case reduction: the test case reducer need not worry about introducing undefined behaviour because before feeding a reduced input to the compilers under test, the input will be reconditioned. This idea was the subject of a recent MSc thesis [2] and is the subject of ongoing work.

My hope is that these ideas may have relevance in the field of database testing, if similar problems of test case validity apply.

References

1 Alastair F. Donaldson, Paul Thomson, Vasyl Teliman, Stefano Milizia, André Perez Maselco, and Antoni Karpinski. Test-case reduction and deduplication almost for free with transformation-based compiler testing. In Stephen N. Freund and Eran Yahav, editors, *PLDI '21: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, Virtual Event, Canada, June 20-25, 2021*, pages 1017–1032. ACM, 2021.

2 Bastien Lecoeur. GLSLsmith: A random generator of OpenGL shader programs. Master's thesis, Imperial College London, 2021.

3 David Maciver and Alastair F. Donaldson. Test-case reduction via test-case generation: Insights from the hypothesis reducer (tool insights paper). In Robert Hirschfeld and Tobias Pape, editors, *34th European Conference on Object-Oriented Programming, ECOOP 2020, November 15-17, 2020, Berlin, Germany (Virtual Conference)*, volume 166 of *LIPIcs*, pages 13:1–13:27. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020.

4 John Regehr, Yang Chen, Pascal Cuoq, Eric Eide, Chucky Ellison, and Xuejun Yang. Test-case reduction for C compiler bugs. In Jan Vitek, Haibo Lin, and Frank Tip, editors, *ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI '12, Beijing, China – June 11 – 16, 2012*, pages 335–346. ACM, 2012.

5 Andreas Zeller and Ralf Hildebrandt. Simplifying and isolating failure-inducing input. *IEEE Trans. Software Eng.*, 28(2):183–200, 2002.

3.4 Formal Verification of Databases

Stefania Dumbrava (ENSIIE – Paris & SAMOVAR – Evry, FR)

We highlight three applications of formal methods to building formally verified specifications of relational, deductive, and graph database query engines. We focus, in particular, on theorem proving with the Coq proof assistant. First, in the relational setting, we give an overview of the state of the art and focus on the formal executable specification of the relational database model [1]. Second, in the deductive setting, we present our experience on developing a modular, reusable library of certified engines for different Datalog dialects (stratified and regular), using the SSReflect extension of Coq [2, 3]. Finally, we discuss the proof engineering aspects related to building an incremental engine, capable of performing view maintenance in the graph database setting. We conclude by outlining future perspectives on using such correct-by-construction specifications as trusted oracles against which one can test commercial implementations.

References

1 Véronique Benzaken, Evelyne Contejean, Stefania Dumbrava. A Coq Formalization of the Relational Data Model. In Zhong Shao, editor, *Programming Languages and Systems – 23rd European Symposium on Programming, ESOP 2014, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2014, Grenoble, France, April 5-13, 2014, Proceedings, volume 8410 of Lecture Notes in Computer Science*, pages 189–208. Springer, 2014.

2 Véronique Benzaken, Evelyne Contejean, Stefania Dumbrava. Certifying Standard and Stratified Datalog Inference Engines in SSReflect. In Mauricio Ayala-Rincón and César A. Muñoz, editors, *Interactive Theorem Proving – 8th International Conference, ITP 2017,*

Brasília, Brazil, September 26-29, 2017, Proceedings, volume 10499 of Lecture Notes in Computer Science, pages 171–188. Springer, 2017.

3 Angela Bonifati, Stefania Dumbrava, Emilio Jesús Gallego Arias. Certified Graph View Maintenance with Regular Datalog. *Theory Pract. Log. Program.*, 18(3-4):372–389, 2018.

3.5 Database Security: Formalization, Verification, and Testing – Challenges and Open Questions

Marco Guarnieri (IMDEA Software – Madrid, ES)

Securing database systems is critical to protect the confidentiality and integrity of the data stored in databases as well as to ensure the security of applications built on top of databases. In this talk, I present an overview of different classes of attacker models for database systems ranging from honest-but-curious attackers that interact with a database following the SQL standard to attackers compromising database-backed applications. For each attacker, I overview existing security mechanisms, attacks bypassing them, and I discuss challenges in testing and verification of security mechanisms.

3.6 Robust Query Execution with Guarantees

Jayant R. Haritsa (Indian Institute of Science – Bangalore, IN)

Robust query processing with strong performance guarantees is an extremely desirable objective in the design of industrial-strength database engines. However, it has proved to be a largely intractable and elusive challenge despite sustained efforts spanning several decades. In this talk, we show how the use of a radically different approach involving application of geometric techniques on execution-time profiles can provide provable guarantees on worst-case performance. Further, these guarantees are independent of data distributions and query structures.

3.7 Isolation Levels and Isolation Level Testing

Kyle Kingsbury (San Francisco, US)

Users who care about their data store it in databases, which (at least in principle) guarantee some form of transactional isolation. However, experience shows that many databases do not provide the isolation guarantees they claim. With the recent proliferation of new distributed databases, demand has grown for checkers that can, by generating client workloads and injecting faults, produce anomalies that witness a violation of a stated guarantee. An ideal checker would be sound (no false positives), efficient (polynomial in history length and concurrency), effective (finding violations in real databases), general (analyzing many patterns of transactions), and informative (justifying the presence of an anomaly with understandable counterexamples). Sadly, we are aware of no checkers that satisfy these goals. We present Elle: a novel checker which infers an Adya-style dependency graph between client-observed transactions. It does so by carefully selecting database objects and operations when generating histories, so as to ensure that the results of database reads reveal information about their version history. Elle can detect every anomaly in Adya et al's formalism (except for predicates), discriminate between them, and provide concise explanations of each.

3.8 Hyper's Reliability and Robustness Challenges

Marcel Kost (Salesforce – München, DE)

Hyper is the data engine of Tableau, an interactive visual analytics application. It transparently aggregates the user's data using the SQL queries generated by the visual frontend, which can become arbitrarily complex.

Hyper is known for its code generation, where SQL queries are compiled to machine code for faster execution. Making this technique robust is a major challenge, since this makes the DBMS prone to hard crashes, and many existing tools like sanitizers can't be applied to run-time generated code.

In addition Hyper is used as part of Tableau's cloud offering, which introduces a large number of cloud-related reliability challenges. The biggest challenge besides general cloud-native problems like deployment and monitoring is efficiently handling multitenancy while still guaranteeing high availability. To master this challenge, a lot of work in resource governance and load balancing is required, probably leading to distributed query processing.

3.9 Testing Consensus Implementations

Burcu Kulahcioglu Ozkan (TU Delft, NL)

Joint work of Cezara Dragoi, Constantin Enea, Burcu Kulahcioglu Ozkan, Rupak Majumdar, Filip Niksic
Main reference Cezara Dragoi, Constantin Enea, Burcu Kulahcioglu Ozkan, Rupak Majumdar, Filip Niksic: "Testing consensus implementations using communication closure", Proc. ACM Program. Lang., Vol. 4(OOPSLA), pp. 210:1–210:29, 2020.
URL http://dx.doi.org/10.1145/3428278

Distributed database systems rely on the coordination and agreement of distributed nodes to provide certain guarantees to the application developers. However, it is difficult to design and implement distributed systems correctly to meet their guarantees. They are error-prone since developers have to reason about a large number of possible event interleaving's due to asynchronous message-based communication, arbitrary message delays, message losses, network partitions, and node failures.

Current testing techniques focus on systematic or randomized exploration of all executions of an implementation while treating the implemented algorithms as black boxes. On the other hand, proofs of correctness of many of the underlying algorithms often exploit semantic properties that reduce reasoning about correctness to a subset of behaviors. For example, the communication-closure property, used in many proofs of distributed consensus algorithms, shows that every asynchronous execution of the algorithm is equivalent to a lossy synchronous execution, thus reducing the burden of proof to only that subset. We formulate the communication-closure hypothesis, which states that bugs in implementations of distributed consensus algorithms will already manifest in lossy synchronous executions, and present a testing algorithm based on this hypothesis. We show that a random testing algorithm based on sampling lossy synchronous executions can empirically find a number of bugs, including previously unknown ones.

In this talk, I introduce the key ideas in our algorithm for testing consensus implementations that are fundamental in many distributed database systems.

3.10 Metamorphic Testing of Datalog Engines

Muhammad Numair Mansur (MPI-SWS – Kaiserslautern, DE), Maria Christakis (MPI-SWS – Kaiserslautern, DE), and Valentin Wüstholz

Datalog is a popular query language with applications in several domains. Like any complex piece of software, Datalog engines may contain bugs. The most critical ones manifest as incorrect results when evaluating queries—we refer to these as query bugs. Given the wide applicability of the language, query bugs may have detrimental consequences, for instance, by compromising the soundness of a program analysis that is implemented and formalized in Datalog. In this talk, I present the first metamorphic-testing approach for detecting query bugs in Datalog engines. We ran our tool on three mature engines and found 13 previously unknown query bugs, some of which are deep and revealed critical semantic issues.

3.11 DuckDB Testing – Present and Future

Mark Raasveldt (CWI – Amsterdam, NL) and Hannes Mühleisen (CWI – Amsterdam, NL)

DuckDB is a fast analytical embedded database system. For users it is crucial that the system is both correct and fast. In this talk we discuss how the different components of the extensive test suite for DuckDB works, and how we use it to ensure the reliability of the system. In addition, we also talk about the fuzzers that we run and the way in which we do performance regression testing.

3.12 A Whirlwind Tour of Automated Database Management System Testing

Manuel Rigger (ETH Zürich, CH)

One prime approach to ensuring the reliability of Database Management Systems (DBMSs) is automated testing. Automated testing can find bugs without user interaction, but not guarantee their absence. Generally speaking, such approaches involve generating a test case, validating the test case's result (which is known as a test oracle), and then reducing the bug to a minimal version. Testing DBMSs typically involves generating a database, a query, and then validating the query's result. In this talk, I will give an overview of the automated testing landscape focusing on the available test oracles. My goal is to convey a simplified overview of the ideas that have been tried and point out works by the seminar's attendees and how they connect. Furthermore, I will do a deep dive into the Ternary Logic Partitioning approach that we designed.

3.13 Fuzz Testing

Abhik Roychoudhury (National University of Singapore, SG)

URL https://www.comp.nus.edu.sg/ abhik/projects/Fuzz/

Fuzz testing, proposed by Barton Miller, is a popular technology for finding bugs in software systems via (biased) random search over the domain of inputs. In this talk, we will review the technology of fuzz testing – specifically its variants in the form of blackbox, greybox and whitebox fuzzing. We will discuss coverage based greybox fuzzing, which conducts a biased random search over inputs guided by a fitness function to approximate code coverage. Recent developments in fuzzing such as directed fuzzing to reach specific code locations, and smart fuzzing to test applications processing structured data will be discussed. Overall, we will discuss the pro-s and con-s of fuzzing technology and take a forward looking view where it can be tuned to find deeper bugs such as violations of non-trivial temporal properties.

3.14 Big Data, Small Testing

Anupam Sanghi (Indian Institute of Science – Bangalore, IN)

Joint work of Anupam Sanghi, Raghav Sood, Jayant R. Haritsa, Srikanta Tirthapura, Shadab Ahmed
Main reference Anupam Sanghi, Raghav Sood, Jayant R. Haritsa, Srikanta Tirthapura: "Scalable and Dynamic Regeneration of Big Data Volumes", in Proc. of the 21st International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, March 26-29, 2018, pp. 301–312, OpenProceedings.org, 2018.
URL http://dx.doi.org/10.5441/002/edbt.2018.27

Synthetic databases are required in a variety of industrial use-cases, ranging from testing and tuning database engines and applications to system benchmarking. In the past decade, several frameworks have advocated modeling data synthesis using a set of cardinality constraints.

Specifically, a cardinality constraint dictates that the output of a given relational expression over the generated database should feature a specified number of rows. We begin this talk with an overview of these frameworks and their current limitations. Then, we present in detail, Hydra, our proposed data generation framework. Hydra constructs a minuscule database summary from the input cardinality constraints, and leverages this summary to dynamically generate data during query execution without explicitly instantiating, storing, and loading the entire database. Adopting this online approach helps to eliminate the time and space overheads typically associated with data synthesis. Finally, to complement dynamic generation, Hydra ensures that the summary generation algorithm is data-scale-free, making it suitable for Big Data environments.

4 Working groups

4.1 Towards Bug-Free DBMS Ecosystems

Mai Zheng (Iowa State University – Ames, US), Jack Clark (ETH Zürich, CH), and Miryung Kim (UCLA, US)

This abstract summarizes the results of a virtual group discussion during the Dagstuhl Seminar 21442 "Ensuring the Reliability and Robustness of Database Management Systems". Participants include Jack Clark (ETH Zürich, CH), Miryung Kim (University of California, Los Angeles, US), and Mai Zheng (Iowa State University, US).

Database Management Systems (DBMS) do not work in isolation. They typically rely on the underlying operating systems (OS) to provide file, networking, and other services. For example, Lightening DB stores a B+ tree as a .mdb file in the file system and uses the mmap syscall to map the .mdb file into memory; moreover, it relies on the fsync and other syscalls to achieve ACID (i.e., atomicity, consistency, isolation, durability) guarantees. Therefore, it is important to take the execution environment of DBMS into account when testing DBMS in practice (i.e., the entire DBMS ecosystem).

Unfortunately, the interactions between DBMS and the surrounding system is complicated. For example, SQLite maintains both data and log files in the file system (FS) and uses a combination of unlink, fsync and other syscalls to implement transactions. But the ACID properties of the transactions may be violated unexpectedly for a number of reasons including the ambiguity in POSIX specifications and DBMS configurations [1, 2]. Similar issues have also been observed on other widely used DBMS ecosystems [1, 2]. How to address the inherent dependency and ambiguity in DBMS ecosystems remains an open question.

The two automatic testing approaches above [1, 2] show the preliminary effectiveness, but they fall short of coverage. Recent efforts have applied formal methods to address the challenge. For example, FSCQ shows that it is possible to formally verify the crash consistency property of a complete file system [3]. However, the verified FS is still much simpler than the POSIX file systems (e.g., Ext4) supporting various DBMS in practice. Also, follow-up research has exposed a bug in the verified FS via fuzzing [4], which implies the difficulty in achieving bug-free DBMS ecosystems. Most recently, Amazon [5] applies lightweight formal methods to validate a key-value (KV) store node in Amazon S3 service, which demonstrates the feasibility of verifying the correctness of an entire storage node.

Unfortunately, while promising, the approach cannot be easily extended to general DBMS ecosystems due to a number of constraints (e.g., it is highly customized for Rust-based S3 KV store).

Addressing the challenges of testing DBMS ecosystems will likely require the expertise and collaboration across different communities including databases, formal methods, file systems, etc. Given the prime importance of DBMS ecosystems, we call for communities' collective efforts in examining the cross-layer challenges and coming up with practical solutions.

References

1 Thanumalayan Sankaranarayana Pillai, Vijay Chidambaram, Ramnatthan Alagappan, Samer Al-Kiswany, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. All file systems are not created equal: On the complexity of crafting crash-consistent applications. In Jason Flinn and Hank Levy, editors, *11th USENIX Symposium on Operating Systems Design and Implementation, OSDI '14, Broomfield, CO, USA, October 6-8, 2014*, pages 433-448. USENIX Association, 2014.

2 Mai Zheng, Joseph Tucek, Dachuan Huang, Feng Qin, Mark Lillibridge, Elizabeth S. Yang, Bill W. Zhao, and Shashank Singh. Torturing databases for fun and profit. In Jason Flinn and Hank Levy, editors, *11th USENIX Symposium on Operating Systems Design and Implementation, OSDI '14, Broomfield, CO, USA, October 6-8, 2014*, pages 449-464. USENIX Association, 2014.

3 Haogang Chen, Daniel Ziegler, Tej Chajed, Adam Chlipala, M. Frans Kaashoek, and Nickolai Zeldovich. Using Crash Hoare logic for certifying the FSCQ file system. In Ajay Gulati and Hakim Weatherspoon, editors, *2016 USENIX Annual Technical Conference, USENIX ATC 2016, Denver, CO, USA, June 22-24, 2016*. USENIX Association, 2016.

4 Seulbae Kim, Meng Xu, Sanidhya Kashyap, Jungyeon Yoon, Wen Xu, and Taesoo Kim. Finding semantic bugs in file systems with an extensible fuzzing framework. In Tim Brecht and Carey Williamson, editors, *Proceedings of the 27th ACM Symposium on Operating Systems Principles, SOSP 2019, Huntsville, ON, Canada, October 27-30, 2019*, pages 147-161. ACM, 2019.

5 James Bornholt, Rajeev Joshi, Vytautas Astrauskas, Brendan Cully, Bernhard Kragl, Seth Markle, Kyle Sauri,Drew Schleit, Grant Slatton, Serdar Tasiran, Jacob Van Geffen, and Andrew Warfield. Using lightweight formal methods to validate a key-value storage node in Amazon S3. In Robbert van Renesse and Nickolai Zeldovich, editors, *SOSP'21: ACM SIGOPS 28th Symposium on Operating Systems Principles, Virtual Event / Koblenz, Germany, October 26-29, 2021*, pages 836-850. ACM, 2021.

5 Findings

In this section, we highlight some of the general conclusions the participants derived from the seminar, and discuss open problems and future work to follow up on.

5.1 General Conclusions

During the discussion, various database developers from both industry and academia shared their quality assurance and testing strategies. In particular, they highlighted both the importance of sophisticated testing strategies for finding defects early, as well as the enormous effort (considering both hardware resources and manual labor) that is put into this part of the system development process. Still, regressions and defects that are only found late in the process or even by customers continue to be a huge challenge even for very mature systems that are in active development for several decades. Despite this obvious importance of the topic, there is **almost no sharing of best practices between database development teams** and only a few publications on the topic. For commercial vendors, their quality assurance strategy is usually seen as a business secret, whereas software quality topics are often not considered interesting enough to publish by open-source software teams and academic development groups.

While **formal methods and verification techniques** have shown their practical benefits in many domains such as aviation, embedded, and real-time system (to only name a few), their **practical application in the database system space is still surprisingly limited**.

Automated testing including sophisticated techniques for defect finding such as semantic fuzzing and error injection plays a key role in most database development teams already today. Still, a **major challenge is to find good oracles that can reliably distinguish between an expected and faulty outcome of a randomly generated, complex testcase**. While the current state of the art is helpful to uncover low-level defects in the system implementations (i.e. software crashes that continue to be an important defect class in many systems), more work is needed in order to uncover more high-level, semantic problems such as wrong query results. In particular, such semantical tests would be of significant help to improve the reliability and robustness of multi-node deployments of databases: These distributed DBMS suffer from additional system complexity by allowing for additional classes of errors such as faulty network communication or requiring distributed coordination protocols between the nodes, to only name two prominent examples.

5.2 Open Problems and Future Work

As mentioned above, the seminar helped the participants to uncover open problems and challenges for the development of reliable and robust database software. Below, we list several topics that participants of the seminar expressed their interest in working on.

There is a clear need to **establish a handbook on best practises for DBMS development teams** that provides comprising guidance of the current state of the art of engineering reliable and robust database systems. We hope that such a reference will allow teams to pick from a bouquet of available techniques depending on their requirements and software complexity. This also includes examples of techniques that were evaluated and did not work out, so that teams can learn from the negative experiences made by other engineering groups.

To address the limited use of formal verification in the DBMS context, we plan to **identify key components in the architecture of DBMS that lend themselves to the use of formal methods**. This potentially includes the query optimizer, where formal methods can help to provide semantic equivalence between optimized and non-optimized query plans, as well as for verifying key characteristics of distributed transaction management, consensus protocols, or replication mechanisms. By providing practical examples from the DBMS domain, we hope to lower the adoption hurdle for system development teams.

Over time, each database management system tends to accumulate a large amount of performance, scalability, and correctness tests. Usually, these tests are in a proprietary format and not shared between systems. By proposing a **standardized format for a high-level test specification** and providing a centralized repository, we hope to foster sharing and re-use between different software development teams.

Participants

- Alexander Böhm
  SAP SE – Walldorf, DE
- Cristian Cadar
  Imperial College London, GB
- Alastair F. Donaldson
  Imperial College London, GB
- Stefania Dumbrava
  ENSIIE – Paris & SAMOVAR – Evry, FR
- Marco Guarnieri
  IMDEA Software – Madrid, ES
- Marcel Kost
  Salesforce – München, DE
- Burcu Kulahcioglu Ozkan
  TU Delft, NL
- Hannes Mühleisen
  CWI – Amsterdam, NL
- Danica Porobic
  Oracle Labs –
  6Redwood Shores, US
- Mark Raasveldt
  CWI – Amsterdam, NL
- Manuel Rigger
  ETH Zürich, CH
- Anupam Sanghi
  Indian Institute of Science –
  Bangalore, IN

Remote Participants

- Artur Andrzejak
  Universität Heidelberg, DE
- Chee-Yong Chan
  National University of Singapore, SG
- Yongheng Chen
  Georgia Institute of Technology –
  Atlanta, US
- Maria Christakis
  MPI-SWS – Kaiserslautern, DE
- Jack Clark
  ETH Zürich, CH
- Jens Dittrich
  Universität des Saarlandes –
  Saarbrücken, DE
- Paolo Guagliardo
  University of Edinburgh, GB
- Jayant R. Haritsa
  Indian Institute of Science –
  Bangalore, IN
- Miryung Kim
  UCLA, US
- Kyle Kingsbury
  San Francisco, US
- Greg Law
  Undo – Cambridge, GB
- Si Liu
  ETH Zürich, CH
- Eric Lo
  The Chinese University of Hong Kong, HK
- Muhammad Numair Mansur
  MPI-SWS – Kaiserslautern, DE
- Zhou Qiang
  PingCAP – Hangzhou, CN
- Tilmann Rabl
  Hasso-Plattner-Institut,
  Universität Potsdam, DE
- Abhik Roychoudhury
  National University of Singapore, SG
- Zhendong Su
  ETH Zürich, CH
- S. Sudarshan
  Indian Institute of Technology –
  Mumbai, IN
- Tao Xie
  Peking University, CN
- Tianyin Xu
  University of Illinois –
  Urbana-Champaign, US
- Mai Zheng
  Iowa State University –
  Ames, US

Report from Dagstuhl Seminar 21451

Managing Industrial Control Systems Security Risks for Cyber Insurance

Edited by

Simon Dejung[1], Mingyan Liu[2], Arndt Lüder[3], and Edgar Weippl[4]

1 **SCOR – Zürich, CH,** sdejung@scor.com
2 **University of Michigan – Ann Arbor, US,** mingyan@umich.edu
3 **Otto-von-Guericke-Universität Magdeburg, DE,** arndt.lueder@ovgu.de
4 **University of Vienna & SBA Research – Wien, AT,** edgar.weippl@univie.ac.at

Abstract

Industrial control systems (ICSs), such as production systems or critical infrastructures, are an attractive target for cybercriminals, since attacks against these systems may cause severe physical damages/material damages (PD/MD), resulting in business interruption (BI) and loss of profit (LOP). Besides financial loss, cyber-attacks against ICSs can also harm human health or the environment or even be used as a kind of weapon. Thus, it is of utmost importance to manage cyber risks throughout the ICS's lifecycle (i.e., engineering, operation, decommissioning), especially in light of the ever-increasing threat level that is accompanied by the progressive digitization of industrial processes. However, asset owners may not be able to address security risks sufficiently, nor adequately quantify them in terms of their potential impact (physical and non-physical) and likelihood. A self-deceptive solution might be using insurance to transfer these risks and offload them from their balance sheet since the underlying problem remains unsolved. The reason for this is that the exposure for asset owners remains and mitigation measures may still not be implemented adequately while the insurance industry is onboarding unassessed risks and covering it often without premium and without managing the potential exposure of accumulated events. The Dagstuhl Seminar 21451 "Managing Industrial Control Systems Security Risks for Cyber Insurance" aimed to provide an interdisciplinary forum to analyze and discuss open questions and current topics of research in this area in order to gain in-depth insights into the security risks of ICSs and the quantification thereof.

Seminar November 7–12, 2021 – http://www.dagstuhl.de/21451

2012 ACM Subject Classification Security and privacy → Economics of security and privacy; Social and professional topics → Information system economics

Keywords and phrases industrial control systems, security, cyber insurance, cyber risk quantification, production systems engineering, risk engineering, SCADA, Industry 4.0

Digital Object Identifier 10.4230/DagRep.11.10.36

Edited in cooperation with Eckhart, Matthias

Managing Industrial Control Systems Security Risks for Cyber Insurance, *Dagstuhl Reports*, Vol. 11, Issue 10, pp. 36–56
Editors: Simon Dejung, Mingyan Liu, Arndt Lüder, and Edgar Weippl
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Matthias Eckhart (SBA Research – Wien, AT, meckhart@sba-research.org)
Simon Dejung (SCOR – Zürich, CH, sdejung@scor.com)
Mingyan Liu (University of Michigan – Ann Arbor, US, mingyan@umich.edu)
Arndt Lüder (Otto-von-Guericke-Universität Magdeburg, DE, arndt.lueder@ovgu.de)
Edgar Weippl (University of Vienna & SBA Research – Wien, AT, edgar.weippl@univie.ac.at)

The security economics community has an ambivalent position on quantifying cyber risk: on the one hand, there is a long-standing interest in establishing cyber risk measurement, while on the other hand, relevant publications report contradictory results [7] or present models with insufficient evidence of validity [6]. As researchers remain cautious about the interpretation of modeling results, the insurance industry's need for gaining a quantitative understanding of cyber risk is more critical than ever. A serious concern for insurers is that a large-scale cyber-attack could result in significant claims arising from transferred security risks. Considering the adverse business implications of cyber-related high-impact, low-frequency events, it is necessary to estimate and control the potential exposure to such losses. Previous industry studies attempted to address this issue by proposing hypothetical worst-case scenarios involving power grids [8], ports [9], and other industrial applications [4]. Furthermore, several industry-led working groups conducted workshops to assess the plausibility, impact, and claim implications of potential ICS-related catastrophic loss events (cf., for instance, [10, 11, 5, 12]). However, it is still not fully understood how the peculiarities of ICSs, technological change in light of strategic initiatives (e.g., Industry 4.0 [3]), and the increasingly sophisticated nature of cyber-physical attacks influence the loss frequency and loss severity [1]. Moreover, a holistic consideration of cyber-physical risk featuring the complete ICS lifecycle calls for an interdisciplinary research approach [2].

Thus, the aim of this Dagstuhl seminar was to bring together different communities to foster research activities that advance the understanding of cyber risks pertaining to ICSs and associated insurance aspects. The concepts developed as part of this seminar are a result of interdisciplinary work conducted by academics and industry professionals, both junior and senior, from the fields of

(i) computer science,
(ii) automation engineering,
(iii) actuarial science, and
(iv) economics.

To address the issues outlined above, the purpose of the seminar was to make the first steps toward a probabilistic cyber catastrophe model that is tailored to the ICS domain. In particular, we planned to achieve the basis of an economic loss model that builds upon worst-case scenarios in which globally and simultaneously many industrial processes in critical infrastructure sectors (e.g., power, petrochemical, transport, logistics) are affected by cyber-attacks. Figure 1 visualizes possible components of such a model, which were discussed and challenged during the seminar. In the first phase, the scenario is formulated, fundamental assumptions are specified, and the theoretical basis of the statistical model is formed. After that, the model is calibrated with data obtained from various sources, such as loss databases or subject-matter experts. Finally, the model and its underlying assumptions are validated.

To set the frame for the seminar, the organizers defined four topics that were covered in plenary sessions and breakout sessions. In each plenary session, lightning talks were held that motivated the collaborative work in the breakout sessions. The working groups studied

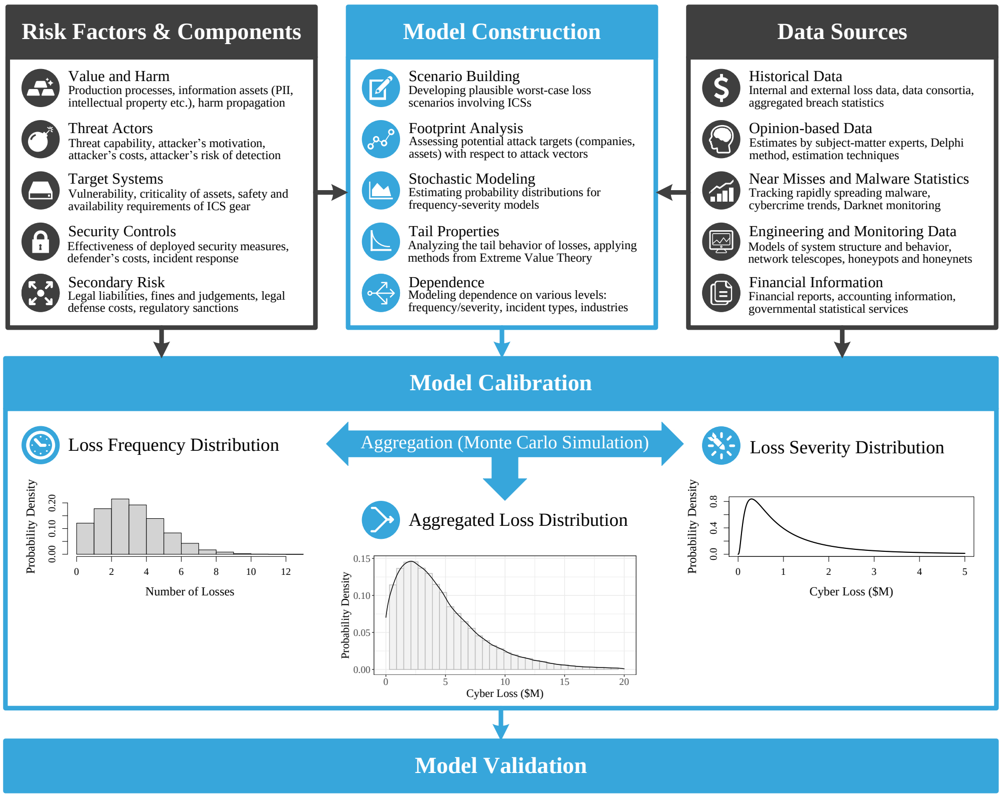

Risk Factors & Components
Value and Harm
Production processes, information assets (PII, intellectual property etc.), harm propagation
Threat Actors
Threat capability, attacker's motivation, attacker's costs, attacker's risk of detection
Target Systems
Vulnerability, criticality of assets, safety and availability requirements of ICS gear
Security Controls
Effectiveness of deployed security measures, defender's costs, incident response
Secondary Risk
Legal liabilities, fines and judgements, legal defense costs, regulatory sanctions
Model Construction
Scenario Building
Developing plausible worst-case loss scenarios involving ICSs
Footprint Analysis
Assessing potential attack targets (companies, assets) with respect to attack vectors
Stochastic Modeling
Estimating probability distributions for frequency-severity models
Tail Properties
Analyzing the tail behavior of losses, applying methods from Extreme Value Theory
Dependence
Modeling dependence on various levels: frequency/severity, incident types, industries
Data Sources
Historical Data
Internal and external loss data, data consortia, aggregated breach statistics
Opinion-based Data
Estimates by subject-matter experts, Delphi method, estimation techniques
Near Misses and Malware Statistics
Tracking rapidly spreading malware, cybercrime trends, Darknet monitoring
Engineering and Monitoring Data
Models of system structure and behavior, network telescopes, honeypots and honeynets
Financial Information
Financial reports, accounting information, governmental statistical services
Model Calibration
Loss Frequency Distribution
Probability Density
0.00 0.10 0.20
0 2 4 6 8 10 12
Number of Losses
Aggregation (Monte Carlo Simulation)
Aggregated Loss Distribution
Probability Density
0.00 0.05 0.10 0.15
0 5 10 15 20
Cyber Loss ($M)
Loss Severity Distribution
Probability Density
0.0 0.4 0.8
0 1 2 3 4 5
Cyber Loss ($M)
Model Validation

Figure 1 Potential components of a probabilistic cyber catastrophe model for the ICS domain (adapted from materials provided by SCOR SE).

the same overarching topic of the breakout session (yet each with a different focus) in order to strengthen interdisciplinary exchange.

Overall, the following topics and motivating research questions were addressed:

1. *ICS Threat Landscape*: How have cyber attacks against ICSs evolved and what should we expect in terms of attack sophistication, persistence, and impact in the future?
2. *Cyber-Physical Risk Quantification*: How can we quantitatively model economic losses caused by ICS-focused cyber risks (i.e., probabilistic cyber catastrophe model)?
3. *Insurance*: What are the opportunities and limitations of transferring ICS-focused cyber risks to insurers?
4. *Management of Security Risks*: Which hard (e.g., technological security measures) and soft (e.g., information sharing, regulations, funding) factors increase or reduce the attack likelihood and severity?

The seminar started with a welcome session to bridge the disciplinary gap. In this session, the organizers presented the seminar program, explained key terms, and discussed core concepts to familiarize attendees with the terminologies used by different communities. Over the following days, several participants gave lightning talks that focused on the following topics:

- cyber-physical systems, security-relevant aspects within their lifecycle, and procurement considerations,

- current ICS security challenges with an emphasis on technological trends (e.g., Industry 4.0, smart manufacturing, Industrial Internet of Things),
- cyber-physical risk assessments, where special attention was given to analysis and quantification methods,
- various aspects of (cyber) insurance (e.g., cyber cat modeling, underwriting, economic problems, regulations), and
- security economics, featuring studies on cybercrime analysis and vulnerability forecasting.

The lightning talks gave participants the opportunity to present new perspectives and challenges, which led to lively discussions that shaped the group sessions. Unfortunately, the restrictions caused by the SARS-CoV-2 pandemic made it not possible to conduct the estimation exercises with all participants, which would have been required for achieving the cyber cat model. However, conducting the seminar in a hybrid format still enabled the participants both on-site and remote to work together on challenging open questions, contribute to group discussions, and forge new research collaborations.

The organizers thank all participants for their valuable contribution. Furthermore, this seminar would not have been possible without the great technical support provided by the Schloss Dagstuhl staff and the considerable effort made by the video conferencing assistants Sejdefa Ibisevic, Markus Maier, and Sara Tajik.

References

1 Matthias Eckhart, Bernhard Brenner, Andreas Ekelhart, and Edgar Weippl. Quantitative security risk assessment for industrial control systems: Research opportunities and challenges. *Journal of Internet Services and Information Security (JISIS)*, 9(3):52–73, August 2019.

2 Gregory Falco, Martin Eling, Danielle Jablanski, Matthias Weber, Virginia Miller, Lawrence A. Gordon, Shaun Shuxun Wang, Joan Schmit, Russell Thomas, Mauro Elvedi, Thomas Maillart, Emy Donavan, Simon Dejung, Eric Durand, Franklin Nutter, Uzi Scheffer, Gil Arazi, Gilbert Ohana, and Herbert Lin. Cyber risk research impeded by disciplinary barriers. *Science*, 366(6469):1066–1069, 2019.

3 Henning Kagermann, Johannes Helbig, Ariane Hellinger, and Wolfgang Wahlster. Recommendations for implementing the strategic initiative INDUSTRIE 4.0 – securing the future of german manufacturing industry. Final report of the Industrie 4.0 working group, acatech – National Academy of Science and Engineering, München, April 2013.

4 Lloyd's of London, Guy Carpenter, and CyberCube Analytics. Cyber risk: The emerging cyber threat to industrial control systems. Technical report, Lloyd's of London, Guy Carpenter, and CyberCube Analytics, February 2021.

5 Lobo, Francis. Upstream oil & gas cyber risk: Insurance technical review. Technical report, Joint Rig Committee, May 2018. A Joint Rig Committee Report.

6 Vilhelm Verendel. Quantified security is a weak hypothesis: A critical survey of results and assumptions. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop*, NSPW '09, pages 37–50, New York, NY, USA, 2009. ACM.

7 Daniel W. Woods and Rainer Böhme. SoK: Quantifying cyber risk. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 211–228, May 2021.

8 Lloyd's of London and Cambridge Centre for Risk Studies. Business blackout: The insurance implications of a cyber attack on the us power grid. Technical report, Lloyd's of London and Cambridge Centre for Risk Studies, July 2015.

9 Lloyd's of London, Cambridge Centre for Risk Studies, and Nanyang Technological University. Shen attack: Cyber risk in asia pacific ports. Technical report, Lloyd's of London, Cambridge Centre for Risk Studies, and Nanyang Technological University, 2019.

10 Dejung, Simon. Economic impact of cyber accumulation scenarios. Technical report, SCOR Global P&C, 2017.

11 Dejung, Simon. Newsletter – risk assessment for ICS/SCADA security in industrial property, engineering, power, oil & gas. Technical report, SCOR Global P&C, March 2018. A joint workshop in March 2018 by LMA, IMIA & OPERA at SCOR (Zurich).

12 IMIA Working Group. Cyber risks: Engineering insurers perspective. Technical Report 98 (16), September 2016. IMIA Annual Conference 2016 – Doha, Qatar.

2 Table of Contents

Executive Summary

Matthias Eckhart, Simon Dejung, Mingyan Liu, Arndt Lüder, Edgar Weippl 37

Overview of Talks

Malware Economics for ICS Risk
Luca Allodi 42

Twin-based Continuous Countermeasure Deployment
Fabrizio Baiardi 42

Quantifying Cyber Risk
Rainer Böhme 44

Are ICS Scenarios Scalable? Ingredients of an Economic Loss Model
Simon Dejung 44

Building Blocks of a Cyber Cat Model
Téodore Iazykoff 45

Towards Joined Cyber Insurance Exercises
Helge Janicke 45

Race-to-the-Bottom: Evolution of Threat Landscape to Industrial Control Systems
Marina Krotofil 46

Vulnerability Forecasting: Theory and Practice
Éireann Leverett 47

Risk Dependency and Cyber Insurance
Mingyan Liu 47

Exploiting production system engineering data to evaluate attacks
Arndt Lüder 48

ICSs in the context of Industry 4.0 - A life cycle consideration
Arndt Lüder 48

Dependency Model of a SCADA System for Goal-Oriented Risk Assessment
Simin Nadjm-Tehrani 49

Cyber Insurance: ICS vs ITS
Galina Schwartz 50

Counterfactual Analysis of Cyber-Physical Risk
Gordon Woo 50

How Insurance Shapes Incident Response
Daniel Woods 51

Working Groups

Analyzing the ICS Threat Landscape
Matthias Eckhart 51

Developing Extreme Cyber-Physical Loss Scenarios
Matthias Eckhart 53

Participants 56

Remote Participants 56

3 Overview of Talks

3.1 Malware Economics for ICS Risk

Luca Allodi (TU Eindhoven, NL)

Joint work of Luca Allodi, Michele Campobasso, Fabio Massacci
Main reference Michele Campobasso, Luca Allodi: "Impersonation-as-a-Service: Characterizing the Emerging Criminal Infrastructure for User Impersonation at Scale", in Proc. of the CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020, pp. 1665–1680, ACM, 2020.
URL http://dx.doi.org/10.1145/3372297.3417892

I bring into focus the relevance of attacker and cybercrime capabilities in the ICS threat scenario. I propose a two-dimensional space to map past ICS incidents over the IT/OT/"logic" knowledge and the process knowledge the attacker needs to engineer and deliver a successful attack. I identify an area of attacks where common cybercriminals ("Dimitry") can operate (and are operating), supported by the underground markets. I discuss implications in terms of attack surface stability, and shared attacker capabilities that together characterize a "baseline risk" for ICS. As such, I argue this baseline risk is systemic to all ICS scenarios, can be quantified, and should be used to identify the gap between what any attacker can achieve, and what sophisticated, resourceful, nation-state level attackers can: how far away from "Dimitry" an attacker has to move to achieve what type of impact?

3.2 Twin-based Continuous Countermeasure Deployment

Fabrizio Baiardi (University of Pisa, IT)

Joint work of Fabrizio Baiardi, Federico Tonelli
Main reference Fabrizio Baiardi, Federico Tonelli: "Twin Based Continuous ICT Risk Management" in Proc. of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference, Venice, Italy, November 1-5, 2020, pp. 2012–2019, Singapore: Research Publishing, 2020.
URL https://www.rpsonline.com.sg/proceedings/esrel2020/html/5003.xml

Digital twins are virtual replicas that simulate the behavior of physical devices before they are built and to support their maintenance. We extend this technology to cybersecurity and integrate it with adversary emulation to define a policy to remediate the vulnerabilities of an ICT before threat actors can exploit them. Distinct twins model, respectively, the infrastructure and threat actors. A twin describes the infrastructure modules, their vulnerabilities, and the elementary attacks actors can implement. The twin of a threat actor describes its attack surface, its goals, how it selects attacks, and it handles attack failures. The Haruspex software platform builds the infrastructure twin and those of the threat actors, and it automates the emulation. In this way, it can discover the attack paths the actor implements without disturbing the infrastructure. In each path, the actor composes elementary attacks to reach its goal. Multiple emulations can discover all the actor paths by covering stochastic factors such as attack success or failure. The knowledge of the paths enables the remediation policy to minimize the countermeasures to deploy. A twin-based approach supports a continuous remediation process to handle changes in the infrastructure, new vulnerabilities, and new threat actors because the platform can update the twins and run adversary emulations. If new attack paths exist, the platform applies the remediation policy. Experimental data confirm the effectiveness of this approach.

References

1 Andy Applebaum, Doug Miller, Blake Strom, Henry Foster, and Cody Thomas. Analysis of automated adversary emulation techniques. In *Proceedings of the Summer Simulation Multi-Conference*, SummerSim '17, San Diego, CA, USA, 2017. Society for Computer Simulation International.

2 Andy Applebaum, Doug Miller, Blake Strom, Chris Korban, and Ross Wolf. Intelligent, automated red team emulation. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, ACSAC '16, pages 363–373, New York, NY, USA, 2016. Association for Computing Machinery.

3 Fabrizio Baiardi. Avoiding the weaknesses of a penetration test. *Computer Fraud & Security*, 2019(4):11–15, 2019.

4 Fabrizio Baiardi and Daniele Sgandurra. Assessing ICT risk through a Monte Carlo method. *Environment Systems and Decisions*, 33(4):486–499, Dec 2013.

5 Fabrizio Baiardi and Federico Tonelli. Twin based continuous ICT risk management. In Piero Baraldi, Francesco Di Maio, and Enrico Zio, editors, *Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*, pages 2012–2019, Singapore, 2020. Research Publishing.

6 Fabrizio Baiardi, Federico Tonelli, and Alessandro Bertolini. CyVar: Extending Var-At-Risk to ICT. In Fredrik Seehusen, Michael Felderer, Jürgen Großmann, and Marc-Florian Wendland, editors, *Risk Assessment and Risk-Driven Testing*, pages 49–62, Cham, 2015. Springer International Publishing.

7 Sean Barnum. Standardizing cyber threat intelligence information with the Structured Threat Information eXpression (STIX$^{\text{TM}}$). techreport, The MITRE Corporation, February 2014.

8 Sarah Brown, Joep Gommers, and Oscar Serrano. From cyber security information sharing to threat management. In *Proceedings of the 2nd ACM Workshop on Information Sharing and Collaborative Security*, WISCS '15, pages 43–49, New York, NY, USA, 2015. Association for Computing Machinery.

9 Matthias Eckhart and Andreas Ekelhart. *Digital Twins for Cyber-Physical Systems Security: State of the Art and Outlook*, chapter 14, pages 383–412. Springer International Publishing, Cham, 2019.

10 Bob Martin. Common Vulnerabilities Enumeration (CVE), Common Weakness Enumeration (CWE), and Common Quality Enumeration (CQE): Attempting to systematically catalog the safety and security challenges for modern, networked, software-intensive systems. *Ada Lett.*, 38(2):9–42, December 2019.

11 Peter Mell, Karen Scarfone, and Sasha Romanosky. The Common Vulnerability Scoring System (CVSS) and its applicability to federal agency systems. *NIST Interagency Report*, 7435, August 2007.

12 Stephen Moskal, Shanchieh Jay Yang, and Michael E. Kuhl. Cyber threat assessment via attack scenario simulation using an integrated adversary and network modeling approach. *The Journal of Defense Modeling and Simulation*, 15(1):13–29, 2018.

13 Blake E. Strom, Andy Applebaum, Doug P. Miller, Kathryn C. Nickels, Adam G. Pennington, and Cody B. Thomas. MITRE ATT&CK®: Design and philosophy. *MITRE Product*, (10AOH08A-JC), March 2020.

14 Fei Tao, He Zhang, Ang Liu, and A. Y. C. Nee. Digital twin in industry: State-of-the-art. *IEEE Transactions on Industrial Informatics*, 15(4):2405–2415, April 2019.

3.3 Quantifying Cyber Risk

Rainer Böhme (Universität Innsbruck, AT)

Joint work of Rainer Böhme, Daniel Woods
Main reference Daniel W. Woods, Rainer Böhme: "SoK: Quantifying Cyber Risk", in Proc. of the 42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021, pp. 211–228, IEEE, 2021.
URL http://dx.doi.org/10.1109/SP40001.2021.00053

This talk introduces a causal model inspired by structural equation modeling that explains cyber risk outcomes in terms of latent factors measured using reflexive indicators. First, we use the model to classify empirical cyber harm studies. We discover cyber harms are not exceptional in terms of typical or extreme losses. The increasing frequency of data breaches is contested and stock market reactions to cyber incidents are becoming less negative over time. Focusing on harms alone breeds fatalism; the causal model is most useful in evaluating the effectiveness of security interventions, which are surveyed in the second half of the talk.

3.4 Are ICS Scenarios Scalable? Ingredients of an Economic Loss Model

Simon Dejung (SCOR – Zürich, CH)

Main reference Simon Dejung: "Newsletter – risk assessment for ICS/SCADA security in industrial property, engineering, power, oil & gas," SCOR Global P&C, Tech. Rep., Mar. 2018, A joint workshop in March 2018 by LMA, IMIA & OPERA at SCOR (Zurich).
URL https://www.imia.com/wp-content/uploads/2021/07/Economic_impact_Cyber_loss_accumulation_scenarios_SVV.pdf

Are scenarios like Industroyer, Havex, Triton/Trisis, Stuxnet scalable? Cyber attacks are the new normal and are correlated with increasing digitalization and interconnectivity. If these attacks are single incidents, they are mainly affecting the attacked victim and companies and/or individuals being linked to these companies having suffered such hostile acts. What we observe more and more is scalability by automation and/or recycling with or without manual adaption of the used malware. Currently we see most incidents in the IT environment and even attacks on critical infrastructures like Colonial pipeline were triggered by office IT ransomware. What if ICS/SCADA/OT scenarios become scalable and e.g., Stuxnet derivatives are more widely used damaging not only non-physical assets, but also physical assets? Current cyber loss models focus on IT damages and its consequences, which are mainly non-physical. Economic loss models breaching the gap to the physical word, considering the probability of material damages and the likelihood of scalability are in its infancy. Interdependencies and interactions of risk factors like e.g., attackers' capabilities, resources, motivation, political and macro-economic environment are not yet sufficiently addressed. Economic loss models properly addressing these factors will serve decision makers on various levels.

References

1 Dejung, Simon. Economic impact of cyber accumulation scenarios. Technical report, SCOR Global P&C, 2017.

2 Dejung, Simon. Newsletter – risk assessment for ICS/SCADA security in industrial property, engineering, power, oil & gas. Technical report, SCOR Global P&C, March 2018. A joint workshop in March 2018 by LMA, IMIA & OPERA at SCOR (Zurich).

3 Ben Hobby and Matthew Hogg. Cyber insurance & business interruption. Technical report, The International Underwriting Association of London Limited and RGL Forensics, July 2018. A report from the IUA's Cyber Underwriting Group in association with RGL Forensics.

4 Matthew Honea, Yoshifumi Yamamoto, Jonathan Laux, Craig Guiliano, and Megan Hart. Silent cyber scenario: Opening the flood gates. Technical report, Aon and Guidewire – Cyence Risk Analytics, October 2018.

5 IMIA Working Group. Cyber risks: Engineering insurers perspective. Technical Report 98 (16), September 2016. IMIA Annual Conference 2016 – Doha, Qatar.

6 Lloyd's of London and Cambridge Centre for Risk Studies. Business blackout: The insurance implications of a cyber attack on the us power grid. Technical report, Lloyd's of London and Cambridge Centre for Risk Studies, July 2015.

7 Lloyd's of London, Cambridge Centre for Risk Studies, and Nanyang Technological University. Shen attack: Cyber risk in asia pacific ports. Technical report, Lloyd's of London, Cambridge Centre for Risk Studies, and Nanyang Technological University, 2019.

3.5 Building Blocks of a Cyber Cat Model

Téodore Iazykoff (SCOR – Paris, FR)

This talk explains basic concepts of cat modeling to build cyber models. A step by step guide provides key principles to enable participants to use a scenario-based approach. Both deterministic and stochastic approaches are compared, and detailed examples for each methodology are presented using a Cloud outage scenario. Participants were given the opportunity to discuss similarities and differences with natural catastrophe models.

3.6 Towards Joined Cyber Insurance Exercises

Helge Janicke (Cyber Security CRS – Joondalup, AU)

Joint work of Helge Janicke, Richard Smith, Allan Cook, Leandros Maglaras, Bil Hallaq
Main reference Richard Smith, Helge Janicke, Ying He, Fenia Ferra, Adham Albakri: "The Agile Incident Response for Industrial Control Systems (AIR4ICS) framework", Comput. Secur., Vol. 109, p. 102398, 2021.
URL http://dx.doi.org/10.1016/j.cose.2021.102398

Insurers need to understand the residual risk and potential consequences of a cyber attack on the businesses they insure. Evaluating documentation of compliance, controls and checklists only goes so far and may not provide a proper picture of an organization's cyber security posture. Many large organizations undertake table-top exercises and have defined incident response plans, but the proof is often in the management of an incident and in the proficiency of the staff responding, that is ultimately responsible for mitigating the consequences. There is an opportunity for cyber insurers to provide training simulations (similar to those mandated for nuclear facilities) as an additional service line to their business, helping inform risk assessments not solely based on controls, policies and plans but also to take into account and organization's cyber capability, capacity and proficiency. The research challenges here are significant, as it is unclear how simulations are co-developed and how capability, capacity and proficiency can be effectively assessed. There are also technical challenges that require more realistic scenarios for ICS. Commercial ICS cyberranges are emerging, but may be underdeveloped, expensive to operate and are not easily adapted to changing technologies. This presentation will incorporate direct experiences in running ICS-specific incident response training from the UK's NCSC funded Agile Incident Response for ICS (AIR4ICS) project. It will also set out some of the challenges to motive the following breakout session.

References

1 Allan Cook, Helge Janicke, Richard Smith, and Leandros Maglaras. The industrial control system cyber defence triage process. *Computers & Security*, 70:467–481, 2017.

2 Bil Hallaq, Andrew Nicholson, Richard Smith, Leandros Maglaras, Allan Cook, Helge Janicke, and Kevin Jones. A novel hybrid cyber range for security exercises on cyber-physical systems. *International Journal of Smart Security Technologies*, 8(1):16–34, January 2021.

3 Richard Smith, Helge Janicke, Ying He, Fenia Ferra, and Adham Albakri. The agile incident response for industrial control systems (AIR4ICS) framework. *Computers & Security*, 109:102398, 2021.

3.7 Race-to-the-Bottom: Evolution of Threat Landscape to Industrial Control Systems

Marina Krotofil (Maersk – Aarhus, DK)

Industrial Control Systems (ICS) threat landscape has changed dramatically over the past years. New threats have emerged to challenge the shock created by Stuxnet. This talk will present the evolution of the ICS exploits and tactics to picture ongoing "race-to-the-bottom" trend between ICS threat actors and defenders. This trend refers to the tendency of the attackers to move their exploits one layer down as soon as security controls are introduced at some layer of the computer or network architecture. While OT asset owners begin to harden operator consoles and embrace ICS network monitoring solutions, the attackers are already moving their exploits into the controllers at the regulatory layer of network architecture. The reason for this strategy is the lack of exploit mitigation and detection capabilities in most of the embedded systems components deployed within ICS and a lack of tools to support compromise assessment and forensic analysis of these systems. This talk will outline the ICS exploitation trends and briefly discuss their implication on defensibility and evaluating risks to ICS environments.

3.8 Vulnerability Forecasting: Theory and Practice

Éireann Leverett (University of Cambridge, GB)

Joint work of Éireann Leverett, Matilda Rhode, Adam Wedgbury
Main reference Éireann Leverett, Matilda Rhode, Adam Wedgbury: "Vulnerability Forecasting: In theory and practice", CoRR, Vol. abs/2012.03814, 2020.
URL https://arxiv.org/abs/2012.03814

It is possible to forecast the volume of CVEs released within a time frame with a given prediction interval. For example, the number of CVEs published between now and a year from now can be forecast within 8% of the actual value. Different predictive algorithms perform well at different lookahead values other than 365 days, such as monthly, quarterly, and half year. It is also possible to estimate the proportions of that total volume belonging to specific vendors, software, CVSS scores, or vulnerability types. Some vendors and products can be predicted with accuracy, others with too much uncertainty to be practically useful. This paper documents which vendors are amenable to being forecasted. Strategic patch management should become much easier with these tools, and further uncertainty reductions can be built from the methodologies in this paper.

3.9 Risk Dependency and Cyber Insurance

Mingyan Liu (University of Michigan – Ann Arbor, US)

Joint work of Mingyan Liu, Mahdi Khalili, Parinaz Naghizadeh, Armin Sarabi, Tongxin Yin

Cyber risks are notoriously interdependent at a firm level: an insured's risk is a function of not only its own conditions, but also that of its vendors and suppliers. Insurers generally try to avoid this type of risk dependency. Within this context, I will discuss our research over the past 7–8 years on shifting the focus from the conventional view of using insurance as primarily a risk management mechanism to one of risk control and reduction by looking for ways to re-align the incentives of parties involved in an insurance contract and exploiting the unique properties of cyber risk. In particular, using a commonly practiced rate-schedule based policy framework, I will analyze and compare three different policy portfolios and make a case for why insurers should actually consider embracing risk dependency in their underwriting. I will also share our most recent work on underwriting ransomware insurance. In doing so, I will draw a number of parallels between ransomware attacks and the centuries-old crime, kidnapping for ransom, discuss how the latter has been an insurable risk, and highlight lessons we can learn in conceptualizing an effective framework around the design and governance of ransomware insurance.

References

1 Mohammad Mahdi Khalili, Mingyan Liu, and Sasha Romanosky. Embracing and controlling risk dependency in cyber-insurance policy underwriting. *Journal of Cybersecurity*, 5(1), October 2019.

2 Mohammad Mahdi Khalili, Parinaz Naghizadeh, and Mingyan Liu. Designing cyber insurance policies: The role of pre-screening and security interdependence. *IEEE Transactions on Information Forensics and Security*, 13(9):2226–2239, September 2018.

3 Mingyan Liu. *Embracing Risk: Cyber Insurance as an Incentive Mechanism for Cybersecurity*, volume 2. Morgan & Claypool Publishers LLC, June 2021.

4 Yang Liu, Armin Sarabi, Jing Zhang, Parinaz Naghizadeh, Manish Karir, Michael Bailey, and Mingyan Liu. Cloudy with a chance of breach: Forecasting cyber security incidents. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 1009–1024, Washington, D.C., August 2015. USENIX Association.

5 Armin Sarabi, Parinaz Naghizadeh, Yang Liu, and Mingyan Liu. Risky business: Fine-grained data breach prediction using business profiles. *Journal of Cybersecurity*, 2(1):15–28, December 2016.

3.10 Exploiting production system engineering data to evaluate attacks

Arndt Lüder (Otto-von-Guericke-Universität Magdeburg, DE)

Main reference Anna-Kristin Behnert, Felix Rinker, Arndt Lüder, Stefan Biffl: "Migrating Engineering Tools Towards an AutomationML-Based Engineering Pipeline", in Proc. of the 19th IEEE International Conference on Industrial Informatics, INDIN 2021, Palma de Mallorca, Spain, July 21-23, 2021, pp. 1–7, IEEE, 2021.
URL http://dx.doi.org/10.1109/INDIN45523.2021.9557517
Main reference Stefan Biffl, Arndt Lüder, Kristof Meixner, Felix Rinker, Matthias Eckhart, Dietmar Winkler: "Multi-view-Model Risk Assessment in Cyber-Physical Production Systems Engineering", in Proc. of the 9th International Conference on Model-Driven Engineering and Software Development, MODELSWARD 2021, Online Streaming, February 8-10, 2021, pp. 163–170, SCITEPRESS, 2021.
URL http://dx.doi.org/10.5220/0010224801630170
Main reference Arndt Lüder, Stefan Biffl, Felix Rinker, Anna-Kristin Behnert: "32 Engineering Data Logistics based on AML", pp. 579–602, De Gruyter Oldenbourg, 2021.
URL http://dx.doi.org/doi:10.1515/9783110745979-034
Main reference Kristof Meixner, Arndt Lüder, J. Herzog, Dietmar Winkler, Stefan Biffl: "Patterns for Reuse in Production Systems Engineering" in Proceedings of the 33th International Conference on Software Engineering and Knowledge Engineering (SEKE)", Pittsburgh, USA, July 1-10, 2021, IEEE, 2021.
URL http://dx.doi.org/10.18293/SEKE2021-150

Engineering data are a vital source of information applicable to evaluate security concerns within production systems.

To make them applicable at first an engineering data logistics is required resulting in an aggregated set of engineering information related to all assets within a production system and at second appropriate analysis methodologies for the collected engineering data treasure are required.

Relevant research question in this directions are the following: What are the right data sources to represent the required knowledge for a security evaluation? What is an appropriate data logistics? How to analyze the collected data?

Attention has to be put on the necessary effort and the potentials of reusing engineering data that can reduce effort significantly.

3.11 ICSs in the context of Industry 4.0 - A life cycle consideration

Arndt Lüder (Otto-von-Guericke-Universität Magdeburg, DE)

Production systems can be considered as a combination of resources intended to be used to execute production processes resulting in products exposint "the right" product properties for beeing valuable for customers. Hence a PPR based consideration of products can help to understand potential impacts on production systems resulting from security attacks.

Such attacks can disable, hamper or disturb production systems. Disable mean permanently prevent a functionality from being applied thus stoping production. Hamper means temporarily prevent a functionality from being applied. Finally disturb means changing function results without notice.

While currently disable and hamper are mainly considered within research disturb is less considered. But such attachs are much more effectfull to companies as they are more difficult to identify than the others by conventional production system quality management.

This leads to the following research questions:

- How to collect and classify assets and attacks to production systems?
- What are motivation, aim, effect, and required knowledge of different attacks?
- Are there similarities and differences?
- Will we find "same" assets?
- Will attacks at engineering time enable / enforce attacks at runtime?
- Which knowledge is required?

3.12 Dependency Model of a SCADA System for Goal-Oriented Risk Assessment

Simin Nadjm-Tehrani (Linköping University, SE)

In this talk I present some reflections from the past two days about how to do ICS risk analysis and relate to work done in my group and other colleagues. The insurance companies and policy makers want to know how to assess economic risks in relation to cyber incidents in ICS. They seem to want an "easy" approach that works without taking months/years to perform. At the other end of the spectrum we have the cybersecurity researchers who study risk in diverse ways and create multiple tools and methods that need feeding with a lot of information. Will these meet the requirements? The meeting has discussed relatively little the contrasts in stakeholders perspectives [3, 2]. I open the talk by some approaches for cyber risk analysis at different levels of granularity and embryos of some (digital) tools to support the analysis activities. I conclude with some recent work [1] that is based on a goal-directed approach to analysis risk in a SCADA context that may be transferable to other areas.

References

1 Yulia Cherdantseva, Pete Burnap, Simin Nadjm-Tehrani, and Kevin Jones. A configurable dependency model of a SCADA system for goal-oriented risk assessment (under submission). 2022.

2 Maria Vasilevskaya and Simin Nadjm-Tehrani. Quantifying risks to data assets using formal metrics in embedded system design. In Floor Koornneef and Coen van Gulijk, editors, *Computer Safety, Reliability, and Security*, pages 347–361, Cham, 2015. Springer International Publishing.

3 Maria Vasilevskaya and Simin Nadjm-Tehrani. Model-based security risk analysis for networked embedded systems. In Christos G. Panayiotou, Georgios Ellinas, Elias Kyriakides, and Marios M. Polycarpou, editors, *Critical Information Infrastructures Security*, pages 381–386, Cham, 2016. Springer International Publishing.

3.13 Cyber Insurance: ICS vs ITS

Galina Schwartz (Cyber Blocks Inc. – Berkeley, US)

Joint work of Galina Schwartz, Carlos Barreto, Alvaro A. Cardenas
Main reference Carlos Barreto, Galina Schwartz, Alvaro A. Cardenas: "Cyber-Risk: Cyber-Physical Systems Versus Information Technology Systems", pp. 319–345, Springer International Publishing, 2021.
URL http://dx.doi.org/10.1007/978-3-030-65048-3_14

This talk introduces a taxonomy of cyber risks for ICS (Industrial control systems) and ITS (information technology systems). Both, ICS and ITS are data rich environments, yet they are plagued by extreme information deficiencies, combined with a high level of information asymmetries. We outline the factors complicating the advancement of ICS cyber-insurance ecosystem, incl.: extreme information scarcity; risk assessment difficulties, exacerbated by the growing complexity of ICS and the intricacies of risk prorogation. We conclude that without improving security relevant information, the cyber-insurance market for ICS may stall. Market advancement requires overcoming data scarcity and lack of standardization. We call for further research in CPS risk management, and specifically design and evaluation of novel technical tools and policies / regulations improving incentives of the ICS decision-makers to collect and share security related data. This talk is loosely based on [2, 1].

References

1 Carlos Barreto, Galina Schwartz, and Alvaro A. Cardenas. *Cyber-Insurance*, pages 347–375. Springer International Publishing, Cham, 2021.

2 Carlos Barreto, Galina Schwartz, and Alvaro A. Cardenas. *Cyber-Risk: Cyber-Physical Systems Versus Information Technology Systems*, pages 319–345. Springer International Publishing, Cham, 2021.

3.14 Counterfactual Analysis of Cyber-Physical Risk

Gordon Woo (Risk Management Solutions – London, GB)

Whenever notable adverse events occur, effort is naturally focused on risk mitigation and disaster prevention. According to psychologists, the great majority of thoughts about the past focus on how things might have been better. These are upward counterfactuals. However, important lessons may also be learned from thoughts about how things might have been otherwise – in particular if they had been worse. These are downward counterfactuals [1]. Most cyber-physical attacks turn out to be near-misses; examples of what might happen, but has not yet happened.

In 2013, an Iranian hacker, working on behalf of the Iranian government, repeatedly obtained unauthorized access to the SCADA systems of the Bowman Dam, in Rye, N.Y.. Although SCADA system access would normally have permitted remote operation and manipulation of the Bowman Dam's sluice gate, by a stroke of good fortune, this had been manually disconnected for maintenance at the time of his intrusion. This was a near-miss. Counterfactually, manipulation of the sluice gate might have led to flooding.

The historical record of cyber attacks is brief, but there is a large database of cyber attacks and their loss impacts. This historical database can be supplemented by a downward counterfactual database, which is currently under development.

References

1 Gordon Woo. Downward counterfactual search for extreme events. *Frontiers in Earth Science*, 7, 2019.

3.15 How Insurance Shapes Incident Response

Daniel Woods (Universität Innsbruck, AT)

Joint work of Daniel Woods, Rainer Böhme
Main reference Daniel W Woods, Rainer Böhme: "How Cyber Insurance Shapes Incident Response: A Mixed Methods Study", 2021.
URL https://weis2021.econinfosec.org/wp-content/uploads/sites/9/2021/06/weis21-woods.pdf

Cyber insurance policies commonly indemnify the cost of incident response services. This creates a multi-layered economic problem in that the policyholder hiring external firms incurs transaction costs and the insurer paying the bill creates a principal-agent problem. We adopted a multistage research design to understand how insurers address the problem. The talk explains how insurers have created a private ordering by controlling which firms are selected, negotiating prices ahead of time, and punishing low service quality by withholding future work. A minority of firms win the majority of work, thereby building trust through repeated interactions.

4 Working Groups

4.1 Analyzing the ICS Threat Landscape

Matthias Eckhart (SBA Research – Wien, AT)

Joint work of All Participants of the Seminar

The objective of this breakout session was to clarify the assumptions regarding ICS-targeting cyber-attacks that underlie loss scenarios. As a first step, the participants determined attributes of threat actors that influence the plausibility of a scenario. The considered attributes are in line with typical attacker properties that are assessed during profiling activities as part of threat modeling, which are often described in security textbooks. Examples include intent, motivation, skills, and resources. Given the evident focus on high-impact events, the participants concentrated on the following attacker archetypes:

(i) state-sponsored actors who want to gain a strategic advantage for their country of origin (e.g., in terms of political, military, and economic power),
(ii) terrorists who aim for maximum visibility to promote their ideologies and attract sponsors, and
(iii) organized cybercrime groups that are financially motivated.

Several assumptions were also made concerning the interactions between attackers and defenders. Most importantly, the participants agreed on an attacker model that considers rational and opportunistic decisions. In other words, adversaries have a limited budget available to execute cyber-physical attacks over a certain time period during which they can also adapt the attack strategy, seeking to minimize their costs while maximizing the defender's losses. Further aspects to consider when defining the attacker model relate to the strategic and tactical dimension, including the preparation phase and required resources (e.g., testbeds, exploits), techniques to gain a foothold, coordination of attack campaigns, and approaches to minimize the risk of detection. In this context, the following questions were discussed by participants:

- What are the resources and skills needed to execute large-scale cyber-attacks against ICSs?
- How does the level of difficulty change along the entire spectrum of ICS-targeting cyber-attacks?
- Which factors drive the scalability of a cyber-physical attack? How scalable is the considered cyber-physical threat?
- How likely are highly coordinated cyber-attacks launched against ICSs?
- Which trends (e.g., IT/OT convergence, Industrial Internet of Things, increasing redundancy in the supply chain) influence risk factors?
- Would such an attacker model have a reasonable stability? Would it be sensible to incorporate it in a loss scenario, considering that the threat landscape and technological trends change so rapidly?
- Which IT security measures have to be adapted to the requirements and characteristics of OT environments?
- How will the regulatory landscape pertaining to ICS security emerge in the next few years?
- How and to what extent is the state of ICS security improved by the standardization and implementation of reference architectures?
- What would be an attacker model that insurance will cover?

Naturally, the search for answers to these questions is guided by the idea that past observations are to a certain extent indicative of the future. Thus, prior ICS-related security incidents and near misses were intensively discussed. A recurring question among participants was why comparatively few noteworthy loss events involving ICSs are known. If the current state of ICS security is as alarming as is often portrayed, why do we not see *more* large-scale ICS-focused attacks that inflict significant damages? Obviously, the majority of successful attacks are promptly handled by incident response teams to limit their impact and even those that caused physical damages often go unreported or are not recognized as security incidents. However, it still seems that the (admittedly sparse) empirical data on cyber-physical attacks are not in line with what we would expect in terms of loss severity. One factor that may explain this discrepancy is the diversity in the context of hardware, software, and industrial processes, limiting the scalability of cyber-physical attacks.

Besides the discussions on how the threat landscape has evolved over the years, the participants also engaged in thought experiments about future attack trends. The groups came up with a set of observations and reflections suggesting that the overall state of ICS security will most likely not improve. Instead, the participants expect that malware will find its way into new generations of ICSs (e.g., renewable energy), cascading risks rise due to the proliferation of cloud-based industrial applications and the Industrial Internet of Things, and supply chain attacks become more sophisticated. Further, they expect that the

ICS threat landscape will be heavily shaped by the global political developments. As for non-state-sponsored threat actors, the participants anticipate that ICSs will become more profitable targets when terrorists and criminals acquire the expertise to scale up their attacks.

4.2 Developing Extreme Cyber-Physical Loss Scenarios

Matthias Eckhart (SBA Research – Wien, AT)

Joint work of All Participants of the Seminar

Following the discussions on the ICS threat landscape, the subsequent breakout sessions focused on developing loss scenarios for the ICS-specific cyber cat model. Ultimately, the objective was to describe and estimate extreme, but plausible cyber-physical attack scenarios featuring cascading effects, which increase the risk that losses could accumulate to a level that exceeds the (re)insurer's capacity to absorb it. The lightning talks given by Simon Dejung and Téodore Iazykoff served as a valuable introduction to catastrophe modeling and stimulated a fruitful exchange of ideas among participants about which approach to take. Two strategies emerged from these activities:

(i) The *top-down* approach seeks to estimate the impact of cyber-physical attacks at the macroeconomic level. Initially, the industry landscape is systematically analyzed to find chokepoints that could harm the global economy in the event of a large-scale attack. In particular, dependencies between companies need to be assessed to identify potential fragile economic conditions, which requires a broad understanding of supply chains and the market structure. Then, the overall problem is decomposed to reduce the complexity of estimating economic losses. An example for such a sub-problem would be the market share of victims in GDP terms. Once the skeleton of the scenario has been constructed, a further drill down to the factors that ultimately drive the frequency and severity can be carried out. The main task at this stage is to determine plausible cyber-physical attacks launched against the considered victims and the consequences that could push PD/MD, BI, and LOP to a realistic maximum.

(ii) The *bottom-up* approach seeks to approximate aggregate losses by using firm-level estimates. Thus, the initial focus lies on the technical and operational aspects from an asset owner's perspective when designing attack scenarios. In this context, the results of business impact analyses and risk assessments are central to understanding the potential consequences of loss of control and loss of safety. Furthermore, a retrospective view upon (non-)cyber-related ICS incidents may be used to identify and discard unrealistic scenarios. It should also be noted that scenarios with more advanced cyber-physical attacks may require careful consideration of the involvement of other roles (i.e., victims) within the ICS lifecycle, such as product suppliers, systems integrators, and service providers. After establishing a solid basis for the analysis, it is necessary to identify victim candidates that would be similarly affected (e.g., due to similarities in their IT/OT architectures) and to assess how the effects of the considered loss event could propagate across sectors.

Depending on the group members' background and confidence to start estimating the loss frequency and loss severity based on macro- or micro-factors, a mixed approach may be beneficial. In this way, complementary views can be incorporated to ensure that assessments coming from both directions meet in the middle. The experience we gained from the breakout sessions was that the bottom-up approach more easily takes subject-matter experts on a journey further down the rabbit hole of cyber-physical attacks, especially if their expertise is predominantly technical.

One group proposed the notion of *proximity* as an indicator for correlated risk, which was well-received and deemed highly useful by other groups. Measures of proximity are relevant to both top-down and bottom-up approaches and take different forms:

- *Logical proximity:* Multiple independent systems that are prone to fail at once due to the use of the same type of components or architectures are considered to be in close logical proximity. Basic examples that may lead to closer logical proximity include shared libraries, (near) identical configurations, similar artifacts in model-driven development, or even common blueprints for engineered systems (e.g., Tesla's Gigafactory concept that heavily relies on similar plans and equipment to accelerate expansion).
- *Temporal proximity:* If this property is present, multiple systems are prone to fail in close succession due to coordinated attacks performed by adversaries to achieve a particular goal. For instance, close temporal proximity exists if a series of well-timed attacks target independent units, one after another, seeking to bring down critical infrastructure.
- *Causal proximity:* This measure reflects the susceptibility of systems caused by their strong dependencies to others. Basically, if an attack compromises an integral component of an infrastructure, the entire services built on top are affected as well. Typical examples that can inflict serious losses due to causal proximity are related to the supply chain (e.g., hardware trojans) and infrastructure (e.g., DNS service disruption that leads to a widespread outage of websites).

In this context, two central questions arise: First, how can we measure logical, temporal, and causal proximity? Second, what actions can be taken to mitigate these forms of proximity? While these questions remain important avenues for future research, the participants suggested that engineering data could enable proximity measurements (at least in greenfield projects) and that different techniques of software diversity may be applicable to the industrial domain. Since data for proximity measurements on a global scale is scarce, the participants attempted to gauge the level of technological heterogeneity in different industrial domains. From a purely anecdotal perspective, it has been suggested that there are few industrial processes relying on highly specialized equipment that can be sourced from just one or two suppliers.

When developing the extreme cyber-physical loss scenarios, the following key questions were considered by the participants:

- Which targets would be profitable for the considered attacker archetypes and which of them could incur significant losses?
- What are possible chokepoints?
- What kind of long-term damages may be incurred by victims of cyber-physical attacks?
- What mechanisms are typically in place that ensure a safe, ultimate shut-down of a plant in case the systems have been compromised and are out of operator control?
- How quickly can asset owners switch to manual operation in the event of an attack to recover the industrial processes?
- What would be an adequate balance of prevention and response measures?
- How will the supplier landscape emerge (e.g., mergers and acquisitions, standardization) and what would be the consequences in terms of correlated risk?
- How can the incident response capacity needed for a given cyber-physical cat event be quantified?
- What are the differences in cyber-physical risk perception from the asset owner's and insurer's perspective?

- How does the sophistication of security measures in ICSs change depending on the level of insurance cover?
- How do asset owners find a balance between risk mitigation and risk transfer? Can we observe regional or sectoral differences?
- Which requirements regarding the implementation of ICS security measures can be imposed by insurers?
- Which information on ICSs should systems integrators provide to support the insurability of cyber-physical risk?
- How can certifications of ICS components shape the insurance industry?

To kick-off scenario building, the organizers have asked the participants about their interest and expertise in answering the aforementioned questions. After establishing common ground, brainstorming sessions were conducted to identify and frame the scenarios, which involved the following domains:

(i) power transmission,
(ii) natural gas (pipelines),
(iii) rail transport, and
(iv) air traffic control.

The participants decided to prioritize loss severity over loss frequency as variables of the former were deemed more stable. Given the limited time available, the participants approached the estimation problem by decomposing it into parameters, rating the parameters per scenario relative to each other, and constructing arguments that support these standpoints. The following list provides on overview of the considered parameters.

- Infrastructure: logical proximity, geographical dispersion, and resilience (in the sense of fault tolerance)
- Adversary: required domain knowledge (with respect to the industrial processes operated by the victim), required knowledge of the victim's IT environment and organizational structure, required knowledge of the victim's OT environment and executed process, and required resources to perform the attack (e.g., testbeds, person hours)
- Impact: negative physical effects and financial losses

While the participants engaged in vivid discussions during parameter ranking that led to important insights into the underlying problems, it became apparent that a more domain-specific setting is needed. Many of the sketched worst-case scenarios portrayed disastrous consequences, but on second thought the impact seemed negligible. For instance, the impact of a gas pipeline outage can be buffered by the storage capacity of the network, allowing time for recovery. The breakout sessions showed that determining such factors requires specialized know-how and a greater focus on a specific scenario. Thus, we plan to intensify our efforts by mobilizing additional domain expertise and initiating follow-up projects. Nevertheless, conducting the scenario building exercises was a valuable experience and marks the first step toward an ICS-focused cyber cat model.

Participants

- Luca Allodi
TU Eindhoven, NL
- Gergely Biczók
Budapest University of Technology & Economics, HU
- Rainer Böhme
Universität Innsbruck, AT
- Carl Denis
Universität der Bundeswehr – München, DE
- Matthias Eckhart
SBA Research – Wien, AT
- Alexander Horch
HIMA – Brühl, DE
- Sejdefa Ibisevic
Universität Wien, AT
- Julia Kittel
FH Emden, DE
- Klaus Kursawe
GridSec – Geneva, CH
- Arndt Lüder
Otto-von-Guericke-Universität Magdeburg, DE
- Fabio Massacci
Vrije Universiteit Amsterdam, NL
- Thomas Steinhaus
Munich Re, DE
- Edgar Weippl
University of Vienna & SBA Research – Wien, AT
- Jens Wiesner
BSI – Bonn, DE
- Daniel Woods
Universität Innsbruck, AT

Remote Participants

- Fabrizio Baiardi
University of Pisa, IT
- Vivien Bilquez
Zurich Insurance Group, CH
- Achim D. Brucker
University of Exeter, GB
- Richard Clayton
University of Cambridge, GB
- Simon Dejung
SCOR – Zürich, CH
- Andreas Ekelhart
SBA Research – Wien, AT
- Barbara Fila
IRISA – Rennes, FR
- Peter Hacker
Distinction.Global – Uetikon am See, CH
- Teodore Iazikoff
SCOR – Paris, FR
- Helge Janicke
Cyber Security CRS – Joondalup, AU
- Ersin Kaplan
HDI Global SE – Hannover, DE
- Marina Krotofil
Maersk – Aarhus, DK
- Éireann Leverett
University of Cambridge, GB
- Mingyan Liu
University of Michigan – Ann Arbor, US
- Markus Maier
Universität Wien, AT
- Jürgen Musil
Netinsurer – Wien, AT
- Simin Nadjm-Tehrani
Linköping University, SE
- Ranjan Pal
University of Michigan – Ann Arbor, US
- Keyun Ruan
Empty Labs Ltd. – London, GB
- Galina Schwartz
Cyber Blocks Inc. – Berkeley, US
- Sara Tajik
SBA Research – Wien, AT
- Josephine Wolff
Tufts Universtity – Medford, US
- Gordon Woo
Risk Management Solutions – London, GB
- Quanyan Zhu
NYU – Brooklyn, US

Report from Dagstuhl Seminar 21452

Unambiguity in Automata Theory

Edited by

Thomas Colcombet[1], Karin Quaas[2], and Michał Skrzypczak[3]

1 **Université Paris Diderot**
2 **Universität Leipzig**
3 **University of Warsaw**

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 21452 "Unambiguity in Automata Theory". The aim of the seminar was to improve the understanding of the notion of unambiguity in automata theory, especially with respect to questions related to the expressive power, succinctness, and the tractability of unambiguous devices. The main motivation behind these studies is the hope that unambiguous machines can provide a golden balance between efficiency – sometimes not worse than for deterministic devices – and expressibility / succinctness, which often is similar to the general nondeterministic machines. These trade-offs become especially important in the models where the expressiveness or the decidability status of unambiguous machines is different from that of nondeterministic ones, as it is the case, e.g., for register automata.

Seminar November 07-12, 2021 – `https://www.dagstuhl.de/21452`
2012 ACM Subject Classification Theory of computation → Regular languages; Theory of computation → Automata over infinite objects; Theory of computation → Algebraic language theory; Theory of computation → Tree languages
Keywords and phrases Unambiguity in Automata Theory, Dagstuhl Seminar
Digital Object Identifier 10.4230/DagRep.11.10.57

1 Executive Summary

Thomas Colcombet
Karin Quaas
Michał Skrzypczak

The Dagstuhl Seminar 21452 "Unambiguity in Automata Theory" was a seminar of five days that took place from November 7th to 12th, 2021, organized by Thomas Colcombet, Karin Quaas, and Michał Skrzypczak. A general goal of the seminar was to bring together experts from different fields of automata theory, to stimulate an exchange of recent results and new proof techniques concerning unambiguity and related topics from automata theory. There were 26 on-site participants from nine different countries (Belgium, Czech Republic, France, Germany, India, Italy, Poland, UK), and further 10 remote participants from seven countries (France, Germany, Poland, Sweden, Switzerland, UK, USA).

The central topic of the seminar was *unambiguous automata*. An automaton is unambiguous if it can make nondeterministic choices, but it is guaranteed that for every input there is *at most one accepting* run. There have recently been numerous new results concerning unambiguous automata; at the same time, a lot of natural and interesting problems have been open for decades. Before the seminar, we identified the following key topics/open problems:

- **Unambiguous Finite Automata** What is the state complexity of the complementation of unambiguous automata? Here, the state complexity refers to how the number of states of the resulting automaton depends on the number of states of the original automaton.
- **Unambiguous versions of infinite state systems, such as vector addition systems with states (VASS) or register automata** Open problems concerning such systems are, for instance: What can be new techniques for proving lower bounds for the containment problem? Are languages accepted by unambiguous register automata with guessing closed under complement?
- **Unambiguous tree automata** One of the most important open questions is how to decide whether a given tree-regular language is recognizable by an unambiguous automaton.
- **Büchi automata and probabilistic automata** What is the computational complexity of the containment of unambiguous Büchi automata?
- **Tropical automata** For this class of weighted automata one of the most important and long standing open questions is whether a given series is polynomially ambiguous.

The seminar was planned to consist of talks and working group sessions, where participants could work on-site on open problems. In order to integrate all participants and to initiate new collaborations, we started the seminar on Monday with introductory talks, where every participant shortly introduced herself to the group. In these introductory sessions, it was also possible to announce open problems the participants were interested to work on during the seminar. We had additionally collected such open problems before the seminar to make them available to the participants in advance.

The second day of the seminar (Tuesday) was dedicated to presentations given by the participants. This day started with an invited talk by Denis Kuperberg on good-for-games automata. Later the day, eight participants of the seminar presented short contributed talks on topics related to unambiguity.

Wednesday began with the invited talk by Gabriele Puppis on register automata. Later, a single contributed talk was given and the whole afternoon was devoted to an excursion and group work.

On Thursday morning, Wojtek Czerwiński gave an invited talk on future-determinisation. After that, four contributed talks were given, and the late afternoon was devoted to work in subgroups.

Finally, on Friday morning we held a closing ceremony. The rest of the day was left to participants to summarise their discussions in subgroups and prepare for departure.

During all days, we have used Schloss Dagstuhl's excellent technical facilities to connect and communicate to remote participants of the seminar. Our experiences regarding such a hybrid Dagstuhl seminar are twofold. On the one hand, it is practical to give remote participants the opportunity to follow the on-site presentations (and Sylvain Lombardi also gave a remote talk). On the other hand, our main aim was to bring together researchers to actually work on concrete problems. It was difficult to integrate participants in group work, when groups gather at different places in the facilities, or when important discussions are led during the excursion or the dinner. We appreciated very much the opportunity to gather on-site at Schloss Dagstuhl after a long time of only non-physical meetings due to the Covid pandemics. As summarized in Session 4, several new collaborations between participants of the seminar have been initiated. We hope that the seminar has inspired new ideas, and interesting new results will be published by the participants.

We would like to warmly thank Schloss Dagstuhl for making this seminar possible. We especially would like to thank for the great help and support in the organization before and during the seminar.

2 Contents

Executive Summary
Thomas Colcombet, Karin Quaas, Michał Skrzypczak . . . 57

Overview of Talks

Regular Tree Algebras
Achim Blumensath . . . 62

Between Deterministic and Nondeterministic Quantitative Automata
Udi Boker . . . 62

Unambiguous automata acceptance?
Dmitry Chistikov . . . 63

Computational complexity of universality and related problems for unambiguous context-free grammars
Lorenzo Clemente . . . 63

On Future-Determinization of Unambiguous Systems (Invited Talk)
Wojciech Czerwiński . . . 63

Alternation as a tool for disambiguation
Simon Jantsch . . . 64

Good-for-Games Automata: State of the Art and Perspectives (Invited Talk)
Denis Kuperberg . . . 64

Quotients, Coverings and Conjugacy of Unambiguous Automata
Sylvain Lombardy . . . 64

Active learning sound negotiations
Anca Muscholl . . . 65

Lower bound for unambiguous arithmetic circuits via Hankel matrix
Pierre Ohlmann . . . 65

Unambiguous Automata for Data Languages (Invited Talk)
Gabriele Puppis . . . 65

On Uniformization in the Full Binary Tree
Alexander Rabinovich . . . 66

State complexity of complementing unambiguous automata
Mikhail Raskin . . . 66

Problems on unambiguous WAs and PAs
Mahsa Shirmohammadi . . . 67

Unambiguity in Transducer Theory
Sarah Winter . . . 67

Working Groups . . . 67

Open Problems . . . 68

Characterizing the counter hierarchy of unambiguous automata
Georg Zetzsche . . . 68

Program synthesis for unambiguous devices
Emmanuel Filiot 68

Deciding efficiently history-determinism for ω-automata
Denis Kuperberg 68

What is the complexity of constructing unambiguous automata
Denis Kuperberg 68

Characterizing classes of languages with atoms from internal closure operations
Antoine Mottet 69

The zeroness problem
Lorenzo Clemente 69

Stronger versions of inclusion of probabilistic automata
Guillermo Alberto Perez 69

The state complexity of unambiguous Büchi automata
Simon Jantsch 69

Universality of register automata over ordered domains
Karin Quaas 70

Decomposition of finitely unambiguous automata
Nathanaël Fijalkow 70

Better bounds on complementing unambiguous automata
Michael Raskin 70

Panel Discussions 70

Participants 71

Remote Participants 71

3 Overview of Talks

3.1 Regular Tree Algebras

Achim Blumensath (Masaryk University – Brno, CZ)

Joint work of Mikołaj Bojańcyck, Thomas Colcombet, Bartek Klin

We introduce a class of algebras that can be used as recognisers for regular tree languages. We show that it is the only such class that forms a pseudo-variety and we prove the existence of syntactic algebras.

3.2 Between Deterministic and Nondeterministic Quantitative Automata

Udi Boker (Reichman University – Herzliya, IL)

Joint work of Udi Boker, Karoliina Lehtinen

There is a challenging trade-off between deterministic and nondeterministic automata, where the former suit various applications better, however at the cost of being exponentially larger or even less expressive.

This gave birth to many notions in between determinism and nondeterminism, aiming at enjoying, sometimes, the best of both worlds. Some of the notions are yes/no ones, for example initial nondeterminism (restricting nondeterminism to allowing several initial states), and some provide a measure of nondeterminism, for example the ambiguity level.

We analyze the possible generalization of such notions from Boolean to quantitative automata, and suggest that it depends on the following key characteristics of the considered notion N – whether it is syntactic or semantic, and if semantic, whether it is word-based or language-based.

A syntactic notion, such as initial nondeterminism, applies as is to a quantitative automaton A, namely $N(A)$. A word-based semantic notion, such as unambiguity, applies as is to a Boolean automaton $t - A$ that is derived from A by accompanying it with some threshold value t, namely $N(t - A)$. A language-based notion, such as history determinism, also applies as is to A, while in addition, it naturally generalizes into two different notions with respect to A itself, by either: i) taking the supremum of $N(t - A)$ over all thresholds t, denoted by $Th - N(A)$; or ii) generalizing the basis of the notion from a language to a function, denoted simply by $N(A)$. While in general $N(A)$ implies $Th - N(A)$ implies $N(t - A)$, we have for some notions that $N(A)$ and $Th - N(A)$ are equivalent and for some not. (For measure notions, "implies" stands for ¿= with respect to the nondeterminism level.)

We classify numerous notions known in the Boolean setting according to their characterization above, generalize them to the quantitative setting and look into relations between them. The generalized notions open new research directions with respect to quantitative automata, and provide insights on the original notions with respect to Boolean automata.

3.3 Unambiguous automata acceptance?

Dmitry Chistikov (University of Warwick – Coventry, GB)

Given a nondeterministic finite automaton (NFA) with m transitions and an input word of length ℓ, one can decide in time $O(m\ell)$ if the word is accepted. If $m \approx n^2$ (where n is the number of states) and $\ell \approx n$, this running time is essentially cubic in n. I don't know if significantly faster algorithms exist for this and several related problems. Can we obtain speed-ups if the automaton is known to be unambiguous?

3.4 Computational complexity of universality and related problems for unambiguous context-free grammars

Lorenzo Clemente (University of Warsaw, PL)

In this talk I recall a classic approach to decide universality of unambiguous context-free grammars. It originates in the work of Chomsky and Schutzenberger, who showed that the (commutative) power series of an unambiguous grammar is algebraic. Based on this fact, one can reduce in PTIME the universality problem to the zeroness problem for a related algebraic power series, and in turn the latter problem can be shown to be PTIME reducible to the existential fragment of the first-order theory of the reals. Since the last problem is in PSPACE by the result of Canny, it follows that universality of unambiguous grammars is in PSPACE. Whether the latter problem actually belongs to a lower complexity class is advertised as an open problem.

3.5 On Future-Determinization of Unambiguous Systems (Invited Talk)

Wojciech Czerwiński (University of Warsaw, PL)

Joint work of Wojciech Czerwiński, Piotr Hofman

I will present you a result based on an on-going work jointed with Piotr Hofman. We have shown that language equivalence is decidable for unambiguous vector addition systems with states (VASS) (acceptance is by state). Id like to focus more on our technique: we have proven that each unambiguous VASS can be determinized in a certain sense (with a use of some additional information about the future), which we call future-determinization. This result makes use of some known regular-separability results. There is a hope that similar techniques can be possible for other unambiguous systems and maybe even point to some high-level connection between separability and unambiguity notions.

3.6 Alternation as a tool for disambiguation

Simon Jantsch (TU Dresden, DE)

Joint work of Simon Jantsch, David Müller, Christel Baier, Joachim Klein, Yufei Liu

In this talk we show how alternating automata can be used as a tool to devise disambiguation algorithms for nondeterministic automata over finite and infinite words. The main idea is to use conjunction and complementation, both of which can be naturally implemented in alternating automata, to restrict nondeterministic branching in a way that preserves the language and makes sure that for any given word only one choice leads to acceptance. A notion of unambiguity for alternating automata is introduced, and we show that standard alternation removal techniques preserve it. The approach works well for automata on finite words and restricted forms of automata (namely very weak ones) but we show that it fails for arbitrary nondeterministic Büchi automata (NBA), and discuss the issues that arise. Finally, we speculate about the relationship between complementation and disambiguation and possible consequences for the state complexity of disambiguating NBA.

3.7 Good-for-Games Automata: State of the Art and Perspectives (Invited Talk)

Denis Kuperberg (ENS – Lyon, FR)

Joint work of Denis Kuperberg, Marc Bagnol, Udi Boker, Emile Hazard, Michał Skrzypczak

In the setting of regular languages of infinite words, Good-for-Games (GFG) automata can be seen as an intermediate formalism between determinism and nondeterminism, with advantages from both worlds. Indeed, like deterministic automata, GFG automata enjoy good compositional properties (useful for solving games and composing automata and trees) and easy inclusion checks. Like nondeterministic automata, they can be exponentially more succinct than deterministic automata. Since their introduction in 2006 by Henzinger and Piterman, there has been a steady research effort to uncover the prop- erties of GFG automata, with some surprises along the way. I will give an overview of the results obtained in this line of research, the proof techniques typically used, and the remaining open problems and conjectures.

3.8 Quotients, Coverings and Conjugacy of Unambiguous Automata

Sylvain Lombardy (University of Bordeaux, FR)

Joint work of Sylvain Lombardy, Marie-Pierre Bal, Jacques Sakarovitch

In this talk, I shall recall the definitions of quotients and coverings, which are useful tools to transform the structure of an automaton while preserving the unambiguity. We shall see that it is always possible to turn an unambiguous automaton to any equivalent one using these tools. The construction of this transformation is based on a more algebraic concept, that is the conjugacy of automata. An open question concerning the transformation of an automaton to another one is the state complexity of the transitional automata. This talk is based on a work with Marie-Pierre Bal and Jacques Sakarovitch.

3.9 Active learning sound negotiations

Anca Muscholl (University of Bordeaux, FR)

Joint work of Anca Muscholl, Igor Walukiewicz

Sound deterministic negotiations are models of distributed systems, a kind of Petri nets or Zielonka automata with additional structure. We show that the additional structure allows to minimize such negotiations. Based on minimisation we present two Angluin-style learning algorithms for sound deterministic negotiations. The two algorithms differ in the kind of membership queries they use, and both have similar (polynomial) complexity as Angluins algorithm.

3.10 Lower bound for unambiguous arithmetic circuits via Hankel matrix

Pierre Ohlmann (CNRS – Paris, FR)

Joint work of Pierre Ohlmann, Nathanaël Fijalkow, Guillaume Lagarde, Olivier Serre

This talk is about arithmetic circuits, for which a major goal is to devise lower bounds: can one find polynomials such that any arithmetic circuit computing them has to be large. I will present a new characterization of the size of the smallest arithmetic circuit computing a given non-associative polynomial, in term of the rank of a so-called Hankel matrix. This generalizes an important result of Nisan (1992); it is based on a result for weighted tree-automata due to Bozapalidis and Loscou-Bozapalidou (1984).

We will then show how the characterization can be used to establish an exponential lower bound for (associative) unambiguous circuits computing the permanent polynomial.

3.11 Unambiguous Automata for Data Languages (Invited Talk)

Gabriele Puppis (University of Udine, IT)

Joint work of Gabriele Puppis, Thomas Colcombet, Michał Skrzypczak

I will present the status of an ongoing research work with Thomas Colcombet and Michał Skrzypczak about unambiguity in register automata (register automata, or finite memory automata, are automata that can describe languages over an infinite alphabet). Differently from finite state automata, the amount of non-determinism allowed in register automata has an impact on the expressive power and the closure properties of the recognized class of languages, as well as on the complexity of some fundamental decision problems. For example, deterministic register automata are strictly less expressive than non-deterministic ones, they are closed under complement, but not under mirroring. On the other hand, non-deterministic register automata (with guessing) are closed under mirroring, but not under complement. It comes natural then to study the intermediate class of unambiguous register automata with guessing. Recently (LICS'21), this class has been shown to enjoy a decidable equivalence problem and is believed to be effectively closed under complement. However, proving this closure property turned out to be more difficult than expected. I will present some ideas and partial results along this goal, mentioning a few other conjectures related to the expressive power of unambiguous register automata.

3.12 On Uniformization in the Full Binary Tree

Alexander Rabinovich (Tel Aviv University, IL)

Joint work of Alexander Rabinovich, Amit Shomrat

Gurevich and Shelah proved that the uniformization property fails for Monadic Second-Order logic (MSO) over the full binary tree, i.e., there is a formula $A(X, Y)$ in MSO such that no MSO formula uniformizes it (over the full binary tree).

The cross-section of a relation $R(X, Y)$ at d is the set of all e such that $R(d, e)$ holds. We prove:
Theorem (Finite-cross Section): If every cross-section of an MSO definable relation is finite then it has an MSO definable uniformizer.
Theorem (Uncountable-cross Section): There is an MSO definable relation R such that every MSO definable relation included in R and with the same domain as R has an uncountable cross-section.

3.13 State complexity of complementing unambiguous automata

Mikhail Raskin (TU München, DE)

Not so long ago, even a polynomial upper bound on state complexity of recognising the complement of the language of an unambiguous finite automaton felt plausible. Now it does not, but what else do we know? Not so much. In this talk I plan to briefly show the approaches that give the best currently known lower and upper bounds for the state complexity of complementation in the unary and binary alphabets; and draw a (straightforward) game reformulation of the large-alphabet problem in the hope it will inspire someone to prove the exponential lower bound in that case.

3.14 Problems on unambiguous WAs and PAs

Mahsa Shirmohammadi (University Paris Diderot, FR)

In this survey talk we recall a proof of the classical results on weighted automata (WAs) over fields, that given a weighted function f realisable with WAs, the size of a minimal canonical WA computing f is equivalent to the rank of the Handle matrix of f. We also briefly talk about recent results of Bell and Smertnig showing that every weighted function taking values in a finitely generated subgroup of a field (and zero) can be realised with an unambiguous WA. We conclude the talk with open problems and directions for future research.

3.15 Unambiguity in Transducer Theory

Sarah Winter (UL – Brussels, BE)

Joint work of Sarah Winter, Emmanuel Filliot

This talk surveys some introductory results regarding unambiguity in transducer theory. Transducers are automata with output; they recognize relations. A transducer is unambiguous if for each word u from its domain there is a unique accepting run with input u.

In more detail, we show that the classes of functions recognized by functional transducers and unambiguous transducers coincide. We also show that unambiguity, while necessary for one-way transducers, can be traded for determinism at the price of two-wayness.

This is a joint work with Emmanuel Filliot.

4 Working Groups

The participants were not formally decomposed into working groups, though many small groups have been interacting and evolving during the program.

- Achim Blumensath and Michał Skrzypczak worked on the *Thin Tree Conjecture*.
- Emmanuel Filiot, Karin Quaas, and Sarah Winter started a new collaboration on synthesis for register automata. A collaboration on a standing open problem related to unambiguity in transducer models emerged during the various discussions. Specifically, the problem concerns the possibility of transforming any streaming string transducer with boundedly many outputs per input into an equivalent finite union of unambiguous functional transducers. The collaboration involved the researchers Emmanuel Filiot, Ismaël Jecker, Christof Löding, Anca Muscholl, Gabriele Puppis, and Sarah Winter, and it is still active. A paper with the outcome of this collaboration will likely be produced in the near future.
- Another research collaboration emerged between Anca Muscholl and Gabriele Puppis on the possibility of having minimal and canonical forms of streaming string transducers, as well as an Angluin-style learning algorithm for these types of transducers.
- Wojciech Czerwiński, Diego Figueira, Gabriele Puppis, Mikhail Raskin, and Georg Zetzsche have collaborated on a decision problem concerning the separability of synchronous relations (i.e. relations represented by letter-to-letter transducers) by means of recognizable relations (i.e. relations obtained as finite unions of products of regular languages).

- Thomas Colcombet and Alexander Rabinovich have been working on the uniformization questions for monadic second-order logic over countable ordinals. The open problems regarding this topic are solved and a paper is under writing.
- Karin Quaas and Narayanan Krishna Shankara have initiated a new collaboration on temporal logics for real-timed systems.

5 Open Problems

5.1 Characterizing the counter hierarchy of unambiguous automata

Georg Zetzsche (MPI Kaiserslautern, DE, georg@mpi-sws.org)

Given a counter language, how many counters does it require to be recognizable by an unambiguous counter machine? This problem is undecidable for non-deterministic counter machines.

5.2 Program synthesis for unambiguous devices

Emmanuel Filiot (UL Bruxelles, BE, efiliot@gmail.com)

Is it the case that for every regular specification $\varphi \in \mathrm{REG}(\Sigma^* \times \Sigma^*)$ there exists an unambiguous transducer which realises this specification?

5.3 Deciding efficiently history-determinism for ω-automata

Denis Kuperberg (LIP, ENS Lyon, FR, denis.kuperberg@ens-lyon.fr)

The $G2$ conjecture states that a parity automaton over ω-words is history-deterministic if and only if there is a winning strategy in a specific two pebbles game (hence the name G2). The conjecture is only known to hold for very low levels of the parity hierarchy.

5.4 What is the complexity of constructing unambiguous automata

Denis Kuperberg (LIP, ENS Lyon, FR, denis.kuperberg@ens-lyon.fr)

Joint work of Denis Kuperberg, Shaull Amalgor

What is the complexity of the following problem: given a non-deterministic finite automaton A and an integer n in binary, does there exist a deterministic (unambiguous) finite automaton B that accepts $L(A)$ and has less than n states?

5.5 Characterizing classes of languages with atoms from internal closure operations

Antoine Mottet (Charles Univ. Prague, CZ, mottet@karlin.mff.cuni.cz)

We say that an operation f of finite arity over the set of data words preserves a language L if $f(L, \ldots, L)$ is a subset of L. For example, if L is recognizable by a register automaton with an atom structure A, then every automorphism of A preserves L. The internal closure properties of data languages have not been considered so far. In particular, can one understand the complexity of a language (i.e., deterministically recognizable, unambiguously recognizable, recognizable, with/without guessing) in terms of the operations preserving a language? This question was answered positively for Turing machines (recognizing several variants of constraint satisfaction problems) where closure properties have been central in characterizing the (descriptive) complexity of problems

5.6 The zeroness problem

Lorenzo Clemente (Univ. of Warsaw, PL, clementelorenzo@gmail.com)

What is the complexity and decidability of the zeroness problem, ie deciding if a machine representing.a function computes the everywhere null constant. The question is of interesting, in particular, for weighted grammars over a field, unary polynomial automata, weighted Parikh automata, weighted vector addition systems with states.

5.7 Stronger versions of inclusion of probabilistic automata

Guillermo Alberto Perez (Univ. Antwerpen, BE, guillermoalberto.perez@uantwerpen.be)

Can one prove decidability of the language containment problem for probabilistic automata with bounded ambiguity without having to assume Schanuel's conjecture?

5.8 The state complexity of unambiguous Büchi automata

Simon Jantsch (TU Dresden, DE, simon.jantsch@tu-dresden.de)

The first asks the general question: can we probe the 2^n lower bound for the problem in the infinite words case? The second and third problems focus on specific LTL formulae over infinite words and asks about the lower bound for unambiguous automata recognising their languages.

5.9 Universality of register automata over ordered domains

Karin Quaas (Univ. Leipzig, DE, quaas@informatik.uni-leipzig.de)

Is the universality problem for unambiguous register automata over the integers with order and constants decidable? If yes, what is the complexity?

5.10 Decomposition of finitely unambiguous automata

Nathanaël Fijalkow (CNRS, Univ. Bordeaux, FR, nathanael.fijalkow@labri.fr)

Is it possible to decompose finitely ambiguous register automata into finitely many unambiguous ones?

5.11 Better bounds on complementing unambiguous automata

Michael Raskin (TU München, DE, raskin@mccme.ru)

It is now known that complementing the language of an n-state unambiguous finite automaton might yield a language not recognisable by some nondeterministic finite automata with fewer than $n^{(\log\log\log n)^{\Omega(1)}}$ for unary alphabet and there is an upper bound of $n^{O(\log n)}$. In the binary case the lower bound is $n^{\Omega(\log n)}$ but the upper bound is still exponential; same for large alphabets. How do we close the non-unary complement gap?

6 Panel Discussions

No panel discussions were organised.

Participants

- Achim Blumensath
Masaryk University – Brno, CZ
- Udi Boker
Reichman University – Herzliya, IL
- Dmitry Chistikov
University of Warwick – Coventry, GB
- Lorenzo Clemente
University of Warsaw, PL
- Thomas Colcombet
CNRS – Paris, FR
- Wojciech Czerwinski
University of Warsaw, PL
- Diego Figueira
CNRS & Université de Bordeaux, FR
- Emmanuel Filiot
UL – Brussels, BE
- Simon Jantsch
TU Dresden, DE
- Ismaël Jecker
University of Warsaw, PL
- Stefan Kiefer
University of Oxford, GB
- Shankaranarayanan Krishna
Indian Institute of Technology – Mumbai, IN
- Denis Kuperberg
ENS – Lyon, FR
- Karoliina Lehtinen
Aix-Marseille University, FR
- Antoine Mottet
Charles University – Prague, CZ
- Anca Muscholl
University of Bordeaux, FR
- Pierre Ohlmann
CNRS – Paris, FR
- Guillermo A. Pérez
University of Antwerp, BE
- Jakob Piribauer
TU Dresden, DE
- Gabriele Puppis
University of Udine, IT
- Karin Quaas
Universität Leipzig, DE
- Alexander Rabinovich
Tel Aviv University, IL
- Michael Raskin
TU München, DE
- Mahsa Shirmohammadi
University Paris Diderot, FR
- Michal Skrzypczak
University of Warsaw, PL
- Sarah Winter
UL – Brussels, BE
- Georg Zetzsche
MPI-SWS – Kaiserslautern, DE

Remote Participants

- Christel Baier
TU Dresden, DE
- Johanna Björklund
University of Umeå, SE
- Michaël Cadilhac
DePaul University – Chicago, US
- Antonio Casares
University of Bordeaux, FR
- Nathanael Fijalkow
University of Bordeaux, FR
- Mika Göös
EPFL Lausanne, CH
- Arthur Jaquard
CNRS – Paris, FR
- Stefan Kiefer
University of Oxford, GB
- Christof Löding
RWTH Aachen, DE
- Sylvain Lombardy
University of Bordeaux, FR
- Radek Piórkowski
University of Warsaw, PL
- Mikhail V. Volkov
Ural Federal University – Ekaterinburg, RU

Report from Dagstuhl Seminar 21461

Descriptive Set Theory and Computable Topology

Edited by

Mathieu Hoyrup[1], Arno Pauly[2], Victor Selivanov[3], and Mariya I. Soskova[4]

1 **LORIA & INRIA Nancy, FR,** mathieu.hoyrup@inria.fr
2 **Swansea University, GB,** arno.m.pauly@gmail.com
3 **A. P. Ershov Institute – Novosibirsk, RU,** vseliv@iis.nsk.su
4 **University of Wisconsin – Madison, US,** msoskova@math.wisc.edu

Abstract

Computability and continuity are closely linked – in fact, continuity can be seen as computability relative to an arbitrary oracle. As such, concepts from topology and descriptive set theory feature heavily in the foundations of computable analysis. Conversely, techniques developed in computability theory can be fruitfully employed in topology and descriptive set theory, even if the desired results mention no computability at all. In this Dagstuhl Seminar, we brought together researchers from computable analysis, from classical computability theory, from descriptive set theory, formal topology, and other relevant areas. Our goals were to identify key open questions related to this interplay, to exploit synergies between the areas and to intensify collaboration between the relevant communities.

Seminar November 14–19, 2021 – http://www.dagstuhl.de/21461
2012 ACM Subject Classification Theory of computation → Computability; Theory of computation → Logic; Mathematics of computing → Topology
Keywords and phrases computable analysis, enumeration degrees, quasi-polish spaces, synthetic topology
Digital Object Identifier 10.4230/DagRep.11.10.72
Edited in cooperation with Ingall, Corrie

1 Executive Summary

Mathieu Hoyrup (LORIA & INRIA Nancy, FR, mathieu.hoyrup@inria.fr)
Arno Pauly (Swansea University, GB, arno.m.pauly@gmail.com)
Victor Selivanov (A. P. Ershov Institute – Novosibirsk, RU, vseliv@iis.nsk.su)
Mariya I. Soskova (University of Wisconsin – Madison, US, msoskova@math.wisc.edu)

Research area and topics

Descriptive set theory traditionally studies the complexity of subsets of and functions between Polish spaces (which are the completely metrizable separable spaces). As a mathematical area, it has well-established interactions with set theory and real analysis. Its canonical textbook is Kechris [11].

Following the developments in (classical) descriptive set theory, also the area of effective descriptive set theory flourished. In a way, this is the result of replacing *continuous* by *computable* everywhere, and by replacing arbitrary countable union by effective ones. Here, the canonical textbook is Moschovakis' [19]. While classical descriptive set theory is trivial on discrete spaces, the results from effective descriptive set theory on $\mathbb{N}$ often generalize results from computability theory. While this is rarely emphasized (see [20] for an exception), one can recover classical descriptive set theory from effective descriptive set theory by relativization – provided that theorems are phrased in the right way.

Recent years have seen a lot of interest in the interplay between descriptive set theory and theoretical computer science going beyond the natural meeting point of effective DST. Four core developments outlined below are particularly relevant for the meeting:

DST on spaces of interest for TCS

Certain classes of topological spaces were revealed as applicable to reasoning about the semantics of programming. The most famous example is domain theory, but Escardo's synthetic topology [6] or the relationship between well-structured transition systems and Noetherian spaces revealed by Goubault-Larrecq [7] were also very influential. The spaces relevant for TCS are often not Hausdorff, and in particular not Polish. Selivanov pioneered the call for a development of descriptive set theory for these spaces [28, 29]. A break-through was achieved by de Brecht [3] with identifying the class of quasi-Polish spaces as a common generalization of Polish spaces and omega-continuous domains, and by showing that many core results of descriptive set theory can be extended to quasi-Polish spaces.

In computable analysis, we typically work with the category of admissible represented spaces (equivalently, with QCB_0-spaces, i.e. T_0-quotients of countably-based spaces) [24]. This is a Cartesian-closed category, meaning that we can form function spaces. This is a very natural requirement from a TCS-perspective, but does not preserve being countably-based. How descriptive set theory works on non-countably-based spaces is still a mystery. de Brecht, Selivanov and Schröder have undertaken initial investigations, in particular into the Kleene-Kreisel spaces in [27, 26, 5]. Hoyrup has shown that even very simple non-countably-based spaces such as $\mathcal{O}(\omega^\omega)$ exhibit very unfamiliar behaviour compared to the usual DST [8].

Synthetic DST

de Brecht and Pauly observed a connection between synthetic topology (which in turn can be seen as the theory of functional programming [6]), models of hypercomputation and descriptive set theory [22, 23, 4]. This connection opens up the opportunity to apply reasoning styles about models of computation to descriptive set theory. Work by Kihara on the Jayne-Rogers conjecture has shown significant potential of this approach for solving open questions in descriptive set theory [12]. There is also a hope that this theory can connect to other parts of TCS such as descriptive complexity.

DST and computability theory

Traditional computability theory, in particular the study of enumeration degrees, was related to the study of topological spaces via the notion of point degree spectrum introduced by Kihara and Pauly [13], building on earlier work by J. Miller [17]. This lets us reason about the degrees of individual points in a topological space, and understand properties of the space in terms of what degrees are realized there. This technique was already used to resolve a long-standing open question by Jayne ([9], also [21]) on the number of sigma-homeomorphism types of Polish spaces in [13].

This connection is bidirectional, and also allows for the application of topological arguments in computability theory. As such, it has inspired a flurry of recent developments in the area of enumeration degrees by J. Miller, M. Soskova and others [2, 18, 1, 16]. Particularly remarkable here is the existence of non-total almost-total enumeration degrees. This is a purely recursion-theoretic statement, but the various known proofs all invoke topological arguments such as Brouwer's Fixed Point theorem, Urysohn's metrization theorem or Hurewicz' and Wallmann's characterization of countably-dimensional Polish spaces.

Of a similar flavour (but the precise connections are still unclear) is the approach to fractal geometry and Hausdorff dimension via *effective dimension* of points, defined via Kolmogorov complexity [14]. This approach has already been demonstrated to provide strengthening of core results of fractal geometry, in many cases by rendering inessential restrictions to measurable sets. This includes a reproof of known answer to the two-dimensional Kakeya-conjecture [15].

coPolish spaces and computational complexity

In general, it seems that computational complexity of algorithms from computable analysis needs second-order complexity (Kawamura and Cook [10]). For certain spaces, however, runtimes of algorithms are still first-order objects [25]. Ongoing work by de Brecht and Schröder has shown that this holds for the coPolish spaces, a dual notion to the quasi-Polish spaces. As such, it seems that "spaces where descriptive set theory is well-behaved" is the dual notion to "spaces where complexity theory is well-behaved". This merits further attention by a broader community.

Seminar structure

As our seminar brought together researchers from previously rather disconnected areas, we included several tutorial talks of one hour each to introduce the various facets of our seminar topics to everyone. The talks covered *Quasi-Polish spaces* (Matthew de Brecht), *Quantitative Coding and Complexity Theory of Continuous Data* (Martin Ziegler), *CoPolish spaces and Effectively Hausdorff spaces* (Matthias Schröder), *New directions in Synthetic Descriptive Set Theory* (Takayuki Kihara), *Categorical aspects of Descriptive Set Theory* (Ruiyuan Chen), *Topology reflected in the enumeration degrees* (Joseph S. Miller), *Point-free Descriptive Set Theory* (Alex Simpson) and *Borel combinatorics fail in HYP* (Linda Westrick).

In addition, we had many short (fifteen minute) talks introducing topics or open questions. The prompt for these talks was "What theorem do you want to prove during/following this workshop?", and we are excited to learn what will come from this in the next months.

Challenges in hybrid Dagstuhl meetings

While the organizers and most participants had grown very accustomed to virtual meetings, the setting for our seminar was decidedly hybrid: About half of the participants were present in person, half were participating remotely. The same split applied to the organizing team.

The Dagstuhl team had equipped our main meeting room with multiple cameras and microphones (including microphones suspended from the ceiling throughout the room to pick up audience contributions). The equipment was controlled by several volunteers amongst the participants, and we are very grateful to Nikolay Bazhenov, Josiah Jacobsen-Grocott and Eike Neumann for having performed this crucial role. This setup made interactions in the lecture theatre between remote and in-person participants almost seamless.

A feature we felt was both crucial for a successful Dagstuhl seminar and difficult to accomplish in a hybrid setting are the informal discussions taking place in smaller groups. Our approach was to make those slightly less informal, and to use the collaboration platform Slack for arranging meetings. Slack also served for asking questions somewhat after the talks. This was somewhat successful, and several fruitful discussions involving both remote and in-person participants took place. It is difficult to ascertain though how much potential for additional discussions remained untapped.

References

1 Uri Andrews, Hristo A. Ganchev, Rutger Kuyper, Steffen Lempp, Joseph S. Miller, Alexandra A. Soskova, and Mariya I. Soskova. On cototality and the skip operator in the enumeration degrees. preprint.

2 Uri Andrews, Gregory Igusa, Joseph S. Miller, and Mariya I. Soskova. Characterizing the continuous degrees. *Israel Journal of Mathematics*, 234:743–767, 2019.

3 Matthew de Brecht. Quasi-Polish spaces. *Annals of Pure and Applied Logic*, 164(3):354–381, 2013.

4 Matthew de Brecht and Arno Pauly. Noetherian Quasi-Polish spaces. In Valentin Goranko and Mads Dam, editors, *26th EACSL Annual Conference on Computer Science Logic (CSL 2017)*, volume 82 of *LIPIcs*, pages 16:1–16:17. Schloss Dagstuhl, 2017.

5 Matthew de Brecht, Matthias Schröder, and Victor Selivanov. Base-complexity classifications of QCB_0-spaces. In Arnold Beckmann, Victor Mitrana, and Mariya Soskova, editors, *Evolving Computability*, pages 156–166. Springer, 2015.

6 Martín Escardó. Synthetic topology of datatypes and classical spaces. *Electronic Notes in Theoretical Computer Science*, 87, 2004.

7 Jean Goubault-Larrecq. *Non-Hausdorff Topology and Domain Theory*. New Mathematical Monographs. Cambridge University Press, 2013.

8 Mathieu Hoyrup. Results in descriptive set theory on some represented spaces. arXiv 1712.03680, 2017.

9 J. E. Jayne. The space of class α Baire functions. *Bull. Amer. Math. Soc.*, 80:1151–1156, 1974.

10 Akitoshi Kawamura and Stephen Cook. Complexity theory for operators in analysis. *ACM Transactions on Computation Theory*, 4(2), 2012.

11 A.S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer, 1995.

12 Takayuki Kihara. Decomposing Borel functions using the Shore-Slaman join theorem. *Fundamenta Mathematicae*, 230, 2015. arXiv 1304.0698.

13 Takayuki Kihara and Arno Pauly. Point degree spectra of represented spaces. arXiv:1405.6866, 2014.

14 Jack Lutz. The dimensions of individual strings and sequences. *Information and Computation*, 187:49–79, 2003.

15 Jack H. Lutz and Neil Lutz. Algorithmic Information, Plane Kakeya Sets, and Conditional Dimension. In Heribert Vollmer and Brigitte Valle´e, editors, *34th Symposium on Theoretical Aspects of Computer Science (STACS 2017)*, volume 66 of *LIPIcs*, pages 53:1–53:13. Schloss Dagstuhl, 2017.

16 Ethan McCarthy. Cototal enumeration degrees and their application to computable mathematics. *Proceedings of the AMS*, 146:3541–3552, 2018.

17 Joseph S. Miller. Degrees of unsolvability of continuous functions. *Journal of Symbolic Logic*, 69(2):555 – 584, 2004.

18 Joseph S. Miller and Mariya I. Soskova. Density of the cototal enumeration degrees. *Annals of Pure and Applied Logic*, 2018.

19 Yiannis N. Moschovakis. *Descriptive Set Theory*, volume 100 of *Studies in Logic and the Foundations of Mathematics*. North-Holland, 1980.

20 Yiannis N. Moschovakis. Classical descriptive set theory as a refinement of effective descriptive set theory. *Annals of Pure and Applied Logic*, 162:243–255, 2010.

21 Luca Motto Ros, Philipp Schlicht, and Victor Selivanov. Wadge-like reducibilities on arbitrary quasi-polish spaces. *Mathematical Structures in Computer Science*, pages 1–50, 11 2014. arXiv 1204.5338.

22 Arno Pauly and Matthew de Brecht. Non-deterministic computation and the Jayne Rogers theorem. *Electronic Proceedings in Theoretical Computer Science*, 143, 2014. DCM 2012.

23 Arno Pauly and Matthew de Brecht. Descriptive set theory in the category of represented spaces. In *30th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 438–449, 2015.

24 Matthias Schröder. Extended admissibility. *Theoretical Computer Science*, 284(2):519–538, 2002.

25 Matthias Schröder. Spaces allowing type-2 complexity theory revisited. *Mathematical Logic Quarterly*, 50(4/5):443–459, 2004.

26 Matthias Schröder and Victor Selivanov. Hyperprojective hierarchy of QCB_0-spaces. *Computability*, 4, 2015. arXiv 1404.0297.

27 Matthias Schröder and Victor L. Selivanov. Some hierarchies of QCB_0-spaces. *Mathematical Structures in Computer Science*, 25(8):1799–1823, 2015. arXiv 1304.1647.

28 Victor L. Selivanov. Difference hierarchy in φ-spaces. *Algebra and Logic*, 43(4):238–248, 2004.

29 Victor L. Selivanov. Towards a descriptive set theory for domain-like structures. *Theoretical Computer Science*, 365(3):258–282, 2006.

2 Table of Contents

Executive Summary

Mathieu Hoyrup, Arno Pauly, Victor Selivanov, and Mariya I. Soskova 72

Tutorial Talks

Categorical aspects of DST
Ruiyuan (Ronnie) Chen 79

Tutorial on Quasi-Polish Spaces
Matthew de Brecht 79

New Directions in Synthetic Descriptive Set Theory
Takayuki Kihara 79

Topology reflected in the enumeration degrees
Joseph S. Miller 80

CoPolish Spaces and Effectively Hausdorff Spaces
Matthias Schröder 81

Tutorial: Quantitative Coding and Complexity Theory of Compact Metric Spaces
Martin Ziegler 81

Borel combinatorics fail in HYP
Linda Westrick 82

Short Talks

Continuity and Computability
Vasco Brattka 82

When does Wadge meet Tang and Pequignot?
Riccardo Camerlo 83

Algorithmic Learning of Structures
Ekaterina Fokina 83

Refuting Selman's theorem in the hyperenumeration degrees
Jun Le Goh 84

A characterization of $\mathbf{\Pi}^0_3$-completeness
Vassilios Gregoriades 84

There is no Good Notion of Quasi-Polish Convergence Spaces
Reinhold Heckmann 84

Descriptive complexity on represented spaces
Mathieu Hoyrup 85

Regularity properties, determinacy, and Solovay models
Daisuke Ikegami 85

Resource-bounded effective dimension and the point-to-set principle
Elvira Mayordomo 85

Computable presentations in topology
Alexander Melnikov 86

Topological spaces of countable structures
Russell G. Miller 86

Computable Endofunctors, Markov-computability and Relativization
Arno Pauly 87

Effective overtness of generalised Cantor spaces
Philipp Schlicht 88

Effective embedding and interpretations
Alexandra A. Soskova 88

The hyper enumeration degrees
Mariya I. Soskova 89

Working groups

Computable categoricity of Polish spaces
Nikolay Bazhenov, Ivan Georgiev, Jun Le Goh, Vassilios Gregoriades, Mathieu Hoyrup, Iskander Shagitovich Kalimullin, Steffen Lempp, Alexander Melnikov, Russell G. Miller, Eike Neumann, Keng Meng Ng, Arno Pauly, Alexandra A. Soskova, and Daniel Turetsky 90

AE-theory of enumeration degree structures
Steffen Lempp, Jun Le Goh, Keng Meng Ng, and Mariya I. Soskova 90

Open problems

Questions on left-c.e. reals
Iskander Shagitovich Kalimullin 91

Which Compact Metric spaces do/don't admit polynomially admissible representations?
Martin Ziegler 91

Participants 92

Remote Participants 92

3 Tutorial Talks

3.1 Categorical aspects of DST

Ruiyuan (Ronnie) Chen (McGill University – Montreal, CA)

We gave an introduction to categorical structures of interest in (classical) descriptive set theory, including axioms on limits and colimits in categories of topological and Borel spaces [3], duality with countably presented algebras, locales and point-free descriptive set theory [4], and connections with infinitary propositional and first-order logic [1, 2].

References

1 R. Chen, *Borel functors, interpretations, and strong conceptual completeness for $\mathcal{L}_{\omega_1\omega}$*, Trans. Amer. Math. Soc. **372** (2019), no. 12, 8955–8983.

2 R. Chen, *Representing Polish groupoids via metric structures*, preprint, `https://arxiv.org/abs/1908.03268`, 2019.

3 R. Chen, *A universal characterization of standard Borel spaces*, preprint, `https://arxiv.org/abs/1908.10510`, 2019.

4 R. Chen, *Borel and analytic sets in locales*, preprint, `https://arxiv.org/abs/2011.00437`, 2020.

3.2 Tutorial on Quasi-Polish Spaces

Matthew de Brecht (Kyoto University, JP)

We give a brief introduction to quasi-Polish spaces and their connections with Descriptive set theory, Domain theory, Computable topology, Geometric logic, and Duality.

3.3 New Directions in Synthetic Descriptive Set Theory

Takayuki Kihara (Nagoya University, JP)

Main reference Takayuki Kihara: "Lawvere-Tierney topologies for computability theorists", CoRR, Vol. abs/2106.03061, 2021.
URL https://arxiv.org/abs/2106.03061
Main reference Takayuki Kihara: "Lawvere-Tierney topologies for computability theorists. arxiv: 2106.03061 (2021). Takayuki Kihara: Rethinking the notion of oracle: A bridge between synthetic descriptive set theory and effective topos theory", in preparation (2022).
Main reference Arno Pauly, Matthew de Brecht: "Descriptive Set Theory in the Category of Represented Spaces", in Proc. of the 30th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2015, Kyoto, Japan, July 6-10, 2015, pp. 438–449, IEEE Computer Society, 2015.
URL https://doi.org/10.1109/LICS.2015.48

Let us reconsider what an *oracle* is. At least three different perspectives of oracle can be presented. The first perspective is the most standard one, which is to think of an oracle as a *blackbox*, represented as a set, a function, an infinite string, etc. If we think of a blackbox as just a container to store an input data (whose data type is stream), as some people say, an oracle is merely an input stream. The latter idea is also quite standard nowadays.

The second perspective is based on a recent approach taken e.g. by de Brecht and Pauly to develop *synthetic descriptive set theory*, which is, according to them, the idea that descriptive set theory can be reinterpreted as the study of certain endofunctors and derived concepts, primarily in the category of *represented spaces*. We interpret this key idea of synthetic descriptive set theory as relativizing topological notions by (higher-type) oracles. In this approach, an oracle is considered to be a functor that allows us to *change the way we access spaces.*

The third perspective of oracle is the one that we promote in this talk. In this third perspective, we consider an oracle to be an *operation on truth-values* that may cause a transformation of one world into another. One might say that this is based on the idea that there is a correspondence between *computations using oracles* and *proofs using transcendental axioms*". Such an idea is used as a very standard technique in, for example, classical reverse mathematics. Our approach is similar, but with a newer perspective that deals more directly with operations on truth-values. More explicitly, it is formulated using topos-theoretic notions such as Lawvere-Tierney topology, which is a kind of generalization of Grothendieck topology to an arbitrary topos.

In this talk, we clarify the connection between these three perspectives of oracle. In this way, we attempt to bridge the gap between computability theory, synthetic description set theory, and effective topos theory.

3.4 Topology reflected in the enumeration degrees

Joseph S. Miller (University of Wisconsin – Madison, US)

This was an expository talk on connections between pure topology and the enumerations degrees.

The continuous degrees were introduced by the speaker (2004) to measure the computability-theoretic content of elements of computable metric spaces. They properly extend the Turing degrees. All known constructions of nontotal (i.e., non-Turing) continuous degrees involve a nontrivial topological component. Indeed, the fact that there is a nontotal continuous degrees in every upper cone is equivalent to the fact that the Hilbert cube is not a countable union of (subspaces homeomorphic to) subspaces of Cantor space.

The continuous degrees naturally embed in the enumeration degrees, where there are more connections to topology. Many of these were described by Kihara and Pauly, who assigned enumeration degrees to the points of any second countable T_0 topological space. This work was continued by Kihara, Ng, and Pauly. Among their many results, they characterized the cototal degrees as the degrees of points in $(\omega_{\mathrm{cof}})^\omega$, where ω_{cof} is ω with the cofinite topology.

It is not know if every continuous degree is graph cototal, but the work above allows us to translate this into a topological question. In particular, the following are equivalent:

- There is a continuous degree that is not graph cototal in every upper cone.
- The Hilbert cube is not a countable union of (subspaces homeomorphic to) subspaces of $(\omega_{\mathrm{cof}})^\omega$.

3.5 CoPolish Spaces and Effectively Hausdorff Spaces

Matthias Schröder (TU Darmstadt, DE)

This talk presented two classes of topological spaces which play a big role in Computable Analysis, namely CoPolish spaces and effectively Hausdorff spaces. CoPolish spaces are a generalisation of locally compact spaces in the realm of QCB-spaces. They form exactly the class of topological spaces which admit Simple Complexity, i.e. the measurement of Time Complexity in terms of a discrete parameter on the input and the desired output precision. Moreover, we show that there exists a universal CoPolish space, which is a CoPolish space into which every other CoPolish space embeds as a closed subspace. Effectively Hausdorff spaces generalise computable metric spaces and yield a better effectivisation of Hausdorffness than the current notion of a computable Hausdorff space. Unlike computable Hausdorff spaces they admit computability of a certain form of overt compact choice. Moreover, we characterise computability of multivalued functions from computable metric spaces to effectively Hausdorff spaces.

3.6 Tutorial: Quantitative Coding and Complexity Theory of Compact Metric Spaces

Martin Ziegler (KAIST – Daejeon, KR)

Joint work of Donghyun Lim, Martin Ziegler
Main reference Donghyun Lim, Martin Ziegler: "Quantitative Coding and Complexity Theory of Compact Metric Spaces", in Proc. of the Beyond the Horizon of Computability – 16th Conference on Computability in Europe, CiE 2020, Fisciano, Italy, June 29 – July 3, 2020, Proceedings, Lecture Notes in Computer Science, Vol. 12098, pp. 205–214, Springer, 2020.
URL https://doi.org/10.1007/978-3-030-51466-2_18
URL http://youtu.be/QGTkZfUzhrI

Specifying a computational problem includes fixing encodings for input and output: encoding graphs as adjacency matrices, characters as integers, integers as bit strings, and vice versa. For such discrete data, the actual encoding is usually straightforward and/or complexity-theoretically inessential (up to linear or polynomial time, say). Concerning continuous data, already real numbers naturally suggest various encodings (formalized as historically so-called *representations*) with very different algorithmic properties, ranging from the computably "unreasonable" binary expansion [doi:10.1112/plms/s2-43.6.544] via qualitatively to polynomially and even linearly complexity-theoretically "reasonable" signed-digit expansion. But how to distinguish between un/suitable encodings of other spaces common in Calculus and Numerics, such as Sobolev?

With respect to qualitative computability over topological spaces, *admissibility* had been identified [doi:10.1016/0304-3975(85)90208-7] as a crucial criterion for a representation over the Cantor space of infinite binary sequences to be 'reasonable': It requires the representation to be (sequentially) continuous, and to be maximal with respect to (sequentially) continuous reduction [doi:10.1007/11780342_48]. Such representations are guaranteed to exist for a large class of spaces. And for (precisely) these does the sometimes so-called *Main Theorem* hold: which characterizes continuity of functions by the continuity of mappings translating codes, so-called *realizers*.

qualitative	computability	topology	(uniform) continuity	compactness	equilogical
quantitative	complexity	metric	modulus of continuity	entropy	ultrametric

Following this ”dictionary”, we refine qualitative computability over topological spaces to quantitative complexity over metric spaces, by developing the theory of *polynomially* and of *linearly admissible* representations. Informally speaking, these are ‘optimally’ continuous, namely linearly/polynomially relative to the space’s entropy; and maximal with respect to relative linearly/polynomially continuous reductions defined below. A large class of spaces is shown to admit a quantitatively admissible representation, including a generalization of the signed-digit encoding; and these exhibit a quantitative strengthening of the qualitative *Main Theorem*, namely now characterizing quantitative continuity of functions by quantitative continuity of realizers. Our quantitative admissibility thus provides the desired criterion for complexity-theoretically ‘reasonable’ encodings.

3.7 Borel combinatorics fail in HYP

Linda Westrick (Pennsylvania State University – University Park, US)

Joint work of Linda Westrick, Henry Towsner, Rose Weisshaar
Main reference H. Towsner, R. Weisshaar, L. Westrick: “Borel combinatorics fail in HYP”, To appear in the Journal of Mathematical Logic
URL http://arxiv.org/abs/2106.13330

Of the principles just slightly weaker than ATR, the most well-known are the theories of hyperarithmetic analysis (THA). By definition, such principles hold in HYP. Motivated by the question of whether the Borel Dual Ramsey Theorem is a THA, we consider several theorems involving Borel sets and ask whether they hold in HYP. To make sense of Borel sets without ATR, we formalize the theorems using completely determined Borel sets. We characterize the completely determined Borel subsets of HYP as precisely the sets of reals which are Δ^1_1 in $L_{\omega_1^{ck}}$. Using this, we show that in HYP, Borel sets behave quite differently than in reality. For example, in HYP, the Borel dual Ramsey theorem fails, every n-regular Borel acyclic graph has a Borel 2-coloring, and the prisoners have a Borel winning strategy in the infinite prisoner hat game. Thus the negations of these statements are not THA.

Funding: NSF grant DMS-1854107

4 Short Talks

4.1 Continuity and Computability

Vasco Brattka (Bundeswehr University Munich, DE)

We discuss relations between continuity and computability. From the folklore fact that LPO is the weakest discontinuous function with respect to the topological version of Weihrauch reducibility, we deduce a characterization of discontinuity as the class of those functions whose parallelization realizes every Turing jump on some cone. We also show that the parallelization of a function being computably reducible to the identity is a condition that sits in between computability and computability with respect to the halting problem and we raise the question whether this condition can be separated from computability.

4.2 When does Wadge meet Tang and Pequignot?

Riccardo Camerlo (University of Genova, IT)

Wadge hierarchy on topological spaces has been introduced by W.W. Wadge to compare subsets according to their complexity. A variation of this hierarchy has been introduced by A. Tang for the Scott domain, and more recently generalized by Y. Pequignot to every T_0 second countable spaces. I discuss the question of when these two hierarchies coincide, presenting what is known and which problems are still open.

4.3 Algorithmic Learning of Structures

Ekaterina Fokina (TU Wien, AT)

Joint work of Nikolay Bazhenov, Ekaterina Fokina, Dino Rossegger, Luca San Mauro, Alexandra Soskova, Mariya Soskova, Stefan Vatev
Main reference Nikolay Bazhenov, Ekaterina B. Fokina, Luca San Mauro: "Learning families of algebraic structures from informant", Inf. Comput., Vol. 275, p. 104590, 2020.
URL https://doi.org/10.1016/j.ic.2020.104590

In this talk we summarize some of the recent results and mention several open questions on algorithmic learning of structures. We combine the ideas of computable structure theory and algorithmic learning theory (inductive inference) to study the question of what classes of structures are learnable under various learning criteria and restrictions. A class of structures is said to be learnable if there is a learner (a function) that correctly learns each structure from the class. This means, that the learner observes larger and larger finite pieces of the structure and makes guesses about which structure it is observing. After finitely many steps the learner must converge to a correct hypothesis. In general, we do not care about the complexity of the learner, but sometimes we do.

In the talk we explain the main result of [1] which gives a syntactic characterization of explanatory learnability of classes of structures from informant and also gives an upper bound on the complexity of the learner. We then mention a similar result for the notion of learning of structures from text (work in progress [3]). Furthermore, we mention results from [2] that reveal an interesting relation between explanatory learning of structures from informant and descriptive set theory. We wonder what other learning criteria can be characterized syntactically and/or in terms of equivalence relations.

References

1 N. Bazhenov, E. Fokina, and L. San Mauro. Learning families of algebraic structures from informant, Information and Computation, 275, 2020.

2 N. Bazhenov, V. Cipriani, and L. San Mauro. Learning structures and Borel equivalence relations, preprint 2021.

3 N. Bazhenov, E. Fokina, D. Rossegger, A. Soskova, M. Soskova, S. Vatev. Vaught's theorem for the Scott topology and a syntactic characterization for learning, work in progress.

4.4 Refuting Selman's theorem in the hyperenumeration degrees

Jun Le Goh (University of Wisconsin – Madison, US)

We report on discussions by the participants in the #e-degrees Slack channel, specifically on hyperenumeration reducibility $\leq_{he}$ (see M. Soskova's abstract in the present report).

We came up with a possible strategy for refuting the analog of Selman's theorem for $\leq_{he}$, i.e., for constructing sets $A \not\leq_{he} B$ such that whenever $B \leq_{he} C \oplus C^c$, we have $A \leq_{he} C \oplus C^c$. The idea is to construct a Δ^1_1-pointed tree $T \subseteq \omega^{<\omega}$ with no dead ends such that $T^c \not\leq_{he} T$. It then suffices to consider $A = T^c$ and $B = T$: If $T \leq_{he} C \oplus C^c$, then T is $\Pi^1_1(C)$, so T has a path P which is $\Pi^1_1(C)$. Since T is Δ^1_1-pointed, it is $\Delta^1_1(P)$, hence $\Delta^1_1(C)$. We conclude that T^c is $\Pi^1_1(C)$, i.e., $T^c \leq_{he} C \oplus C^c$ as desired.

Josiah Jacobsen-Grocott has made progress on implementing the above strategy.

4.5 A characterization of Π^0_3-completeness

Vassilios Gregoriades (National Technical University of Athens, GR)

Main reference Vassilios Gregoriades: "Intersections of ℓ^q spaces in the Borel hierarchy", Journal of Mathematical Analysis and Applications, Vol. 498(1), p. 124922, 2021.
URL https://doi.org/10.1016/j.jmaa.2021.124922

Given $0 < a < q$, the intersection of all spaces ℓ^p for $p > a$ is a Π^0_3-complete subset of ℓ^q. This answers a question by Nestoridis [1]. The proof motivates a characterization of Π^0_3-completeness of sets in Polish spaces.

References

1 Vassili Nestoridis. A project about chains of spaces, regarding topological and algebraic genericity and spaceability. https://arxiv.org/abs/2005.01023, 2020.

4.6 There is no Good Notion of Quasi-Polish Convergence Spaces

Reinhold Heckmann (AbsInt – Saarbrücken, DE)

We looked for a full subcategory QP-CONV of the category CONV of convergence spaces that is closed under countable product, equalizers, and exponentials and whose topological spaces are exactly the quasi-Polish spaces. A natural candidate is QPE, the least full subcategory of CONV that contains the Sierpinski space and is closed under isomorphism, countable products, equalizers, and exponentials. Yet QPE contains the subspace Q of R, which is not quasi-Polish, and this implies that there is no category QP-CONV with the desired properties. Nevertheless, we think that QPE is an interesting category for further study.

4.7 Descriptive complexity on represented spaces

Mathieu Hoyrup (Loria, Inria – Nancy, FR)

Our goal is to better understand the relationship between two notions of descriptive complexity for subsets of a represented space, one using the topology, the other one using the representation.

4.8 Regularity properties, determinacy, and Solovay models

Daisuke Ikegami (Shibaura Institute of Technology – Tokyo, JP)

Main reference Daisuke Ikegami: "I-regularity, determinacy, and ∞-Borel sets of reals", CoRR, Vol. 2108.06632, 2021.
URL https://arxiv.org/abs/2108.06632

Regularity properties for sets of reals have been extensively studied since the early 20th century. A set of reals with a regularity property can be approximated by simple sets (such as Borel sets) modulo some small sets. Typical examples of regularity properties are Lebesgue measurability, the Baire property, the perfect set property, and Ramseyness.

For each σ-ideal I on the Baire space, Khomskii introduced a regularity property called I-regularity, and developed a general theory of I-regularity. Khomskii asked if strong axioms of determinacy (such as the Axiom of Determinacy) imply every set of reals is I-regular for any I such that the associated preorder P_I is proper.

In this talk, we discuss some results and questions concerning I-regularity, determinacy of infinite games, and Solovay models.

4.9 Resource-bounded effective dimension and the point-to-set principle

Elvira Mayordomo (University of Zaragoza, ES)

Joint work of Jack H. Lutz, Neil Lutz, Elvira Mayordomo
Main reference Jack H. Lutz, Neil Lutz, Elvira Mayordomo: "Dimension and the Structure of Complexity Classes", CoRR, Vol. abs/2109.05956, 2021.
URL https://arxiv.org/abs/2109.05956

In this short talk I review the recent results on the point to set principle for resource-bounded dimensions [1] stating that if Δ is a resource bound more general than Γ then Δ-dimension can be characterized in terms of Γ-dimension relativized to oracles dependent on Δ. I also include a few questions on the optimality and complexity of the corresponding oracles for different resource-bounds and gauge functions.

References

1 Jack H. Lutz, Neil Lutz, and Elvira Mayordomo. *Dimension and the Structure of Complexity Classes*. Arxiv arXiv:2109.05956, 2021

4.10 Computable presentations in topology

Alexander Melnikov

Joint work of Harrison-Trainor, Melnikov, and Ng
Main reference Matthew Harrison-Trainor, Alexander G. Melnikov, Keng Meng Ng: "Computability of Polish Spaces up to Homeomorphism", J. Symb. Log., Vol. 85(4), pp. 1664–1686, 2020.
URL https://doi.org/10.1017/jsl.2020.67

Computable presentations in effective algebra have been studied extensively for over 60 years. Classical results of Turing, Novikov, Boone, Feiner, and Khisamiev (in chronological order) illustrate that the standard notions of computable presentability for discrete algebraic structures differ in the standard classes such as semigroups, finitely presented groups, Boolean algebras, and abelian groups, respectively. Similar results are well-known for other common classes of structures such as, e.g., linear orders.

Similarly, investigations into the algorithmic content of abstract topological structures can be traced back to Maltcev in the 1960s. There are many definitions in the literature of what it means for a Polishable space to be computably presentable. These include computable complete metrization, computable topological presentation, and an effectively compact (completely metrized) presentation. These three notions seem to be the most commonly used notions throughout the literature. Nonetheless, in contrast with effective algebra, until very recently it was not known whether these notions of computable presentability differed (up to homeomorphism). We discuss several very recent works in which, using classical and advanced modern techniques, these notions have been separated in several common classed of compact spaces.

4.11 Topological spaces of countable structures

Russell G. Miller (CUNY Queens College – Flushing, US)

Main reference Russell Miller: "Isomorphism and classification for countable structures", Comput., Vol. 8(2), pp. 99–117, 2019.
URL https://doi.org/10.3233/COM-180095

We describe a natural way to view a collection of (isomorphism types of) countable structures as a topological space. The space is T_0 provided that the structures all have distinct existential theories: sometimes it is useful to adjoin definable predicates to the signature to achieve this. The notion of a (boldface) *Turing-computable embedding*, developed by Knight et al., is simply a continuous injective map from one such space to another.

We consider the specific example of algebraic field extensions of the rational numbers. Here the topology turns out to be that of a spectral space, meaning that (by a theorem of Hochster) there is some commutative ring R whose spectrum of prime ideals, under the Zariski topology, is homeomorphic to this space and thus can serve as a classification of these fields.

The main point of this talk is to raise questions. First, what is this ring R whose spectrum classifies the algebraic fields? (Well-known polynomial rings and other obvious guesses at R have all so far turned out to be wrong.) Second, the procedure above gives rise to many more computable topological spaces, some of which are spectral and others not. In what ways do the separate, well-developed disciplines of computable topology and computable structure theory interact here, and how can we use the interaction to develop these disciplines further and to link them together?

References

1 M. Hochster, Prime Ideal Structure in Commutative Rings, *Transactions of the American Mathematical Society* 142 (1969), 43–60.

2 J.F. Knight, S. Miller, & M. Vanden Boom, Turing Computable Embeddings, *Journal of Symbolic Logic* 72 3 (2007), 901–918.

3 R. Miller, Isomorphism and Classification for Countable Structures, *Computability* 8 (2019) 2, 99–117.

4.12 Computable Endofunctors, Markov-computability and Relativization

Arno Pauly (Swansea University, GB)

Joint work of Arno Pauly, Matthew de Brecht

The notion of a computable endofunctor was introduced by Pauly and de Brecht [4] in order to give a somewhat unified and principled approach to develop descriptive set theory for arbitrary represented spaces. The technology was used in [3] to obtain a computable version of the Jayne Rogers theorem, and in [1] to effectivize the property of being a Noetherian topological space (in a way that revealed it to be a higher-order analogue of both compactness and overtness).

(As pointed out by Neumann in [2], the terminology “locally computable endofunctor” would be more appropriate.)

If the endofunctors generating the usual notions of interest for descriptive set theory had left adjoints, we could use abstract category theory to draw conclusions in a way that generalizes retopologization arguments. Alas, adjoints seem to be rare over the category of represented spaces and continuous functions. If instead, we take Markov-computable maps as morphisms, adjoints become abundant. A challenging question now is whether we can incorporate relativization arguments into the category-theoretic framework in a way that links the Markov-computable setting with the usual one.

References

1 Matthew de Brecht and Arno Pauly. Noetherian Quasi-Polish spaces. In Valentin Goranko and Mads Dam, editors, *26th EACSL Annual Conference on Computer Science Logic (CSL 2017)*, volume 82 of *LIPIcs*, pages 16:1–16:17. Schloss Dagstuhl, 2017.

2 Eike Neumann. *Universal Envelopes of Discontinuous Functions.* PhD thesis, Aston University, 2018.

3 Arno Pauly and Matthew de Brecht. Non-deterministic computation and the Jayne Rogers theorem. *Electronic Proceedings in Theoretical Computer Science*, 143, 2014. DCM 2012.

4 Arno Pauly and Matthew de Brecht. Descriptive set theory in the category of represented spaces. In *30th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 438–449, 2015.

4.13 Effective overtness of generalised Cantor spaces

Philipp Schlicht (University of Bristol, GB)

Joint work of Arno Pauly, Philipp Schlicht

The generalised Cantor space 2^κ for an uncountable regular cardinal κ is the space of binary sequences of length κ. One can translate the notion of representable space to this context, since 2^κ comes with a natural notion of computability with time bound κ. While 2^κ need not have a κ-computable dense subset of size κ, we discuss the weaker notion of effective overtness for these spaces.

4.14 Effective embedding and interpretations

Alexandra A. Soskova (University of Sofia, BG)

Joint work of Rachael Alvir, Wesley Calvert, G. Goodman, Valentina S. Harizanov, Julia F. Knight, R. Miller, Andrei S. Morozov, Alexandra A. Soskova, Stefan Vatev, Rose Weisshaar

Main reference Julia F. Knight, Alexandra A. Soskova, Stefan V. Vatev: “Coding in graphs and linear Orderings”, J. Symb. Log., Vol. 85(2), pp. 673–690, 2020.

URL https://doi.org/10.1017/jsl.2019.91

Friedman and Stanley [2] introduced Borel embeddings as a way of comparing classification problems for different classes of structures. A Borel embedding for a class K in a class K' represents a uniform procedure for coding structures from K in structures from K'. Many Borel embeddings are actually Turing computable. When a structure $\mathcal{A}$ is coded in a structure $\mathcal{B}$, effective decoding is represented by a Medvedev reduction of $\mathcal{A}$ to $\mathcal{B}$. Harrison-Trainor, Melnikov, Miller, and Montalbán [3] defined a notion of effective interpretation of $\mathcal{A}$ in $\mathcal{B}$ and proved that this is equivalent with the existing of computable functor.

The class of undirected graphs and the class of linear orderings both lie “on top” under Turing computable embeddings. The standard Turing computable embeddings of structures in undirected graphs come with uniform effective interpretations. We [4] give examples of graphs that are not Medvedev reducible to any linear ordering, or to the jump of any linear ordering. Any graph can be interpreted in a linear ordering using computable Σ_3 formulas. Friedman and Stanley gave a Turing computable embedding L of directed graphs in linear orderings. We show that there does not exist a Borel interpretation, i.e. there are no $L_{\omega_1\omega}$ formulas that, for all graphs G, interpret G in $L(G)$. Our conjecture is: *For any Turing computable embedding Θ of graphs in orderings, there do not exist $L_{\omega_1\omega}$ formulas that, for all graphs G, define an interpretation of G in $\Theta(G)$.*

We [1] succeed to find an effective interpretation of a field in its Heisenberg group without parameters, generalising an old result of Maltsev, who gave a definition of a field in its Heisenberg group with a pair of parameters. We could define an algebraically closed field C in the group $SL_2(C)$ using finitary existential formulas with a pair of parameters. The question is: *Are there formulas that, for all algebraically closed fields C of characteristic 0, define an effective interpretation of C in $SL_2(C)$? Are there existential formulas that serve?*

The work was partially supported by BNSF, KP-06-Austria-04/06.08.2019, FNI-SU 80-10-136/26.03.2021.

References

1 R. Alvir, W. Calvert, G. Goodman, V. Harizanov, J. Knight, A. Morozov, R. Miller, A. Soskova, and R. Weisshaar. *Interpreting a field in its Heisenberg group*. J. Symbolic Logic, 2021

2 H. Friedman and L. Stanley. *Borel reducibility theory for classes of countable structures*. J. Symbolic Logic, 54, 894–914, 1989

3 M. Harrison-Trainor, A. Melnikov, R. Miller, and A. Montalbán. *Computable functors and effective interpretability*. J. Symbolic Logic, 82, 77–97, 2017

4 J. Knight, A. Soskova, and S. Vatev. *Coding in graphs and linear orderings*. J. Symbolic Logic, 85 (2), 673– 690, 2020

4.15 The hyper enumeration degrees

Mariya I. Soskova (University of Wisconsin – Madison, US)

In this talk I outlined some main aspects of the enumeration degrees and their relationship to the Turing degrees, so that I can draw a parallel between the enumeration degrees and the hyperenumeration degrees. We say that A is hyper enumeration reducible to B if there is a c.e. set W such that $x \in A$ if and only if for every $f \in \omega^{<\omega}$ there is some n and some finite set D such that $(f \upharpoonright n, x, D) \in W$ and $D \subseteq B$. This notion was introduced and studied by Sanchis [1], who showed that in many ways hyper enumeration reducibility relates to hyperarithmetic reducibility in the same way that enumeration reducibility relates to Turing reducibility.

I focused on two open questions:

1. Do we have an analog of Selman's theorem for hyper-enumeration reducibility: Is it true that $A \leq_{he} B$ if and only if for every X if B is $\Pi^1_1(X)$ then A is $\Pi^1_1(X)$?
2. Is there a way to stratify hyper-enumeration reducibility: We know that $A \leq_h B$ if and only if $A \leq_T B^{(\alpha)}$ for some B-computable ordinal α. Do we have some analogous result for hyper enumeration reducibility, perhaps using the skip instead of the jump?

References

1 Luis Sanchis. *Hyperenumeration reducibility*. Notre Dame Journal of Formal Logic, Volume XIX, Number 3, July 1978.

5 Working groups

5.1 Computable categoricity of Polish spaces

Nikolay Bazhenov (Sobolev Institute of Mathematics – Novosibirsk, RU), Ivan Georgiev (Sofia University "St. Kliment Ohridski", BG), Jun Le Goh (University of Wisconsin – Madison, US), Vassilios Gregoriades (National Technical University of Athens, GR), Mathieu Hoyrup (LORIA & INRIA Nancy, FR), Iskander Shagitovich Kalimullin (Kazan Federal University, RU), Steffen Lempp (University of Wisconsin – Madison, US), Alexander Melnikov (Victoria University – Wellington, NZ), Russell G. Miller (CUNY Queens College – Flushing, US), Eike Neumann (MPI für Informatik – Saarbrücken, DE), Keng Meng Ng (Nanyang TU – Singapore, SG), Arno Pauly (Swansea University, GB), Alexandra A. Soskova (University of Sofia, BG), and Daniel Turetsky (Victoria University – Wellington, NZ)

This working group followed talks by Ng and Melnikov.

Galicki, Melnikov and Ng have studied categoricity of Polish spaces. A Polish space X is computably categorical if all computable presentations of X are computably homeomorphic. More generally, a set A is the degree of categoricity of a space X if A is the minimal oracle such that all computable copies of X are A-computably homeomorphic.

They proved the following results, among others:

- The space of natural numbers N has degree 0',
- The Cantor space has degree 0',
- The Baire space is not 0'-computably categorical,
- The unit interval [0,1] has degree 0".

They also have a sketch proof that no compact Polish space is computably categorical.

In this group we have discussed the case of compact Polish spaces, trying to complete the proof, and obtained that X a compact space is not computably categorical in the following cases:

- If X has a computable copy containing a nowhere dense non-empty Pi01-set,
- If X has a computable copy such that N computably embeds in the isolated points of X.

The arguments also extend to sigma-compact spaces.

It remains open whether there is computably categorical compact Polish space, more generally if there is a computably categorical Polish space.

5.2 AE-theory of enumeration degree structures

Steffen Lempp (University of Wisconsin – Madison, US), Jun Le Goh (University of Wisconsin – Madison, US), Keng Meng Ng (Nanyang TU – Singapore, SG), and Mariya I. Soskova (University of Wisconsin – Madison, US)

This is to follow up on the short talk I gave on progress toward deciding the AE-fragments of the first-order theories of two degree structures, the global enumeration degrees and the local Σ^0_2-enumeration degrees. For the global structure, significant progress was already reported on from the paper [1]. Plans are in place to extend our results toward a full solution. For the local structure, significant progress has been made during and since the workshop by the four of us: We now have a working conjecture for 1-point extensions of antichains, which we hope to check and write up carefully over the next few months, whereas at the time of the workshop, we only had an analysis of the very special case where the antichain has size 3!

References

1 Lempp, Steffen; Soskova, Mariya I.; and Slaman, Theodore A., *Fragments of the theory of the enumeration degrees*, Advances in Mathematics, Vol. 383, 2021, paper 107686, 39 pages.

6 Open problems

6.1 Questions on left-c.e. reals

Iskander Shagitovich Kalimullin (Kazan Federal University, RU)

Joint work of Iskander Shagitovich Kalimullin, Marat Kh. Faizrahmanov
Main reference Marat Kh. Faizrahmanov, Iskander Sh. Kalimullin: "Limitwise monotonic sets of reals", Math. Log. Q., Vol. 61(3), pp. 224–229, 2015.
URL https://doi.org/10.1002/malq.201400015

The talk is devoted to possible applications and problems in computable topology related to the paper [1]. In this paper the authors found a countable subset of the reals which is not left-c.e. but is non-uniformly left-c.e. relative to any non-computable oracle. This has an applications in computable structure theory, but it is interesting also to know what effects we have studying uncountable subsets of the reals.

References

1 Marat Kh. Faizrahmanov, Iskander Sh. Kalimullin, *Limitwise monotonic sets of reals*. Math. Log. Q. 61(3): 224-229 (2015)

6.2 Which Compact Metric spaces do/don't admit polynomially admissible representations?

Martin Ziegler (KAIST – Daejeon, KR)

Donghyun Lim and Martin Ziegler [arXiv:2002.04005v5] have quantitatively refined the qualitative notion of "admissible representation" [Kreitz&Weihrauch'85]; see the tutorial in this very seminar.

Many spaces admit polynomially admissible representations, the reals even a linearly admissible (namely the signed-digit) representations.

We wonder about compact metric spaces that provably do NOT admit a polynomially admissible representatios; and perhaps even a characterization of those that do.

Participants

- Vasco Brattka
Bundeswehr University
Munich, DE
- Riccardo Camerlo
University of Genova, IT
- Raphael Carroy
University of Torino, IT
- Jacques Duparc
University of Lausanne, CH
- Ivan Georgiev
Sofia University "St. Kliment
Ohridski", BG
- Jun Le Goh
University of Wisconsin –
Madison, US
- Vassilios Gregoriades
National Technical University of
Athens, GR
- Mathieu Hoyrup
LORIA & INRIA Nancy, FR
- Steffen Lempp
University of Wisconsin –
Madison, US
- Elvira Mayordomo
University of Zaragoza, ES
- Russell G. Miller
CUNY Queens College –
Flushing, US
- Eike Neumann
MPI für Informatik –
Saarbrücken, DE
- Adam Ó Conghaile
University of Cambridge, GB
- Arno Pauly
Swansea University, GB
- Marcin Sabok
McGill University –
Montreal, CA
- Matthias Schröder
TU Darmstadt, DE
- Alexandra A. Soskova
University of Sofia, BG
- Martin Ziegler
KAIST – Daejeon, KR

Remote Participants

- Alessandro Andretta
University of Torino, IT
- Nikolay Bazhenov
Sobolev Institute of Mathematics
– Novosibirsk, RU
- Ruiyuan (Ronnie) Chen
McGill University –
Montreal, CA
- Matthew de Brecht
Kyoto University, JP
- Damir D. Dzhafarov
University of Connecticut –
Storrs, US
- Olivier Finkel
University of Paris, FR
- Ekaterina Fokina
TU Wien, AT
- Johanna N. Y. Franklin
Hofstra University –
Hempstead, US
- Reinhold Heckmann
AbsInt – Saarbrücken, DE
- Daisuke Ikegami
Shibaura Institute of Technology
– Tokyo, JP
- Corrie Ingall
University of Connecticut –
Storrs, US
- Josiah Jacobsen-Grocott
University of Wisconsin –
Madison, US
- Iskander Shagitovich
Kalimullin
Kazan Federal University, RU
- Takayuki Kihara
Nagoya University, JP
- Margarita Korovina
Universität Trier, DE

- Davorin Lesnik
University of Ljubljana, SI
- Neil Lutz
Swarthmore College, US
- Alexander Melnikov
Victoria University – Wellington, NZ
- Joseph S. Miller
University of Wisconsin – Madison, US
- Luca Motto Ros
University of Torino, IT
- Takako Nemoto
Hiroshima Institute of Technology, JP
- Keng Meng Ng
Nanyang TU – Singapore, SG
- André Otfrid Nies
University of Auckland, NZ
- Alexey Ostrovsky
SUAI – Sankt-Peterburg, RU
- Jan Reimann
Pennsylvania State University – University Park, US
- Philipp Schlicht
University of Bristol, GB
- Victor Selivanov
A. P. Ershov Institute – Novosibirsk, RU
- Svetlana Selivanova
KAIST – Daejeon, KR
- Alex Simpson
University of Ljubljana, SI
- Theodore A. Slaman
University of California – Berkeley, US
- Mariya I. Soskova
University of Wisconsin – Madison, US
- Daniel Turetsky
Victoria University – Wellington, NZ
- Linda Westrick
Pennsylvania State University – University Park, US

Report from Dagstuhl Seminar 21462

Foundations of Persistent Programming

Edited by

Hans-J. Boehm[1], Ori Lahav[2], and Azalea Raad[3]

1 **Google – Mountain View, US,** boehm@acm.org
2 **Tel Aviv University, IL,** orilahav@tau.ac.il
3 **Imperial College London, GB,** azalea.raad@imperial.ac.uk

Abstract

Although early electronic computers commonly had persistent core memory that retained its contents with power off, modern computers generally do not. DRAM loses its contents when power is lost. However, DRAM has been difficult to scale to smaller feature sizes and larger capacities, making it costly to build balanced systems with sufficient amounts of directly accessible memory. Commonly proposed replacements, including Intel's Optane product, are once again persistent. It is however unclear, and probably unlikely, that the fastest levels of the memory hierarchy will be able to adopt such technology. No such non-volatile (NVM) technology has yet taken over, but there remains a strong economic incentive to move hardware in this direction, and it would be disappointing if we continued to be constrained by the current DRAM scaling.

Since current computer systems often invest great effort, in the form of software complexity, power, and computation time, to "persist" data from DRAM by rearranging and copying it to persistent storage, like magnetic disks or flash memory, it is natural and important to ask whether we can leverage persistence of part of primary memory to avoid this overhead. Such efforts are complicated by the fact that real systems are likely to remain only partially persistent; some memory components, like processor caches and device registers. may remain volatile.

This seminar focused on various aspects of programming for such persistent memory systems, ranging from programming models for reasoning about and formally verifying programs that leverage persistence, to techniques for converting existing multithreaded programs (particularly, lock-free ones) to corresponding programs that also directly persist their state in NVM. We explored relationships between this problem and prior work on concurrent programming models.

Seminar November 14–17, 2021 – http://www.dagstuhl.de/21462

2012 ACM Subject Classification Theory of computation → Concurrency; Hardware → Non-volatile memory; Theory of computation → Program semantics

Keywords and phrases concurrency; non-volatile-memory; persistency; semantics; weak memory models

Digital Object Identifier 10.4230/DagRep.11.10.94

Edited in cooperation with Michalis Kokologiannakis

Foundations of Persistent Programming, *Dagstuhl Reports*, Vol. 11, Issue 10, pp. 94–110
Editors: Hans-J. Boehm, Ori Lahav, and Azalea Raad
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Hans-J. Boehm (Google – Mountain View, US, boehm@acm.org)
Ori Lahav (Tel Aviv University, IL, orilahav@tau.ac.il)
Azalea Raad (Imperial College London, GB, azalea.raad@imperial.ac.uk)

We brought together 15 in-person attendees at Schloss Dagstuhl, with a roughly equal number of remote attendees. Remote attendance was challenging, particularly for attendees from very different time-zones. It nonetheless provided the opportunity for us to hear from a wider selection of participants.

We decided up-front not to try to cater the schedule to remote participation. Given the number of time zones covered by the participants, we continue to believe that, although it clearly had adverse impacts, it was the right decision.

We had a number of remote presentations that included interactions with the speakers. Otherwise the discussion tended to happen mostly among the in-person participants, in spite of the excellent AV systems at Dagstuhl. Many remote participants were limited in attendance due to time-zone issues.

Our area is perhaps unique, in that it includes deep theoretical work, but is also very dependent on technological developments. Accordingly, the participants of the seminar were from a spectrum of topics ranging from theory of distributed systems to hardware specification and design. The seminar gave many of us the opportunity to catch up on both theory and practice, including input from some participants with more direct insights into industrial developments.

We began the seminar by reviewing some of the underlying assumptions that were made by prior work in this field, often without certainty about their correctness. We were actually able to get much more shared clarity on a few of these as a result of audience discussion during the seminar. Some of us learned that non-volatile caches are being publicly discussed by Intel, and that there also is similar agreement that memory encryption, to restore volatility when needed, is desirable. We also learned that writes to the same cache line are not just believed by software researchers to reach memory in the correct order, but at least one hardware vendor also agrees. Though this last fact is rather obscure, it is important for some NVM algorithms, and not normally reflected in hardware manuals.

The rest of the seminar consisted of talks and group discussions. Three talks were longer overview talks on different aspects (Michael Scott on buffered persistency, Parosh Aziz Abdulla on verification, and Erez Petrank on persistent lock-free data structures). We did not feel the need for smaller break-out sessions, since the in-person group was quite small, largely due to our timing with respect to Covid waves. Much of the benefit here appears to have been in listening to discussions that often were either significantly more theoretical or significantly more practical than our own research.

NVM programming is both complicated by, and often synergistic with concurrent programming. Much of our focus was on the interaction between the two. Due to these close interactions, we asked several speakers to talk about concurrency issues that seemed particularly relevant (e.g., Peter Sewell on Armv8-A virtual memory model, Mark Batty on novel solution to the "thin air problem", and Paul McKenney on weak memory schemes used in the Linux kernel).

Several talks raised questions on the foundations of the field, such as what are the hardware-supplied programming models, or what it means for a persistent program to be correct. It is entirely possible that most future NVM programmers will be more concerned with something like the persistent transactions discussed in Michael Bond's talk. However, given the relative immaturity of the area and foundational uncertainty, the emphasis seemed appropriate.

2 Table of Contents

Executive Summary

Hans-J. Boehm, Ori Lahav, and Azalea Raad . . . 95

Overview of Talks

Recoverable Self-Implementations of Primitive Operations
Hagit Attiya . . . 99

Consistency and Persistency: Challenges and Opportunities in Program Verification
Parosh Aziz Abdulla . . . 99

Isolating the Thin Air Problem: Semantic Dependency for Optimised Concurrency
Mark Batty . . . 100

Persistent Transactions: Desirable Semantics and Efficient Designs
Michael D. Bond . . . 100

Model Checking Persistent Memory Programs
Brian Demsky . . . 101

View-Based Owicki–Gries Reasoning for Persistent x86-TSO
Brijesh Dongol . . . 101

General Constructions for Non-Volatile Memory
Michal Friedman . . . 102

Formal Foundations for Intermittent Computing
Limin Jia . . . 102

Revamping Hardware Persistency Models: View-Based and Axiomatic Persistency Models for Intel-X86 and Armv8
Jeehoon Kang . . . 103

Abstraction for Crash-Resilient Objects
Artem Khyzha . . . 103

Taming x86-TSO Persistency
Artem Khyzha . . . 104

PerSeVerE: Persistency Semantics for Verification under Ext4
Michalis Kokologiannakis . . . 104

Weak Memory Schemes Used in the Linux Kernel
Paul McKenney . . . 105

Hazard Pointer Synchronous Reclamation
Maged M. Michael . . . 105

Persistent Lock-Free Data Structures for Non-Volatile Memory
Erez Petrank . . . 105

TSOPER: Efficient Coherence-Based Strict Persistency
Konstantinos Sagonas . . . 106

The Case for Buffered Persistence
Michael Scott . . . 106

A Taste of Armv8-A Relaxed Virtual Memory
Peter Sewell . . . 107

The ISA Semantics / Concurrency Model Interface
Peter Sewell 107

Non-Temporal Stores and their Semantics
Viktor Vafeiadis 108

Architectural Support for Persistent Programming
William Wang 109

Modularizing Verification of Durable Opacity
Heike Wehrheim 109

Participants 110

Remote Participants 110

3 Overview of Talks

3.1 Recoverable Self-Implementations of Primitive Operations

Hagit Attiya (Technion – Haifa, IL)

Joint work of Liad Nahum, Hagit Attiya, Ohad Ben-Baruch, Denny Hendler
Main reference Liad Nahum, Hagit Attiya, Ohad Ben-Baruch, Danny Hendler: "Recoverable and Detectable Fetch&Add". OPODIS 2021, LIPIcs Vol. 217
URL https://doi.org/10.4230/LIPIcs.OPODIS.2021.29

An attractive way to derive recoverable programs is to substitute every invocation of a primitive operation with a recoverable *self-implementation* of the same primitive.

We explore the properties that should be satisfied by such implementations for this approach to work. We also present some positive (implementations) and negative (impossibility) results, for primitives such as `test&set`, `fetch&add`, and `compare&swap`.

3.2 Consistency and Persistency: Challenges and Opportunities in Program Verification

Parosh Aziz Abdulla (Uppsala University, SE)

Joint work of Parosh Aziz Abdulla, Mohamed Faouzi Atig, Ahmed Bouajjani, K. Narayan Kumar, Prakash Saivasan
Main reference Parosh Aziz Abdulla, Mohamed Faouzi Atig, Ahmed Bouajjani, K. Narayan Kumar, Prakash Saivasan: "Deciding reachability under persistent x86-TSO", Proc. ACM Program. Lang., Vol. 5(POPL), pp. 1–32, 2021.
URL http://dx.doi.org/10.1145/3434337

Nowadays, most application platforms offer more relaxed semantics than the classical Sequential Consistency (SC) semantics. There are two primary sources of relaxation, namely weak consistency and weak persistence. Weakly consistent platforms are present at all level of the system design: at the hardware level, e.g., multiprocessors such as x86-TSO, SPARC, IBM POWER, and ARM; at the language level, e.g., C11 or Java; and at the application level, e.g., distributed databases, and geo-replicated systems. All these platforms sacrifice SC to provide stronger efficiency guarantees.

Weak persistence means that the order in which data persist over system crashes is inconsistent with the order in which the data is generated by the application. Weakly persistent systems arise in intermittent computing, file systems, and (more recently) architectures that employ Nonvolatile memories (NVRAMs).

Concurrent programs exhibit entirely new behaviors compared to SC when running on platforms with relaxed semantics. Even textbook programs such as small mutual exclusion protocols or concurrent data structures that are provably correct under SC can now show counter-intuitive behaviors. Hence, the verification community is currently facing new exciting, complicated, and practically motivated challenges.

To make the ideas concrete, I will present the semantics of concurrent programs that run on the Persistent Intel x86 architecture (Px86), implemented in Intel's Optane memory chip. We investigate the state reachability problem and show how to prove its decidability for finite-state programs. To achieve that, we provide a new formal model that is equivalent to Px86, and that has the feature of being a well-structured system. Deriving this new model results from a deep investigation of the properties of Px86 and the interplay of its component.

3.3 Isolating the Thin Air Problem: Semantic Dependency for Optimised Concurrency

Mark Batty (University of Kent – Canterbury, GB)

Joint work of Marco Paviotti, Simon Cooksey, Anouk Paradis, Daniel Wright, Scott Owens, Brijesh Dongol, Alan Jeffrey, James Riely, Ilya Kaysin, Anton Podkopaev, Mark Batty
Main reference Alan Jeffrey, James Riely, Mark Batty, Simon Cooksey, Ilya Kaysin, Anton Podkopaev: "The leaky semicolon: compositional semantic dependencies for relaxed-memory concurrency", Proc. ACM Program. Lang., Vol. 6(POPL), pp. 1–30, 2022.
URL http://dx.doi.org/10.1145/3498716

Languages like C/C++ (and Java) include both aggressive sequential optimisation and unmediated concurrent access to memory. These features collude to produce the out-of-thin-air problem – the language definition allows the conjuring of erroneous values in concurrent executions that can never be seen in practice, wrecking our ability to reason about concurrent code. Previous solutions involve disallowing optimisations at a cost to performance, or rebasing the language semantics on a very different model – an unlikely route for a language specification like C/C++. Here we present a less intrusive fix: an oracle takes the program and calculates semantic dependency, a record of which thread-local dependencies are preserved, and this dependency relation is used in the standard concurrency model of C/C++. We highlight recent work presenting two options for the calculation of semantic dependency, and a prospective operational model and program logic built upon semantic dependency. We believe this approach is reusable in the context of persistent memory.

3.4 Persistent Transactions: Desirable Semantics and Efficient Designs

Michael D. Bond (Ohio State University – Columbus, US)

Joint work of Kann Genç, Guoqing Harry Xu, Michael D. Bond

I'll argue that persistent transactions should provide failure atomicity and strict durability and respect inter-thread dependencies, under an assumption of data race freedom. Is this behavior reasonable and sufficient? And how do we implement such persistent transactions efficiently? I'll suggest that a combination of shadow memory, redo logs, and reference-counting-based dependence tracking yields a simple design that is likely more efficient than prior designs, although the performance story is complicated by tradeoffs of using shadow memory. Looking forward to your feedback and discussion.

3.5 Model Checking Persistent Memory Programs

Brian Demsky (University of California – Irvine, US)

Joint work of Hamed Gorjiara, Guoqing Harry Xu, Brian Demsky
Main reference Hamed Gorjiara, Guoqing Harry Xu, Brian Demsky: "Jaaru: efficiently model checking persistent memory programs", in Proc. of the ASPLOS '21: 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Virtual Event, USA, April 19-23, 2021, pp. 415–428, ACM, 2021.
URL http://dx.doi.org/10.1145/3445814.3446735

Persistent memory (PM) technologies combine near DRAM performance with persistency and open the possibility of using one copy of a data structure as both a working copy and a persistent store of the data. Ensuring that these persistent data structures are crash consistent is a major challenge. Stores to persistent memory are not immediately made persistent — they initially reside in processor cache and are only written to PM when a flush occurs due to space constraints or explicit flush instructions. It is more challenging to test crash consistency for PM than for disks given the PM's byte-addressability that leads to significantly more states.

I present Jaaru, a fully-automated and efficient model checker for PM programs. Key to Jaaru's efficiency is a new technique based on constraint refinement that can reduce the number of executions that must be explored by many orders of magnitude. This exploration technique effectively leverages commit stores, a common coding pattern, to reduce the model checking complexity from exponential in the length of program executions to quadratic. We have evaluated Jaaru with PMDK and RECIPE, and found 25 persistency bugs, 18 of which are new.

3.6 View-Based Owicki–Gries Reasoning for Persistent x86-TSO

Brijesh Dongol (University of Surrey – Guildford, GB)

Joint work of Eleni Bila, Brijesh Dongol, Ori Lahav, Azalea Raad, John Wickerson
Main reference Eleni Bila, BBrijesh Dongol, Ori Lahav, Azalea Raad, John Wickerson: "View-Based Owicki–Gries Reasoning for Persistent x86-TSO". ESOP 2022 (To appear)

This work develops a program logic for reasoning about persistent x86 code that uses low-level operations such as memory accesses, fences, and flushes. Our logic, called Pierogi, benefits from an underlying operational semantics by Cho et al that is based on views. Pierogi is able to handle optimised flush operations that previous program logics for persistency could not. Pierogi is mechanised in the Isabelle/HOL proof assistant, which serves as a semi-automated verification tool. This talk will discuss the basics of Pierogi, its use in program verification, and its encoding in Isabelle/HOL.

3.7 General Constructions for Non-Volatile Memory

Michal Friedman (Technion – Haifa, IL)

Joint work of Michal Friedman, Erez Petrank, Guy E. Blelloch, Pedro Ramalhete, Naama Ben David, Yuanhao Wei
Main reference Michal Friedman, Erez Petrank, Pedro Ramalhete: "Mirror: making lock-free data structures persistent", in Proc. of the PLDI '21: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, Virtual Event, Canada, June 20-25, 2021, pp. 1218–1232, ACM, 2021.
URL http://dx.doi.org/10.1145/3453483.3454105

With the recent launch of the Intel Optane memory platform, non-volatile main memory in the form of fast, dense, byte-addressable non-volatile memory has now become available. Nevertheless, designing crash-resilient data structures is complex and error-prone, especially when caches and machine registers are still volatile and the data residing in memory after a crash might not reflect a consistent view of the program state. This talk will focus on NVTraverse and Mirror, which are two different general transformations that adds durability in an automatic manner to lock-free data structures, with a low performance overhead.

3.8 Formal Foundations for Intermittent Computing

Limin Jia (Carnegie Mellon University – Pittsburgh, US)

Joint work of Milijana Surbatovich, Brandon Lucia, Limin Jia
Main reference Milijana Surbatovich, Brandon Lucia, Limin Jia: "Towards a formal foundation of intermittent computing", Proc. ACM Program. Lang., Vol. 4(OOPSLA), pp. 163:1–163:31, 2020.
URL http://dx.doi.org/10.1145/3428231

Intermittently powered devices enable new applications in harsh or remote environments, e.g., space or in-body implants, but also introduce problems in programmability and correctness. A variety of intermittent systems to save and restore program state have been proposed to ensure complete program execution despite power failures. For such systems to execute programs correctly, non-volatile memory locations, whose writes may cause intermittent executions to diverge from continuously powered executions, need to be handled carefully. In this talk, I will present our work on formalizing and proving the correctness properties of intermittent systems and discuss our ongoing work targeting Rust.

3.9 Revamping Hardware Persistency Models: View-Based and Axiomatic Persistency Models for Intel-X86 and Armv8

Jeehoon Kang (KAIST – Daejeon, KR)

Joint work of CKyeongmin Cho, Sung Hwan Lee, Azalea Raad, Jeehoon Kang
Main reference Kyeongmin Cho, Sung Hwan Lee, Azalea Raad, Jeehoon Kang: "Revamping hardware persistency models: view-based and axiomatic persistency models for Intel-x86 and Armv8", in Proc. of the PLDI '21: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, Virtual Event, Canada, June 20-25, 2021, pp. 16–31, ACM, 2021.
URL http://dx.doi.org/10.1145/3453483.3454027

Non-volatile memory (NVM) is a cutting-edge storage technology that promises the performance of DRAM with the durability of SSD. Recent work has proposed several persistency models for mainstream architectures such as Intel-x86 and Armv8, describing the order in which writes are propagated to NVM. However, these models have several limitations; most notably, they either lack operational models or do not support persistent synchronization patterns.

We close this gap by revamping the existing persistency models. First, inspired by the recent work on promising semantics, we propose a unified operational style for describing persistency using views, and develop view-based operational persistency models for Intel-x86 and Armv8, thus presenting the first operational model for Armv8 persistency. Next, we propose a unified axiomatic style for describing hardware persistency, allowing us to recast and repair the existing axiomatic models of Intel-x86 and Armv8 persistency. We prove that our axiomatic models are equivalent to the authoritative semantics reviewed by Intel and Arm engineers. We further prove that each axiomatic hardware persistency model is equivalent to its operational counterpart. Finally, we develop a persistent model checking algorithm and tool, and use it to verify several representative examples.

This talk is based on a conference paper [1].

References

1 Kyeongmin Cho, Sung-Hwan Lee, Azalea Raad, and Jeehoon Kang. *Revamping hardware persistency models: view-based and axiomatic persistency models for Intel-x86 and Armv8*. PLDI 2021). DOI: `https://doi.org/10.1145/3453483.3454027`

3.10 Abstraction for Crash-Resilient Objects

Artem Khyzha (Arm – Cambridge, GB)

Joint work of Artem Khyzha, Ori Lahav
Main reference Artem Khyzha, Ori Lahav: "Abstraction for Crash-Resilient Objects (Extended Version)", CoRR, Vol. abs/2111.03881, 2021.
URL https://arxiv.org/abs/2111.03881

In this talk we discuss formal compositional reasoning under non-volatile memory. We develop a library correctness criterion that is sound for ensuring contextual refinement in this setting, thus allowing clients to reason about library behaviors in terms of their abstract specifications, and library developers to verify their implementations against the specifications abstracting away from particular client programs.

We employ a recent NVM model, called Persistent Sequential Consistency, as a semantic foundation, and extend its language and operational semantics with useful specification constructs. The proposed correctness criterion accounts for NVM-related interactions between client and library code due to explicit persist instructions, and for calling policies enforced by libraries.

3.11 Taming x86-TSO Persistency

Artem Khyzha (Arm – Cambridge, GB)

Joint work of Artem Khyzha, Ori Lahav
Main reference Artem Khyzha, Ori Lahav: "Taming x86-TSO persistency", Proc. ACM Program. Lang., Vol. 5(POPL), pp. 1–29, 2021.
URL http://dx.doi.org/10.1145/3434328

We study the formal semantics of non-volatile memory in the x86-TSO architecture. We show that while the explicit persist operations in the recent model of Raad et al. from POPL'20 only enforce order between writes to the non-volatile memory, it is equivalent, in terms of reachable states, to a model whose explicit persist operations mandate that prior writes are actually written to the non-volatile memory. The latter provides a novel model that is much closer to common developers' understanding of persistency semantics. We further introduce a simpler and stronger sequentially consistent persistency model, develop a sound mapping from this model to x86, and establish a data-race-freedom guarantee providing programmers with a safe programming discipline. Our operational models are accompanied with equivalent declarative formulations, which facilitate our formal arguments, and may prove useful for program verification under x86 persistency.

3.12 PerSeVerE: Persistency Semantics for Verification under Ext4

Michalis Kokologiannakis (MPI-SWS – Kaiserslautern, DE)

Joint work of Michalis Kokologiannakis, Ilya Kaysin, Azalea Raad, Viktor Vafeiadis
Main reference Michalis Kokologiannakis, Ilya Kaysin, Azalea Raad, Viktor Vafeiadis: "PerSeVerE: persistency semantics for verification under ext4", Proc. ACM Program. Lang., Vol. 5(POPL), pp. 1–29, 2021.
URL http://dx.doi.org/10.1145/3434324

Although ubiquitous, modern filesystems have rather complex behaviours that are hardly understood by programmers and lead to severe software bugs such as data corruption.

As a first step to ensure correctness of software performing file I/O, we have formalized the semantics of the Linux ext4 filesystem, which we have integrated with the weak memory consistency semantics of C/C++. In addition, we have developed an effective model checking approach for verifying programs that use the filesystem. While doing so, we discovered bugs in commonly-used text editors such as vim, emacs and nano.

My talk gives an overview of ext4's persistency semantics and our formalization, as well as of some editor bugs we found.

3.13 Weak Memory Schemes Used in the Linux Kernel

Paul McKenney (Facebook – Beaverton, US)

This talk gives background on persistent-memory mechanisms provided by the Linux kernel. The kernel generally does not rely on persistent memory internally because this would make it difficult to run the kernel on systems lacking persistent memory.

This talk then explains why many researchers have encountered significant performance penalties associated with strong ordering and persistent memory, showing how this is due to the finite speed of light (especially in solids), the non-zero size of atoms, protocol overheads (for example, due to cache coherence protocols), overheads from electrical fundamentals, and, in some cases, chemistry. These penalties are of special interest to organizations having large numbers of systems, where even a 1% increase in overhead can be quite expensive.

Further, the finite speed of light can make it impossible to determine the order in which external events occurred, in which case strong ordering might not be particularly helpful.

The talk concludes with a survey of a few of the weakly ordered concurrent algorithms used in the Linux kernel.

3.14 Hazard Pointer Synchronous Reclamation

Maged M. Michael (Facebook – New York, US)

Deferred reclamation techniques, such as hazard pointers, do not guarantee the timing of reclamation of reclaimable objects by default. This talk describes the problem of synchronous deferred reclamation, where users require guarantees for the timing of reclamation. This talk also reviews the semantics of various synchronous reclamation guarantees, and outlines cohort-based reclamation, a novel scalable algorithm for hazard pointer synchronous reclamation.

3.15 Persistent Lock-Free Data Structures for Non-Volatile Memory

Erez Petrank (Technion – Haifa, IL)

In this talk I will discuss the design of lock-free (concurrent) data structures adequate for non-volatile RAM. I will shortly review constructions of persistent queues and sets, mention general transformation and discuss the basic techniques behind all.

3.16 TSOPER: Efficient Coherence-Based Strict Persistency

Konstantinos Sagonas (Uppsala University, SE)

Joint work of Per Ekemark, Yuan Yao, Alberto Ros, Konstantinos Sagonas, Stefanos Kaxiras
Main reference Per Ekemark, Yuan Yao, Alberto Ros, Konstantinos Sagonas, Stefanos Kaxiras: "TSOPER: Efficient Coherence-Based Strict Persistency", in Proc. of the IEEE International Symposium on High-Performance Computer Architecture, HPCA 2021, Seoul, South Korea, February 27 – March 3, 2021, pp. 125–138, IEEE, 2021.
URL http://dx.doi.org/10.1109/HPCA51647.2021.00021

We propose a novel approach for hardware-based strict TSO persistency, called TSOPER. We allow a TSO persistency model to freely coalesce values in the caches, by forming atomic groups of cachelines to be persisted. A group persist is initiated for an atomic group if any of its newly written values are exposed to the outside world. A key difference with prior work is that our architecture is based on the concept of a TSO persist buffer, that sits in parallel to the shared LLC, and persists atomic groups directly from private caches to NVM, bypassing the coherence serialization of the LLC. To impose dependencies among atomic groups that are persisted from the private caches to the TSO persist buffer, we introduce a sharing-list coherence protocol that naturally captures the order of coherence operations in its sharing lists, and thus can reconstruct the dependencies among different atomic groups entirely at the private cache level without involving the shared LLC. The combination of the sharing-list coherence and the TSO persist buffer allows persist operations and writes to non-volatile memory to happen in the background and trail the coherence operations. Coherence runs ahead at full speed; persistency follows belatedly. Our evaluation shows that TSOPER provides the same level of reordering as a program-driven relaxed model, hence, approximately the same level of performance, albeit without needing the programmer or compiler to be concerned about false sharing, data-race-free semantics, etc., and guaranteeing all software that can run on top of TSO, automatically persists in TSO.

3.17 The Case for Buffered Persistence

Michael Scott (University of Rochester, US)

Joint work of Wentao Cai, Haosen Wen, Vladimir Maksimovski, Mingzhe Du, Rafaello Sanna, Shreif Abdallah, Michael L. Scott, Benjamin Valpey, Louis Jenkins, H. Alan Beadle, Mohammad Hedayati, Joseph Izraelevitz, Hammurabi Mendes
Main reference Wentao Cai, Haosen Wen, Vladimir Maksimovski, Mingzhe Du, Rafaello Sanna, Shreif Abdallah, Michael L. Scott: "Fast Nonblocking Persistence for Concurrent Data Structures", in Proc. of the 35th International Symposium on Distributed Computing, DISC 2021, October 4-8, 2021, Freiburg, Germany (Virtual Conference), LIPIcs, Vol. 209, pp. 14:1–14:20, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.
URL http://dx.doi.org/10.4230/LIPIcs.DISC.2021.14

For machines with nonvolatile memory (NVM) but volatile caches, most work on persistent data structures has assumed the need for strict durable linearizability – for operations whose effects are guaranteed to survive any crash that occurs after their return to the caller. Most programmers, however, aren't interested in persisting existing transient structures: they're interested in avoiding serialization and deserialization of structures currently kept long-term in block-structured files and databases. For such structures, programmers are comfortable with the familiar concept of *buffering*, which separates persistence from atomicity and consistency, allowing data to persist some time in the (not-too-distant) future or in response to an explicit `sync` operation.

Given the high latency of fence instructions, buffered persistence has the potential to significantly shorten the critical path of the application. To evaluate this potential, we have created a general-purpose persistence system, Montage, that divides time into multi-millisecond *epochs*. Each epoch boundary is guaranteed to represent a consistent cut across the happens-before graph of operations. On a crash, recovery restores state as of the second-to-last epoch boundary. Experiments with both micro and macro benchmarks confirm that Montage dramatically outperforms existing general-purpose persistence systems, rivals or exceeds the performance of special-purpose persistent structures, and indeed approaches the performance of *nonpersistent* structures placed in NVM.

3.18 A Taste of Armv8-A Relaxed Virtual Memory

Peter Sewell (University of Cambridge, GB)

Joint work of Ben Simner, Alasdair Armstrong, Jean Pichon-Pharabod, Christopher Pulte, Peter Sewell

Virtual memory is an essential mechanism for enforcing security boundaries, but its relaxed-memory concurrency semantics has not previously been investigated in detail. The concurrent systems code managing virtual memory has been left on an entirely informal basis, and OS and hypervisor verification has had to make major simplifying assumptions.

This talk will give a taste of work in progress on relaxed virtual memory semantics for the Armv8-A architecture, to support future system-software verification. We identify many design questions, in discussion with Arm; develop a test suite, including use cases from the pKVM production hypervisor under development by Google; delimit the design space with axiomatic-style concurrency models; prove that under simple stable configurations our architectural model collapses to previous "user" models; develop tooling to compute allowed behaviours in the model integrated with the full Armv8-A ISA semantics; and develop a hardware test harness.

This lays out some of the main issues in relaxed virtual memory, bringing these security-critical systems phenomena into the domain of programming-language semantics and verification, with foundational architecture semantics, for the first time.

3.19 The ISA Semantics / Concurrency Model Interface

Peter Sewell (University of Cambridge, GB)

Joint work of Peter Sewell, Jean Pichon-Pharabod, Alasdair Armstrong, Ben Simner, Brian Campbell, Christopher Pulte, Thomas Bauereiss, Shaked Flur

We want to establish a standard interface between instruction-set architecture (ISA) semantics and concurrency models, to support the many things one would like to do above them – especially as one moves to concurrency models that handle more systems features, and to complete ISA definitions rather than idealised fragments. In this talk I'll spell out some of the desiderata and our current sketch design; hopefully it will stimulate discussion with potential users.

The interface shouldn't be large or complex, but it does have to harmonise several different usages and cope with all the many phenomena we care about, and it may underlie a lot of future work, so it's worth polishing it as much as we reasonably can up-front (though it will surely also have to change as it's used). It builds on our existing ISA/concurrency integrations in the rmem and isla-axiomatic tools, for full Armv8-A and RISC-V ISAs in Sail, and for user, ifetch, and virtual-memory concurrency, and on our Sail-generated Isabelle and Coq ISA semantics. We want to clean these up and generalise them, both for tools and for theorem-prover reasoning.

3.20 Non-Temporal Stores and their Semantics

Viktor Vafeiadis (MPI-SWS – Kaiserslautern, DE)

Joint work of Azalea Raad, Luc Maranget, Viktor Vafeiadis
Main reference Azalea Raad, Luc Maranget, Viktor Vafeiadis: "Extending Intel-x86 consistency and persistency: formalising the semantics of Intel-x86 memory types and non-temporal stores", Proc. ACM Program. Lang., Vol. 6(POPL), pp. 1–31, 2022.
URL http://dx.doi.org/10.1145/3498683

In the research community, there are several formalisations of the sequential semantics of various fragments of the Intel-x86 architecture and a few that also describe their weakly consistent concurrency semantics and/or their persistency semantics. As far as the concurrency semantics is concerned, Intel-x86 is widely believed to follow the x86-TSO model of Sewell et al. [1]. The persistency semantics is similarly believed to follow the Px86 model of Raad et al. [2].

Nevertheless, this is not the full story. The x86-TSO and Px86 models cover only a small fragment of its available features that are relevant for the consistency semantics of multithreaded programs and the persistency semantics of programs interfacing with non-volatile memory. In particular, besides normal store instructions, Intel-x86 supports non-temporal stores, which provide higher performance and are used to ensure that updates are flushed to memory. Furthermore, Intel-x86 allows declaring regions of memory with different types—uncacheable, write-back, write-combined, write-protected, and write-through—and existing models cover only the semantics of the (default) write-back memory regions.

Since both non-temporal stores and memory regions with different types are widely used in systems code, Azalea Raad, Luc Maranget, and I have extended the existing x86-TSO and Px86 formalizations to cover these features and the interaction between them. Our formal model is published at POPL'22 [3]. My talk at the seminar presented an overview of these additional features: how they can be used and what semantics they have.

References

1 Peter Sewell, Susmit Sarkar, Scott Owens, Francesco Zappa Nardelli, and Magnus O. Myreen. *x86-TSO: a rigorous and usable programmer's model for x86 multiprocessors*. Commun. ACM 53(7), 2010. DOI: `https://doi.org/10.1145/1785414.1785443`

2 Azalea Raad, John Wickerson, Gil Neiger, and Viktor Vafeiadis. *Persistency semantics of the Intel-x86 architecture*. Proc. ACM Program. Lang. 4(POPL), 2020. DOI: `https://doi.org/10.1145/3371079`

3 Azalea Raad, Luc Maranget, and Viktor Vafeiadis. *Extending Intel-x86 consistency and persistency: formalising the semantics of Intel-x86 memory types and non-temporal stores*. Proc. ACM Program. Lang. 6(POPL), 2022. DOI: `https://doi.org/10.1145/3498683`

3.21 Architectural Support for Persistent Programming

William Wang (Arm – Cambridge, GB)

The talk opens with the use cases of NVM, including "more" memory and persistent memory, leveraging its density and persistence. Then the talk focuses on the persistent use, which requires software changes that bring programming challenges.

To reduce the barrier of entry for persistent memory and simplify persistent programming, we ask whether the Arm architecture has sufficient support for programming persistent memory. In searching for an answer, two problems related to ensuring persistent ordering across threads and within a thread are uncovered, and two solutions are outlined. Specifically, the persistent transitive stores at the instruction set architectural level and the battery-backed buffers at the microarchitectural level.

To wrap up, the talk briefly mentions about other persistent programming challenges, including failure atomicity, persistent addressing that can also benefit from architectural support.

3.22 Modularizing Verification of Durable Opacity

Heike Wehrheim (Universität Oldenburg, DE)

Joint work of Eleni Bila, John Derrick, Simon Doherty, Brijesh Dongol, Gerhard Schellhorn, Heike Wehrheim
Main reference Eleni Bila, John Derrick, Simon Doherty, Brijesh Dongol, Gerhard Schellhorn, Heike Wehrheim: "Modularising Verification Of Durable Opacity", CoRR, Vol. abs/2011.15013, 2020.
URL https://arxiv.org/abs/2011.15013

Opacity is the standard correctness condition for Software Transactional Memory Algorithms. One way of proving opacity is by showing a refinement relationship to hold between an abstract sequential specification and an implementation. Durable opacity is the analogue of opacity for settings with non-volatile memory. It can similarly be shown by refinement, now using a durable abstract specification. In the talk, I will present work towards (a) reusing existing refinement proofs of opacity for proving durable opacity, and (b) modularizing such proofs by capsulating all accesses to shared state in a library, and only showing this library to be durable linearizable.

Participants

- Mark Batty
University of Kent – Canterbury, GB
- Hans-J. Boehm
Google – Mountain View, US
- Michael D. Bond
Ohio State University – Columbus, US
- Brijesh Dongol
University of Surrey – Guildford, GB
- Michal Friedman
Technion – Haifa, IL
- Michalis Kokologiannakis
MPI-SWS – Kaiserslautern, DE
- Ori Lahav
Tel Aviv University, IL
- Maged M. Michael
Facebook – New York, US
- Erez Petrank
Technion – Haifa, IL
- Azalea Raad
Imperial College London, GB
- Konstantinos Sagonas
Uppsala University, SE
- Michael Scott
University of Rochester, US
- Viktor Vafeiadis
MPI-SWS – Kaiserslautern, DE
- William Wang
Arm – Cambridge, GB
- Heike Wehrheim
Universität Oldenburg, DE

Remote Participants

- Hagit Attiya
Technion – Haifa, IL
- Parosh Aziz Abdulla
Uppsala University, SE
- Piotr Balcer
Intel Technology – Gdansk, PL
- Eleni Bila
University of Surrey – Guildford, GB
- Dhruva Chakrabarti
AMD – Santa Clara, US
- Soham Chakraborty
TU Delft, NL
- Ricardo Ciríaco da Graca
MPI-SWS – Kaiserslautern, DE
- Brian Demsky
University of California – Irvine, US
- Derek Dreyer
MPI-SWS – Saarbrücken, DE
- João F. Ferreira
INESC-ID – Lisboa, PT
- Maurice Herlihy
Brown University – Providence, US
- Joseph Izraelevitz
University of Colorado – Boulder, US
- Limin Jia
Carnegie Mellon University – Pittsburgh, US
- Jeehoon Kang
KAIST – Daejeon, KR
- Artem Khyzha
Arm – Cambridge, GB
- Umang Mathur
National University of Singapore, SG
- Paul McKenney
Facebook – Beaverton, US
- Nikos Nikoleris
Arm – Cambridge, GB
- Andy Rudoff
Intel – Boulder, US
- Peter Sewell
University of Cambridge, GB
- John Wickerson
Imperial College London, GB

Report from Dagstuhl Seminar 21471

Geometric Modeling: Interoperability and New Challenges

Edited by

Falai Chen[1], Tor Dokken[2], and Géraldine Morin[3]

1 **University of Science & Technology of China – Anhui, CN,** chenfl@ustc.edu.cn
2 **SINTEF – Oslo, NO,** tor.dokken@sintef.no
3 **IRIT – University of Toulouse, FR,** geraldine.morin@irit.fr

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 21471 "Geometric Modeling: Interoperability and New Challenges". This seminar was initially planned on May 2021, and was delayed due to the pandemic. The seminar took place as a hybrid version with on site and remote participants. It provided a great opportunity for exchanges which, as pointed out by participants, were very appreciated in this period where international scientific interactions have been diminished.

This report summarizes the seminar communications, first by providing the abstracts of the talks which present recent results in geometric modeling. Moreover, the scientific exchanges during the seminar provided a great basis for scientific discussions that resulted to the included five reports which highlight the new and future challenges in Geometric Modeling.

Seminar November 21–26, 2021 – http://www.dagstuhl.de/21471

2012 ACM Subject Classification Mathematics of computing → Mathematical software; Computing methodologies → Artificial intelligence; Computing methodologies → Computer graphics; Applied computing → Physical sciences and engineering; Mathematics of computing → Numerical analysis

Keywords and phrases Additive Manufacturing; Computer Graphics; Design Optimization; Geometric Modeling; Geometry; Geometry Processing; Isogeometric Analysis; Shape Design; Computer-Aided Design

Digital Object Identifier 10.4230/DagRep.11.10.111

Edited in cooperation with Konstantinos Gavriil

1 Executive Summary

Falai Chen (Univiversity of Science & Technology of China – Anhui, CN)
Tor Dokken (SINTEF – Oslo, NO)
Géraldine Morin (IRIT – University of Toulouse, FR)

The Dagstuhl seminar, initially planned in May 2020, took place as a hybrid conference in November 2021. Eighteen participants were on site, and thirty three participated remotely out of which five from East Asia and twelve from America.

Due to the pandemic, getting together for a conference has been an important event, and an outstanding exchange time between researchers (compared to a two years pandemic context where interaction has greatly been reduced). In particular, having a significant part of the participants on site has been a real asset compared to the full online conferences. Note also that this has been particularly true for young researchers that are in the process of developing networks and developing collaborations.

Geometric Modeling: Interoperability and New Challenges, *Dagstuhl Reports*, Vol. 11, Issue 10, pp. 111–150
Editors: Falai Chen, Tor Dokken, and Géraldine Morin
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

48 talks were given including 18 on site and 30 remotely. The program was organized into topics and structured to the extent possible to minimize the challenges posed by the time difference between on-site and remote participants. Speakers from East-Asia were assigned time slots in the morning and speakers from America in the afternoon. The beginning of the afternoon was a privileged time for all participants to meet. The social afternoon was canceled, as it would not have been inclusive for remote participants.

The time freed allowed us to extend the time assigned to topic focused groups sessions. This triggered, under the supervision of five on-site participants, development of topic focused reports. Two of these reports, *The Future of CAD* (group led by Tom Grandine) and *Design Optimization* (group led by Konstantinos Gavriil) address the evolution of the application fields in Geometric Modeling, closely linked to its use in Industry. Three other reports on emerging topics have also been based on the group working sessions. *Additive Manufacturing* (a group led by Sylvain Lefebvre) has been identified as a disruptive technology and has triggered the emergence of new geometric models and materials. *Isogeometric Analysis* (group led by Carla Manni) addresses how the gap between geometric modeling and simulation can be bridged by replacing the traditional shape functions of Finite Element Analysis by B-splines that cross element boundaries. It thus supplies continuous models connecting the representations of Computer Aided Design and Finite Element Analysis. *Geometric Machine Learning* (group led by Rene Hiemstra) is a fast evolving domain. Deep learning approaches have already changed the field of Computer Vision, and the contribution into Geometric Modeling is becoming more pregnant. These reports offer to the participants, and beyond, a perspective of the coming challenges in the field of Geometric Modeling.

On top of the communications done in Dagstuhl, a special issue of the journal Graphical Models, has been planned. Submission to the journal is pending.

2 Table of Contents

Executive Summary

Falai Chen, Tor Dokken, Géraldine Morin 111

Overview of Talks

On the new class of spatial PH B-Spline curves
Gudrun Albrecht 116

Exploring challenges in shape analysis, generation, and optimization with neural networks
Arturs Berzins 116

Constructing planar domain parameterization with HB-splines via quasi-conformal mapping
Falai Chen 117

AI and Beyond in the World's Largest 3D Capture Stage
Ilke Demir 117

On the effect of scaled B-splines for different approaches to locally refined splines
Tor Dokken 118

Volumetric Representations: Design, Analysis, Optimization, and Fabrication of Porous/Heterogeneous Artifacts
Gershon Elber 118

Computational Design of Cold Bent Glass Façades
Konstantinos Gavriil 119

Recent advances on adaptive isogeometric methods with hierarchical spline models
Carlotta Giannelli 119

If I could do it over, I would. . .
Thomas A. Grandine 120

Geometric construction and fabrication of auxetic metamaterials
Stefanie Hahmann 120

Learning quadrature for implicit domains
Rene Hiemstra 121

Geometric Regularizations and Representations for Neural 3D Synthesis
Qi-xing Huang 121

Complete Classification and Efficient Determination of Arrangements Formed by Two Ellipsoids
Xiaohong Jia and Wenping Wang 121

Numerical integration on trimmed planar domains via high-order transport theorems for implicit curves
Bert Jüttler 122

Supporting Expensive Physical Models With Geometric Moment Invariants to Accelerate Sensitivity Analysis for Shape Optimisation
Panagiotis Kaklis 122

Efficient Multimodal Belief Propagation for Robust SLAM Using Clustering Based Reparameterization
Tae-wan Kim 124

On Triangular Splines: CAD and Quadrature
Jiri Kosinka 124

Cyclidic splines and kinematic interpretation of quaternionic curves/surfaces
Rimvydas Krasauskas 125

Generating oriented structures and trajectories within part volumes
Sylvain Lefebvre 125

Deep Implicit Moving Least-Squares Functions for 3D Reconstruction
Yang Liu 125

Outlier-free isogeometric discretizations
Carla Manni 126

Scale-Space for Machine Learning on 3D point clouds
Nicolas Mellado 127

Tubular parametric volume objects
Géraldine Morin 127

nTopology: A Design System Based on Implicit Modeling
Suraj R. Musuvathy 128

Polyhedral net spline modeling
Jörg Peters 128

Topology Optimization for Additive Manufacturing
Xiaoping Qian 129

Interoperability of Geometric Models and Numerical Analysis by Immersed Boundary Methods
Ernst Rank 129

ABC-Surfaces
Ulrich Reif 130

3D printed metamaterials in industry
Elissa Ross 131

CAD Model Details via Curved Knot Lines and Truncated Powers
Malcolm A. Sabin 131

Geometric interpolation of Euler-Rodrigues frames with G2 Pythagorean-hodograph curves of degree 7
Maria Lucia Sampoli 131

Explicit error estimates for isogeometric discretizations of partial differential equations
Espen Sande 132

C^1 isogeometric spaces
Giancarlo Sangalli 132

Smooth polar spline representations suited for design and analysis
Hendrik Speleers 133

Segmentation of X-Ray CT Volume of Binned Parts by Constructing Morse Skeleton Graph of Distance Transform
Hiromasa Suzuki 134

Multi-sided surface patches over curved, multi-connected domains
Tamás Várady 134

On Modeling Neural Implicit Surfaces with Detailed Features
Wenping Wang 135

A Deep Learning Approach for Non-rigid Registration
Juyong Zhang 135

An Isogeometric Analysis Based Topology Optimization Framework for Additive Manufacturing of 2D Cross-Flow Heat Exchangers
Yongjie Jessica Zhang 136

Working groups

The Future of CAD
Arturs Berzins, Tor Dokken, Nira Dyn, Gershon Elber, Konstantinos Gavriil, Carlotta Giannelli, Ron Goldman, Hyunsun Alicia Kim, Jiri Kosinka, Rimvydas Krasauskas, Tom Lyche, Carla Manni, Géraldine Morin, Suraj R. Musuvathy, Jeff Poskin, Ernst Rank, Maria Lucia Sampoli, Espen Sande, Hendrik Speleers, Deepesh Toshniwal, Nelly Villamizar 137

Design Optimization
Konstantinos Gavriil, Panagiotis Kaklis, Hyunsun Alicia Kim, Jeff Poskin, Helmut Pottmann 140

Additive Manufacturing
Gershon Elber, Sylvain Lefebvre, Géraldine Morin, Suraj R. Musuvathy, Stefanie Hahmann, Xiaoping Qian, Ernst Rank, Elissa Ross, Yongjie Jessica Zhang 143

Isogeometric Analysis
Carlotta Giannelli, Panagiotis Kaklis, Tom Lyche, Carla Manni, Malcolm Sabin, Espen Sande, Giancarlo Sangalli, Hendrik Speleers, Deepesh Toshniwal 145

Geometric Machine Learning
Arturs Berzins, Ilke Demir, Rene Hiemstra, Qi-xing Huang, Yang Liu, Nicolas Mellado, Géraldine Morin, Wenping Wang, Juyong Zhang 146

Acknowledgments 148

Participants 149

Remote Participants 149

3 Overview of Talks

3.1 On the new class of spatial PH B-Spline curves

Gudrun Albrecht (Universidad Nacional de Colombia – Medellin, CO)

Joint work of Gudrun Albrecht, Carolina Beccari, Lucia Romani

In this talk we present the spatial counterpart of the recently introduced class of planar Pythagorean-Hodograph (PH) B–Spline curves. Spatial Pythagorean-Hodograph B–Spline curves are odd-degree, non-uniform, parametric spatial B–Spline curves whose arc length is a B–Spline function of the curve parameter and can thus be computed explicitly without numerical quadrature. We provide the general construction of these curves using quaternion algebra and formulate the problem of point interpolation by clamped and closed PH B–Spline curves of arbitrary odd degree. In particular, we provide closed form solutions for the cubic and the quintic cases, and discuss how degree-$(2n+1)$, C^n-continuous PH B–Spline curves can be computed by optimizing several scale-invariant fairness measures with interpolation constraints. Finally, we define Rational B-Spline Euler Rodrigues Frames (RBSERF) for regular PH B-Spline curves as well as rational tensor product B-Spline pipe surfaces. A functional is introduced to minimize the rotation of the RBSERF, and the results are illustrated on the corresponding rational pipe surface.

References

1 G. Albrecht, C.V. Beccari, L. Romani. *Spatial Pythagorean-Hodograph B–Spline curves and 3D point data interpolation.* Comput. Aided Geom. Des. 80: 101868, 2020. `https://doi.org/10.1016/j.cagd.2020.101868`

2 G. Albrecht, C.V. Beccari, J.-C. Canonne, L. Romani. *Planar Pythagorean-Hodograph B-Spline curves.* Comput. Aided Geom. Des. 57: 57-77, 2017. `https://doi.org/10.1016/j.cagd.2017.09.001`

3.2 Exploring challenges in shape analysis, generation, and optimization with neural networks

Arturs Berzins (SINTEF – Oslo, NO)

Triangular surface meshes are a widespread representation in 3D shape applications and a variety of neural network (NN) architectures have been developed in the recent years to handle them directly. However, with the advent of additive manufacturing, shapes with internal structures gain ever more significance, motivating the need for NN architectures capable of processing representations that admit disconnected boundaries. For analysis tasks, a potential solution is to extend existing NN architectures to handle tetrahedral volumetric meshes. On the other hand, generative tasks pose a greater challenge with both surface and volumetric meshes being restricted to a fixed topology. NNs representing implicit level-set functions offer greater topological flexibility but can also produce unwanted disconnected components. This talk will explore early-stage ideas on NN architectures for handling tetrahedral volumetric meshes, imposing shape connectivity on NNs representing implicit level-set functions and, finally, interoperability of NNs and PDE solvers for shape and topology optimization as a particular generative task.

3.3 Constructing planar domain parameterization with HB-splines via quasi-conformal mapping

Falai Chen (University of Science & Technology of China – Anhui, CN)

Joint work of Falai Chen, Maodong Pan
Main reference Maodong Pan, Falai Chen, Weihua Tong: "Low-rank parameterization of planar domains for isogeometric analysis", Comput. Aided Geom. Des., Vol. 63, pp. 1–16, 2018.
URL https://doi.org/10.1016/j.cagd.2018.04.002

Constructing a high-quality parameterization of a computational domain is a fundamental research problem in isogeometric analysis, which has been extensively investigated so far. However, most of the current approaches employ non-uniform rational B-splines (NURBS) as the geometric representation of the physical domain. NURBS introduce redundant degrees of freedom due to their tensor-product structure. In this paper, we propose a new parameterization method for planar domains by adopting hierarchical B-splines (HB-splines) as the geometric representation that possess local refinement abilities. Starting from an initial parameterization such as a harmonic map, our method repeats the following two steps until a bijective parameterization with low distortion is achieved. First, a non-linear optimization model is proposed to compute a quasi-conformal map represented by HB-splines, and an efficient algorithm is provided to deal with this model by alternatively solving two quadratic optimization problems. Second, the parameterization is refined locally through HB-splines based on the bijectivity and conformal distortion of the parameterization. Several examples are demonstrated to verity the effectiveness and advantages of the proposed approach.

3.4 AI and Beyond in the World's Largest 3D Capture Stage

Ilke Demir (Intel – Hermosa Beach, US)

One picture is worth a thousand words, so what have been told with videos? What about 100 simultaneous videos to reconstruct every frame of life in a 10.000 sq. ft dome? Is it enough to reconstruct and digitize us realistically? Similar to other industries, entertainment industry is also being reshaped by AI, especially towards AR/VR consumption. Before democratization of AI and data, such immersive experiences were lacking an essential element: photorealism. As the amount of data increased, our models got deeper, and the reality became decipherable.

This talk will introduce recent deep learning advancements in 3D vision, reconstruction, and shape understanding techniques with a focus on generative models to digitize performances and scenes. Then we will shift gears with an overview of such models in 3D, and their progression on voxels, point clouds, meshes, graphs, and other 3D representations. Back to our studio, in addition to a discussion about how to process such large visual data, the challenges of scaling 10x over current capture platforms, and over 200x over state-of-the-art datasets will be presented. The talk will conclude with a sneak peek of upcoming VR/AR productions from the world's largest volumetric capture stage at Intel Studios, as an example of real-world use cases of such AI approaches.

3.5 On the effect of scaled B-splines for different approaches to locally refined splines

Tor Dokken (SINTEF – Oslo, NO)

Both Truncated Hierarchical B-splines and LR B-splines use scaled B-splines as part of the construction to ensure partition of unity. Certain configurations of refinements have a surprisingly big effect on the scaling factor. As the scaling factors have a direct effect on magnitude of the elements in the mass and stiffness matrices scaling consequently directly impacts the condition numbers of these matrices. The talk addressed how these effects can be observed already for bi-degree $(3, 3)$, and that the effect grows with growing polynomial degrees.

3.6 Volumetric Representations: Design, Analysis, Optimization, and Fabrication of Porous/Heterogeneous Artifacts

Gershon Elber (Technion – Haifa, IL)

The needs of modern (additive) manufacturing technologies can be satisfied no longer by Boundary representations (B-reps), as they requires the representation and manipulation of interior (material) properties and fields. Further, while the need for a tight coupling between the design and analysis stages has been recognized as crucial almost since geometric modeling (GM) was conceived, contemporary GM systems only offer a loose link between the two, if at all.

For about half a century, (trimmed) Non Uniform Rational B-spline (NURBs) surfaces has been the B-rep of choice for virtually all the GM industry. Fundamentally, B-rep GM has evolved little during this period. In this talk, we will present a kernel of a volumetric representation (V-rep) that is based on (trimmed) trivariate NURBs and successfully confront the existing and anticipated design, analysis, and manufacturing foreseen challenges, toward porous, heterogeneous and anisotropic representation. With a V-rep kernel that supports all fundamental B-rep GM operations, such as primitive constructors and Boolean operations, we present a tight link to (Isogeometric) analysis on one hand and the full support of (heterogeneous) additive manufacturing on the other.

In this talk, we will present numerous examples that exemplify the portrayed advantages of V-reps over B-reps.

Work in collaboration with many others, including Ben Ezair, Fady Massarwi, Boris van Sosin, Jinesh Machchhar, Ramy Masalha, Q Youn Hong, Sumita Dahiya, Annalisa Buffa, Giancarlo Sangalli, Pablo Antolin, Massimiliano Martinelli, Bob Haimes and Stefanie Elgeti.

3.7 Computational Design of Cold Bent Glass Façades

Konstantinos Gavriil (SINTEF – Oslo, NO)

Joint work of Konstantinos Gavriil, Ruslan Guseinov, Jesús Pérez, Davide Pellis, Paul Henderson, Florian Rist, Helmut Pottmann, Bernd Bickel

Main reference Konstantinos Gavriil, Ruslan Guseinov, Jesús Pérez, Davide Pellis, Paul Henderson, Florian Rist, Helmut Pottmann, Bernd Bickel: "Computational design of cold bent glass façades", ACM Trans. Graph., Vol. 39(6), pp. 208:1–208:16, 2020.

URL https://doi.org/10.1145/3414685.3417843

Cold bent glass is a promising and cost-efficient method for realizing doubly curved glass façades. They are produced by attaching planar glass sheets to curved frames and require keeping the occurring stress within safe limits. However, it is very challenging to navigate the design space of cold bent glass panels due to the fragility of the material, which impedes the form-finding for practically feasible and aesthetically pleasing cold bent glass façades. We propose an interactive, data-driven approach for designing cold bent glass façades that can be seamlessly integrated into a typical architectural design pipeline. Our method allows non-expert users to interactively edit a parametric surface while providing real-time feedback on the deformed shape and maximum stress of cold bent glass panels. Designs are automatically refined to minimize several fairness criteria while maximal stresses are kept within glass limits. We achieve interactive frame rates by using a differentiable Mixture Density Network trained from more than a million simulations. Given a curved boundary, our regression model is capable of handling multistable configurations and accurately predicting the equilibrium shape of the panel and its corresponding maximal stress. We show predictions are highly accurate and validate our results with a physical realization of a cold bent glass surface.

3.8 Recent advances on adaptive isogeometric methods with hierarchical spline models

Carlotta Giannelli (University of Firenze, IT)

Joint work of Cesare Bracco, Carlotta Giannelli, Tadej Kanduč, Mario Kapl, Massimiliano Martinelli, Giancarlo Sangalli, Mattia Tani, Rafael Vázquez

The design and analysis of adaptive isogeometric methods with hierarchical spline constructions has attracted remarkable interest in the last few years. In order to increase the flexibility of the hierarchical approximation framework, while simultaneously preserving the performance of the overall adaptive scheme, particular attention is currently devoted to address the fast formation of system matrices arising from hierarchical discretization as well as to the development of effective multi-patch extensions. The talk will present recent results on these directions.

3.9 If I could do it over, I would...

Thomas A. Grandine (Seattle, US)

I recently retired after a nearly 35 year career researching, developing, and deploying geometric modeling and processing tools for The Boeing Company. The year I've spent away from the daily demands of the job have given me an opportunity to reflect on what mathematical technologies I wish I had made more effective use of, those which I wish I had been able to push more deeply into the company's well-established bag of tricks, and those I didn't pursue at all, but wish I had. Additionally, there are things I did pursue that turned out not to be very good ideas. This talk will provide a quick summary of the lessons learned.

3.10 Geometric construction and fabrication of auxetic metamaterials

Stefanie Hahmann (INRIA Grenoble Rhône-Alpes, FR)

Joint work of Stefanie Hahmann, Georges-Pierre Bonneau, Johana Marku
Main reference Georges-Pierre Bonneau, Stefanie Hahmann, Johana Marku: "Geometric construction of auxetic metamaterials", Comput. Graph. Forum, Vol. 40(2), pp. 291–304, 2021.
URL https://doi.org/10.1111/cgf.142633

Recent advances in digital manufacturing, where computational design, materials science and engineering meet, offer whole new perspectives for tailoring mechanical properties and fabrication of new meta-materials with applications as diverse as product design, architecture, engineering and art. A meta-material is a material whose microstructure can be controlled to achieve the desired macroscopic deformation behavior.

This presentation is devoted to a category of metamaterials called auxetic structures, or auxetic networks. Auxetic materials are characterized by a negative Poisson's ratio. They do not behave like usual materials, because when they are stretched in one direction, they expand in the perpendicular direction.Whereas regular auxetic networks are well studied, our focus is on disordered auxetic networks. In particular, we are exploring geometrical strategies to generate 2-dimensional random auxetic meta-materials. Starting from a dense irregular network, we seek to reduce the Poisson's ratio, by pruning bonds (edges) based solely on geometric criteria. To this end, we first deduce some prominent geometric features from regular auxetic networks and then introduce a strategy combining a pure geometric pruning algorithm followed by a physics-based testing phase to determine the resulting Poisson's ratio of our networks. We provide numerical results and statistical validation. We also show physical tests with both laser-cut rubber networks and 3D-printed networks showing auxetic behaviour.

3.11 Learning quadrature for implicit domains

Rene Hiemstra (Leibniz Universität Hannover, DE)

Implicit geometry representations are becoming increasingly widespread in applications, driven by a need and interplay of complex geometry, topology and physics. In this work I develop a supervised learning approach to "learn" quadrature for implicitly defined domains by means of feed-forward artificial neural networks. Once optimized, these networks can accurately and efficiently evaluate quadrature rules for implicit domains. This greatly simplifies and accelerates numerical quadrature for fictitious domain methods, which is illustrated by several numerical benchmarks.

3.12 Geometric Regularizations and Representations for Neural 3D Synthesis

Qi-xing Huang (University of Texas – Austin, US)

Synthesizing 3D shapes and scenes is a core problem in visual computing. Recent advances in deep generator models have pushed this field to a new ear. However, existing approaches predominately generalize machine learning algorithms developed on images to the 3D domain. They do not fully utilize geometric properties of 3D shapes, e.g., in shape deformation and 3D representations. This talk covers recent works that construct geometric regularizations and hybrid representations for 3D synthesis. Specifically, I will first discuss an as-rigid-as possible regularization loss that regularizes the tangent spaces defined by the generator for training deformable shape generators. I will then move to the domain of 3D scenes and present a recent work that combines the strengths of neural generators and traditional approaches for modeling uncertainties of geometric attributes.

3.13 Complete Classification and Efficient Determination of Arrangements Formed by Two Ellipsoids

Xiaohong Jia (Chinese Academy of Sciences, CN) and Wenping Wang (Texas A&M University – College Station, US)

Joint work of Xiaohong Jia, Changhe Tu, Bernard Mourrain, Wenping Wang
Main reference Xiaohong Jia, Changhe Tu, Bernard Mourrain, Wenping Wang: "Complete Classification and Efficient Determination of Arrangements Formed by Two Ellipsoids", ACM Trans. Graph., Vol. 39(3), pp. 27:1–27:12, 2020.
URL https://doi.org/10.1145/3388540

Arrangements of geometric objects refer to the spatial partitions formed by the objects and they serve as an underlining structure of motion design, analysis and planning in CAD/CAM, robotics, molecular modeling, manufacturing and computer-assisted radio surgery. Arrangements are especially useful to collision detection, which is a key task in various applications such as computer animation, virtual reality, computer games, robotics, CAD/CAM and computational physics.

Ellipsoids are commonly used as bounding volumes in approximating complex geometric objects in collision detection. In this paper we present an in-depth study on the arrangements formed by two ellipsoids. Specifically, we present a classification of these arrangements and propose an efficient algorithm for determining the arrangement formed by any particular pair of ellipsoids. A stratification diagram is also established to show the connections among all the arrangements formed by two ellipsoids.

3.14 Numerical integration on trimmed planar domains via high-order transport theorems for implicit curves

Bert Jüttler (Johannes Kepler Universität Linz, AT)

Joint work of Felix Scholz, Bert Jüttler
Main reference Felix Scholz, Bert Jüttler: "Using High-Order Transport Theorems for Implicitly Defined Moving Curves to Perform Quadrature on Planar Domains", SIAM J. Numer. Anal., Vol. 59(4), pp. 2138–2162, 2021.
URL https://doi.org/10.1137/20M1341283

We study numerical integration over a planar domain that is cut by an implicitly defined boundary curve. This important problem arises, for example, in unfitted finite element methods and in isogeometric analysis on trimmed computational domains. We present a general version of the transport theorem for moving domains defined by implicitly defined curves. This result is then used to derive an efficient and accurate quadrature rule for this class of domains. Numerical experiments are presented in order to demonstrate that the method achieves a high rate of convergence.

3.15 Supporting Expensive Physical Models With Geometric Moment Invariants to Accelerate Sensitivity Analysis for Shape Optimisation

Panagiotis Kaklis (The University of Strathclyde – Glasgow, GB)

Joint work of Panagiotis Kaklis, Shahroz Khan, Andrea Serani, Matteo Diez

Parametric Sensitivity Analysis (PSA) investigates the sensitivity of parameters, defining the design space of a shape-optimisation problem, for tackling the challenges of the curse of dimensionality or decreasing the uncertainty in design's performance. However, the analytical implementation of PSA can often be tricky, especially if the chosen method requires the evaluation of high-dimensional integrals or if the baseline simulation codes do not provide an analytical solution to design performance. Therefore, PSA needs to be implemented with sampling methods, such as Monte Carlo sampling, which is highly susceptible to slow convergence and necessitates a sufficiently large number of samples for stable results, especially for high-dimensional problems.

In this work, we aim to address above the mentioned challenges associated with PSA by offloading the evaluation of parametric sensitivities from physical quantities to quantities, which are relatively inexpensive but, like physical metrics, provide important clues about the form, distribution and validity of the design. It is well known that shape's integral properties, such as geometric moments and their invariants serve as a geometric foundation for different designs' physical analyses. In this connection, we propose a geometric moment-dependent PSA approach, that harnesses the geometric variation of designs in the design space using geometric moments as a quantity of interest (QoI) to identify parametric sensitivities. These results can serve as prior estimates of parametric sensitivities with respect to physics. The selection of geometric moments in our work is motivated by the following baseline insights:

- It is very likely that physics analysis requires the computation of such integral properties of the geometry such as the stiffness and mass matrix, and moments of a domain are sufficient to ensure accurate integration of a large class of integrands.
- Like physics, geometric moments can also act as a compact shape signature or descriptor to a specific design falling in a specific category, which facilitates various shape processing tasks.

To validate our approach and experimentally demonstrate the effectiveness of geometric moments, we used two ship hulls parameterised with 27 and 26 parameters using two different techniques based on *procedural deformation* (PD) and global modification function (GMF), respectively. In this setting, we use the *wave-resistance coefficient* (C_w) as the physical QoI, as it plays a crucial role in ship hull design. The longitudinal distribution of the hulls' geometry has a similar impact on geometric moments as C_w.

To commence, we construct the so-called shape-signature vector ($\mathcal{MI}^s$), that will be used as shape descriptor and contains all the geometric moments up to s order. To align better with C_w, all moments in this vector are taken invariant with respect to translation and scaling. A Global Variance-Based Sensitivity Analyses (GVBSA) is performed for learning parametric sensitivities with respect to $\mathcal{MI}^s$ and C_w. Here $\mathcal{MI}^s$ is purely a vector quantity containing the moments of various orders while C_w is a scalar and computationally expensive one. Therefore, learning sensitivities to $\mathcal{MI}^s$ requires implementing a multivariate extension of GVBS, such as covariance decomposition, which provides generalised sensitivity indices of design parameters to all moments in $\mathcal{MI}^s$.

The results from the experiments conducted in this study show a good correlation between the sensitive parameters obtained from C_w and $\mathcal{MI}^s$, specifically with the fourth-order shape-signature vector $\mathcal{MI}^4$. In the case of the PD-based hull, 7 parameters sensitive to $\mathcal{MI}^4$ are also among the 8 parameters sensitive to C_w. Interestingly, similar results are obtained for the GMF-based hull, where 6 out of 7 sensitive parameters to C_w are also sensitive to $\mathcal{MI}^4$. Afterwards, two different design spaces are constructed for both hull models, one with sensitive parameters obtained with C_w and the other with $\mathcal{MI}^4$. Shape optimisation is performed in both spaces performed with a meta-heuristic optimisation approach. Final optimisation results showed that the design generated from design space constructed with sensitive parameters of C_w and $\mathcal{MI}^4$ for both types of hulls offer similar performance. These results indicate that PSA performed with moments can reasonably estimate parameters' sensitivity to the design's physics with considerably reduced computational cost.

3.16 Efficient Multimodal Belief Propagation for Robust SLAM Using Clustering Based Reparameterization

Tae-wan Kim (Seoul National University, KR)

Joint work of Seungwon Choi, Tae-Wan Kim
Main reference Seungwon Choi, Taewan Kim: "Efficient Multimodal Belief Propagation for Robust SLAM Using Clustering Based Reparameterization", in Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 – Oct. 1, 2021, pp. 3354–3360, IEEE, 2021.
URL https://doi.org/10.1109/IROS51168.2021.9636040

Due to the presence of ambiguities caused by sensor noise and structural similarity, simultaneous localization and mapping (SLAM) observation models are typically multimodal. The multimodal inference process can be directly dealt with by belief propagation (BP) using weighted Gaussian mixture messages, but for efficiency, a combinatorial explosion of the complexity must be suitably relaxed. In this study, we present an effective multimodal BP SLAM for robust inference with ambiguities. Using Gaussian bandwidth mean shift and cluster-based reparameterization, we reduce the number of Gaussian components in each message due to the BP nature. The proposed algorithm reduces the number of components of the product by summarizing indistinguishable modes in weighted Gaussian mixtures and keeping only the significant modes, making BP computationally efficient.

3.17 On Triangular Splines: CAD and Quadrature

Jiri Kosinka (University of Groningen, NL)

Joint work of Gerben Jan Hettinga, Michael Barton, Jiri Kosinka

The standard representation of CAD (computer aided design) models is based on the boundary representation (B-reps) with trimmed and (topologically) stitched tensor-product NURBS patches. Due to trimming, this leads to gaps and overlaps in the models. While these can be made arbitrarily small for visualisation and manufacturing purposes, they still pose problems in downstream applications such as (isogeometric) analysis and 3D printing.

It is therefore worthwhile to investigate conversion methods which (necessarily approximately) convert these models into water-tight or even smooth representations. After briefly surveying existing conversion methods, we will focus on techniques that convert CAD models into triangular spline surfaces of various levels of continuity. In the second part, we will investigate efficient quadrature rules for triangular spline spaces.

References

1 Gerben Jan Hettinga, Jiří Kosinka. Conversion of B-rep CAD models into globally G1 triangular splines. Computer Aided Geometric Design, Volume 77, 2020, 101832. `https://doi.org/10.1016/j.cagd.2020.101832`

2 Jiří Kosinka, Michael Bartoň. Gaussian quadrature for C1 cubic Clough–Tocher macro-triangles. Journal of Computational and Applied Mathematics, Volume 351, 2019, pp. 6-13. `https://doi.org/10.1016/j.cam.2018.10.036`

3.18 Cyclidic splines and kinematic interpretation of quaternionic curves/surfaces

Rimvydas Krasauskas (Vilnius University, LT)

Joint work of Rimvydas Krasauskas, Jean Michel Menjanahary

Regular circular quad meshes produce smooth cyclidic splines composed of principal patches of Dupin cyclides with simply computable offsets. This property is crucial in CAD and 3D printing applications. We increase flexibility of such splines by extending the variety of blended patches: now they can be bounded not only by principal circles but also by diagonals of principal patches, i.e. quartic curves in general. Our methods are based on kinematic interpretation of quaternionic-Bezier formulas and Moebius invariant constructions. In the case of cubic cyclidic splines possibility of foldings and branchings of the Gaussian map is demonstrated. Topological restrictions are detected for general cyclidic splines without spherical or planar patches.

3.19 Generating oriented structures and trajectories within part volumes

Sylvain Lefebvre (LORIA & INRIA – Nancy, FR)

Joint work of Sylvain Lefebvre, Jonàs Martinez, Cédric Zanni, Thibault Tricard, Samuel Hornus, Jimmy Etienne, Adrien Bedel, Yoann Coudert-Osmont, Rahnuma Islam Nishat, Sue Whitesides, Vincent Tavernier, Fabrice Neyret, Pierre-Alexandre Hugron, Tim Kuipers

Generating trajectories for Additive Manufacturing processes is typically performed under conflicting objectives. In particular, it is often desirable for the trajectories to follow specific directions within the part volume, aligning with shape features or following a user-specified control field. This directionality may for instance result in controlled anisotropic properties (elasticity, specularity), which could not be easily obtained with traditional manufacturing processes.

However, such an objective conflicts with that of producing equally spaced trajectories – avoiding porosities or excess material curing/deposition – and process constraints such as solidification radii bounds, admissible overhangs or continuity of deposition.

In this presentation I will discuss some of our latest research in controlling deposition orientation within part volumes – as well as their potential applications – for both sparse and dense infills.

3.20 Deep Implicit Moving Least-Squares Functions for 3D Reconstruction

Yang Liu (Microsoft Research – Beijing, CN)

Point set is a flexible and lightweight representation widely used for 3D deep learning. However, their discrete nature prevents them from representing continuous and fine geometry, posing a major issue for learning-based shape generation. In this work, we turn the discrete point sets into smooth surfaces by introducing the well-known implicit moving least-squares (IMLS) surface formulation, which naturally defines locally implicit functions on point sets. We incorporate IMLS surface generation into deep neural networks for inheriting both the flexibility of point sets and the high quality of implicit surfaces. I do see that there are many opportunities to bridge classic and well-defined shape representation and machine learning for improving the interoperability of shape learning and synthesis and making classic methods more robust and applicable.

3.21 Outlier-free isogeometric discretizations

Carla Manni (University of Rome "Tor Vergata", IT)

Joint work of Carla Manni, Espen Sande, Hendrik Speleers
Main reference Carla Manni, Espen Sande, Hendrik Speleers: "Application of optimal spline subspaces for the removal of spurious outliers in isogeometric discretizations", Computer Methods in Applied Mechanics and Engineering, Vol. 389, p. 114260, 2022.
URL https://doi.org/10.1016/j.cma.2021.114260

Spectral analysis can be used to study the error in each eigenvalue and eigenfunction of a numerical discretization of an eigenvalue problem. For a large class of boundary and initial-value problems the total discretization error on a given mesh can be recovered from its spectral error. This is of primary interest in engineering applications.

The isogeometric approach for eigenvalue problems has been widely investigated in the literature. Maximally smooth spline spaces on uniform grids are an excellent choice for addressing eigenvalue problems. Yet, they still present a flaw: a very small portion of the eigenvalues are poorly approximated and the corresponding computed values are much larger than the exact ones. These spurious values are usually referred to as outliers. The number of outliers increases with the degree p. However, for fixed p, it is independent of the degrees of freedom for univariate problems, while a thin layer of outliers is observed in the multivariate setting.

Outlier-free discretizations are appealing, not only for their superior description of the spectrum of the continuous operator, but also for their beneficial effects in various contexts, such as an efficient selection of time-steps in (explicit) dynamics and robust treatment of wave propagation. For a fixed degree, the challenge is to remove outliers without loss of accuracy in the approximation of all eigenfunctions.

In this talk we discuss isogeometric Galerkin discretizations of eigenvalue problems related to the Laplace operator subject to any standard type of homogeneous boundary conditions conditions in certain optimal spline subspaces. Roughly speaking, these optimal subspaces are obtained from the full spline space defined on specific uniform knot sequences by imposing specific additional boundary conditions. The spline subspaces of interest have been introduced in the literature some years ago when proving their optimality with respect to Kolmogorov n-widths. For a fixed number of degrees of freedom, all the eigenfunctions and the corresponding eigenvalues are well approximated, without loss of accuracy in the whole spectrum when compared to the full spline space. Moreover, there are no spurious values in the approximated spectrum. In other words, the considered subspaces provide accurate outlier-free discretizations in the univariate and in the multivariate tensor-product case.

The role of such spaces as accurate discretization spaces for addressing general problems is discussed as well.

3.22 Scale-Space for Machine Learning on 3D point clouds

Nicolas Mellado (CNRS – Toulouse, FR)

Applying Machine Learning algorithms to acquired point clouds with hundreds of millions of points remains a challenging task. First, networks have to handle the size of the point cloud, which is out of reach of most approaches found in the literature, only demonstrated on point clouds with a few thousands points. Second, the methods need to handle the complexity of the represented shapes (e.g., millimeter-scale acquisition of entire buildings), while being robust to acquisition artifacts (noise, sampling, holes). In this talk we introduce a new approach where point samples are used to compute implicit surface representation at multiple scales using state of the art algorithms, known to be robust, fast and stable. A neural network is then used to analyze the differential properties of the reconstructed surfaces, and perform a given task (here illustrated on pointwise edge classification). Instead of trying to learn how to be robust to acquisition defects, we propose to hide the complexity of the acquired data using surface reconstruction, and to define small, and fast networks requiring very small amounts of training data. We demonstrate the benefit of our approach on very large point clouds (e.g., buildings with dozens of millions of points) and also on collections of CAD geometry with thousands of objects (ABC dataset). We also show that low processing time and low data requirements unlock the definition of interactive learning applications, where the user can interactively show localized classification examples, train the network in seconds, and classify the entire dataset in seconds.

3.23 Tubular parametric volume objects

Géraldine Morin (IRIT – University of Toulouse, FR)

Joint work of Samuel Peltier, Géraldine Morin, Damien Aholou
Main reference Samuel Peltier, Géraldine Morin, Damien Aholou: "Tubular parametric volume objects: Thickening a piecewise smooth 3D stick figure", Comput. Aided Geom. Des., Vol. 85, p. 101981, 2021.
URL https://doi.org/10.1016/j.cagd.2021.101981

A volume parametric model is computed from a piecewise smooth skeleton. Generating a volume model from a stick figure S defined in 3D is an intuitive process: given S whose topology is a pseudo-graph and whose edges are embedded as Bézier curves in $\mathbb{R}^3$, we propose a method for creating a thick volume parametric model "around" S. The volume model we generate is based on semi-simploidal sets, which guarantees a proper topology and provides a 3D parametric domain for Bézier spaces. This volume is a continuous piecewise Bézier representation which boundary corresponds to a B-Rep made of tensor product Bézier patches.

3.24 nTopology: A Design System Based on Implicit Modeling

Suraj R. Musuvathy (nTopology – New York, US)

URL https://ntopology.com

Additive manufacturing enables the design and manufacturing of objects with unprecedented complexity and customizability. Advances in generative design exploration methods such as topology optimization, multi-disciplinary optimization, and AI have enabled rapid exploration of large design spaces. Most, if not all, major CAD software systems today are built on Boundary Representations (B-Reps). In our view B-Reps have reached their limits on addressing the design opportunities available today due to limitations in scalability and reliability. B-Reps require explicit representations of geometry and topology, thereby falling short on the ability to model complex structures that are manufacturable today by several orders of magnitude. B-Rep based modeling algorithms (Booleans, offsets, blends, etc.) are fragile thereby limiting automation required for generative techniques and mass customization applications. nTopology has developed a design system based on procedural implicit modeling that addresses these limitations. Solid geometry is represented as an implicit scalar-valued expression consisting of standard mathematical operators (arithmetic, logic, trigonometric, etc.) and custom operators. The custom operators enable interfacing any kind of data, including other representations of geometric data, with the solid modeling system by implementing a set of queries including computing scalar-field values, and optionally gradients and intervals. For example, a custom operator can be built for querying physics simulation analysis results in order to drive solid model geometry directly. The most common modeling algorithms are simple mathematical expressions (e.g., Booleans can be defined with min/max functions or other R-function variants), and are therefore robust and computationally fast. Procedural implicit functions can represent geometry of arbitrary complexity, and fundamental geometry processing algorithms can be easily parallelized thereby enabling effective use of modern multi-core CPUs and GPUs. There have been advances in visualization, physics simulation, topology optimization, and machine learning approaches that work directly with implicits. As these applications mature, an implicit representation approach holds great promise in delivering design engineering applications required for product development with unprecedented capabilities.

3.25 Polyhedral net spline modeling

Jörg Peters (University of Florida – Gainesville, US)

Joint work of Jörg Peters, Kestutis Karčiauskas, Kyle Lo
URL https://uf-cise-surflab.github.io/blender-polyhedral-splines-web/

Piecewise polynomial splines with polyhedral control nets allow for merging parameter directions and transitioning between coarse and fine meshes in a unified fashion. This talk summarizes recent work and available software for interpreting quad-dominant polyhedral control nets as control nets of smooth polyhedral splines of degree at most bi-3.

References

1 Karčiauskas, K., & Peters, J. (2015). Smooth multi-sided blending of biquadratic splines. Computers & Graphics, 46, 172-185.
`https://doi.org/10.1016/j.cag.2014.09.004`

2 Karčiauskas, K., & Peters, J. (2020). Smooth polar caps for locally quad-dominant meshes. Computer Aided Geometric Design, 81, 101908.
`https://doi.org/10.1016/j.cagd.2020.101908`

3 Karčiauskas, K., & Peters, J. (2020). Low degree splines for locally quad-dominant meshes. Computer Aided Geometric Design, 83, 101934.
`https://doi.org/10.1016/j.cagd.2020.101934`

3.26 Topology Optimization for Additive Manufacturing

Xiaoping Qian (University of Wisconsin – Madison, US)

URL https://qian.me.wisc.edu/

Topology optimization (TO) and additive manufacturing (AM) are twin-technologies that can be synergistically integrated to exploit shape flexibility in part design and fabrication. AM processes can fabricate parts of complex geometric shapes. However they also pose geometric constraints for part design. Some of these constraints such as overhang angle and support volume depend on part build orientation, and are not differentiable to build orientation. In topology optimization where a large number of optimization variables, often on the orders of millions, are used, gradient based optimization is preferred. How to formulate differentiable AM constraints for TO has been a challenge.

In this talk, I will present several formulations of AM constraints in density-based TO that are differentiable with respect to both density variables and build orientation. These differentiable formulations have explicit geometric meanings including projected undercut perimeter, overhang angle for self-support constraint, and support volume. These formulations thus enable simultaneous topology and build orientation optimization, yielding designs that meet AM constraints. Two essential elements of such differential formulations are implicit treatment of density representation, and the use of advection-diffusion equations for discerning accessibility and occlusion. Numerical examples will be given to demonstrate the efficacy of the proposed formulations.

3.27 Interoperability of Geometric Models and Numerical Analysis by Immersed Boundary Methods

Ernst Rank (TU München, DE)

Joint work of Ernst Rank, Davide d'Angella, Stefan Kollmannsberger, Nina Korshunova, Lászloó Kudela, Beatrice Wassermann
URL https://www.cms.bgu.tum.de/en/team/rank

Immersed Boundary Methods (IBM) like the Finite Cell or the CutFEM method have gained large interest in the mathematical as well as in the engineering community. A domain of computation is embedded in a larger, typically simply shaped fictitious domain, which is meshed e.g. in a simple Cartesian grid. The exact shape of the original domain is only observed on the level of element matrices by using a point membership test (PMT), which can be specifically designed for a given geometric model. By construction, Immersed Boundary Methods do NOT need any generation of body fitted meshes and thus relieve from one of the major practical obstacles of true geometry-analysis interoperability. This advantage comes yet with significant challenges, like precise numerical integration of cut cells, adequate imposition of boundary conditions, (pre)-conditioning of the arising algebraic systems or local refinement of the approximation. Successful solutions for these problems have recently been obtained, without compromising the accuracy or computational efficiency of the corresponding Finite Element or Isogeometric Element approximations. This presentation will focus on the principles of IBM and then show examples with various types of geometric models. Among these examples are (flawed) BRep- and Constructive Solid Geometry models, VReps, Computer Tomograms and Point Cloud models. Also evolving domains relevant in process simulation for additive manufacturing profit from the non-boundary conforming discretization of IBMs.

References

1 Düster, A., Parvizian, J., Yang, Z., and Rank, E. The finite cell method for three-dimensional problems of solid mechanics. Computer methods in applied mechanics and engineering, 197(45-48):37683782, 2008.
`https://doi.org/10.1016/j.cma.2008.02.036`

2 Korshunova N., Alaimo G., Hosseini SB., Carraturo M., Reali A., Niiranen J., Auricchio F., Rank E., Kollmannsberger S., Bending behavior of octet-truss lattice structures: Modelling options, numerical characterization and experimental validation. Materials & Design 205, 109693, 2021.
`https://doi.org/10.1016/j.matdes.2021.109693`

3.28 ABC-Surfaces

Ulrich Reif (TU Darmstadt, DE)

Joint work of Florian Martin, Ulrich Reif
Main reference Florian Martin, Ulrich Reif: "Trimmed Spline Surfaces with Accurate Boundary Control". In: C. Manni and H. Speleers (eds.) Geometric Challenges in Isogeometric Analysis, Springer INdAM Series, to appear.

Composite trimmed NURBS surfaces are a standard tool in industrial free form modeling. However, they are typical discontinuous along boundaries of neighboring patches. In this talk, we present a solution of the problem, called ABC-surfaces. It is based on an appropriate blend of a given surfaces patch, representing the overall shape, and so-called ribbons, representing the shape near the segments of the boundary. Using ABC-surfaces, it is possible to model composite spline surfaces of arbitrary smoothness. Another important feature is the fact that all building blocks and also the resulting surfaces can be represented in terms of standard NURBS elements, what is crucial for a potential integration in commercial CAD systems.

ABC-surfaces can also be used for single-patch parametrizations of planar domains bounded by spline curves. These parametrizations are close to the identity and can be constructed in a systematic way, avoiding the notorious problems of meshing. Thus, ABC-surfaces are a promising new tool for the simulation of boundary value problems. A generalization to the 3D case is possible.

3.29 3D printed metamaterials in industry

Elissa Ross (Metafold 3D – Toronto, CA)

Additive manufacturing has opened doors to the physical realization of geometry that cannot be fabricated by traditional methods. Examples include 3D lattice structures, which may have elements composed of either beams or surfaces, and may be arbitrarily complex or have extremely high surface area. By varying the lattice geometry together with the 3D printing medium, it is possible to achieve a vast spectrum of material behaviour, ranging from fully flexible forms to completely stiff examples with high strength. This range of material expression makes lattice geometry extremely promising for industrial applications, yet there remain numerous obstacles to their adoption. In this talk I will discuss recent work to address this through the development of 3D printing software and hardware specifically to print lattices, microstructures, metamaterials, and procedurally generated geometry.

3.30 CAD Model Details via Curved Knot Lines and Truncated Powers

Malcolm A. Sabin (Cambridge, GB)

Joint work of Malcolm A. Sabin, Chris Fellows, Jiří Kosinka
Main reference Malcolm A. Sabin, Chris Fellows, Jirí Kosinka: "CAD Model Details via Curved Knot Lines and Truncated Powers", Comput. Aided Des., Vol. 143, p. 103137, 2022.
URL https://doi.org/10.1016/j.cad.2021.103137

This presentation describes the background of and concepts underlying the work in [1].

In particular it covered the requirements for automotive body shell design and how they can be met better than by the current conventions and workflow practises. Other talks in this meeting which address this issue are those by Ulrich Reif (Talk 3.28) and by Tamás Várady (Talk 3.36).

References

1 Malcolm A. Sabin, Chris Fellows, Jiří Kosinka. *CAD Model Details via Curved Knot Lines and Truncated Powers*. Computer-Aided Design 143, 2022.
`https://doi.org/10.1016/j.cad.2021.103137`

3.31 Geometric interpolation of Euler-Rodrigues frames with G2 Pythagorean-hodograph curves of degree 7

Maria Lucia Sampoli (University of Siena, IT)

Joint work of Marjeta Knez, Maria Lucia Sampoli
Main reference Marjeta Knez, Maria Lucia Sampoli: "Geometric interpolation of ER frames with G^2 Pythagorean-hodograph curves of degree 7", Comput. Aided Geom. Des., Vol. 88, p. 102001, 2021.
URL https://doi.org/10.1016/j.cagd.2021.102001

In this talk a novel construction of spatial curves interpolating assigned positions and boundary frames is presented. The proposed construction results in Pythagorean-Hodograph (PH) curve segments of degree 7 with G^2 continuity and having the associated Euler-Rodrigues frame G^1 continuous. Therefore it can be used to form a spline curve whose frame is varying continuously, which is a feature very useful in motion design applications. Exploiting the relation between rotational matrices and quaternions on the unit sphere, geometric continuity conditions on the frames are expressed through conditions on the corresponding quaternion polynomials. This leads to a nonlinear system of equations whose solvability is investigated, and asymptotic analysis of the solutions in the case of data sampled from a smooth parametric curve and its general adapted frame is derived. It is shown that there exist PH interpolants with optimal approximation order 6, except for the case of the Frenet frame, where the approximation order is at most 4. Several numerical examples are presented, which confirm the theoretical results.

3.32 Explicit error estimates for isogeometric discretizations of partial differential equations

Espen Sande (EPFL – Lausanne, CH)

Joint work of Espen Sande, Carla Manni, Hendrik Speleers
Main reference Espen Sande, Carla Manni, Hendrik Speleers: "Explicit error estimates for spline approximation of arbitrary smoothness in isogeometric analysis", Numerische Mathematik, Vol. 144(4), pp. 889–929, 2020.
URL https://doi.org/10.1007/s00211-019-01097-9

In this talk we discuss techniques to obtain error estimates with explicit constants for Ritz-type projections onto spline spaces of arbitrary smoothness defined on arbitrary grids. The presented error estimates indicate that smoother spline spaces exhibit better approximation per degree of freedom, even for low regularity of the function to be approximated. This is in complete agreement with the numerical evidence found in the literature.

The extension of these error estimates to the case of mapped geometries (both single-patch and multi-patch) will also be mentioned.

3.33 C^1 isogeometric spaces

Giancarlo Sangalli (University of Pavia, IT)

Joint work of Giancarlo Sangalli, Andrea Benvenuti, Annabelle Collin, Gabriele Loli, Mario Kapl, Thomas Takacs
Main reference Mario Kapl, Giancarlo Sangalli, Thomas Takacs: "A family of C^1 quadrilateral finite elements", Adv. Comput. Math., Vol. 47(6), p. 82, 2021.
URL https://doi.org/10.1007/s10444-021-09878-3

In this talk I have presented results (from papers in collaboration) about the construction of C^1 isogeometric quadrilateral elements that could be seen as extensions of the the classical Argyris triangular element. The structure is different (triangular and quadrilateral elements are structurally different) but the d.o.f.s and space contraints have some similarities. When the quadrilateral is a spline patch, the optimal order of approximation requires some constraints of the parametrization that needs to be "analysis-suitable G^1". An alternative is to enforce the C^1 interelement continuity in a weak sense, e.g. by the mortar method.

References

1 Argyris, J. H., I. Fried, and D. W. Scharpf (1968). The TUBA family of plate elements for the matrix displacement method. The Aeronautical Journal 72(692), 701–709.

2 Benvenuti, A. (2016). Isogeometric Analysis for C^1-continuous Mortar Method. Ph. D. thesis, University of Pavia.

3 Brenner, S. C. and L.-Y. Sung (2005). C^0 interior penalty methods for fourth order elliptic boundary value problems on polygonal domains. Journal of Scientific Computing 22(1-3), 83–118.

4 Collin, A., G. Sangalli, and T. Takacs (2016). Analysis-suitable G^1 multi-patch parametrizations for C^1 isogeometric spaces. Computer Aided Geometric Design 47, 93 – 113.

5 Kapl, M., G. Sangalli, and T. Takacs (2021). A family of C^1 quadrilateral finite elements. Advances in Computational Mathematics 47 (6), 1–38.

3.34 Smooth polar spline representations suited for design and analysis

Hendrik Speleers (University of Rome "Tor Vergata", IT)

Joint work of Hendrik Speleers, Deepesh Toshniwal
Main reference Hendrik Speleers, Deepesh Toshniwal: "A General Class of C1 Smooth Rational Splines: Application to Construction of Exact Ellipses and Ellipsoids", Comput. Aided Des., Vol. 132, p. 102982, 2021.
URL https://doi.org/10.1016/j.cad.2020.102982

One of the upshots of CAD representations of arbitrary genus surfaces with finite number of polynomial patches is the introduction of holes surrounded by periodic configurations. Such holes can then be filled by means of polar spline surfaces, where the basic idea is to use periodic spline patches with one collapsed boundary (polar singularity).

In this talk, keeping in mind applications to design as well as analysis, we focus on C^k polar spline surfaces. We present a simple, geometric construction of basis functions over such polar configurations possessing interesting properties as nonnegativity and partition of unity. The polar basis functions are assembled by transforming sets of B-splines/NURBS via compatible extraction matrices. To increase flexibility, one could also start from different sets of B-splines/NURBS that are allowed to have different polynomial degrees and weight functions.

The polar spline representations are suited for geometric modeling of geometries with one or more polar singularities, and in particular allow for compact, smooth, low degree descriptions of ellipsoids. Moreover, the constructed splines show optimal approximation behavior, even at the polar singularity. These properties make them also attractive for isogeometric analysis. Thanks to their construction in terms of B-splines/NURBS, they can be readily implemented and used in CAD or CAE software.

3.35 Segmentation of X-Ray CT Volume of Binned Parts by Constructing Morse Skeleton Graph of Distance Transform

Hiromasa Suzuki (University of Tokyo, JP)

Joint work of Hiromasa Suzuki, Tatsuya Yatagawa, Yutaka Ohtake

Industrial X-ray CT scanners have delivered non-destructive evaluation of industrial products with its capability of inspecting even inside the body of products. This paper introduces a new approach to accelerate inspection of a large number of the same mechanical parts by scanning their heap in a bin at once. The scanning result is a CT volumetric image containing all of these parts out of which each part is segmented for inspection. This segmentation is a kind of template matching problem. However, random postures and dense contacts of the binned parts prohibit extracting the parts one-by-one using a traditional template matching due to its high computational complexity. To reduce the computational complexity, we convert both the scanned volumetric images of the template and the binned parts to simpler graph structures, and then, we solve well-studied graph matching problem to distinguish each part. We convert a discrete volume data to a distance field by the distance transform, and then, construct a graph consisting of nodes at extremum points of the distance field based on the Morse theory. The experimental evaluation demonstrates that our method without manual arrangement of the target parts works even for the scan of a heap of 50 binned parts in CT volumes of about 800^3 voxels, and an average processing time is as short as 30 minutes.

3.36 Multi-sided surface patches over curved, multi-connected domains

Tamás Várady (Budapest University of Technology and Economics, HU)

Joint work of Márton Vaitkus, Tamás Várady, Péter Salvi, Ágoston Sipos
Main reference Márton Vaitkus, Tamás Várady, Péter Salvi, Ágoston Sipos: "Multi-sided B-spline surfaces over curved, multi-connected domains", Comput. Aided Geom. Des., Vol. 89, p. 102019, 2021.
URL https://doi.org/10.1016/j.cagd.2021.102019

A new control point based parametric surface representation is presented, that interpolates a collection of surface ribbons, i.e. boundary curves and cross-derivatives, given in Bézier or B-spline form. A single surface equation can describe complex, multi-connected free-from surfaces, that are compatible with tensor-product surfaces. The scheme is defined over a planar domain with curved boundaries that mimic the shape of the 3D boundary curves, and it is capable to handle strongly concave boundaries and periodic hole loops in the interior.

We discuss (i) an algorithm for curved domain generation, (ii) the methods of defining local parameterizations using harmonic functions, (iii) the blending functions associated with the control points of the ribbons, (iv) the composition of the surface equation and (v) options to edit the interior of the patch.

The main area of application is curve network based design, hole filling (vertex blending) and general lofting, in particular when watertight connections are important. The strength of the scheme is its flexibility to define complex shapes; its main weakness is that it cannot be given in standard form. Several examples will be given to compare the difficulties of classical surfacing approaches and the benefits of the new multi-sided scheme.

References

1 T. Várady, P. Salvi, Gy. Karikó, A Multi-sided Bézier patch with a simple control structure. Computer Graphics Forum, Vol. 35(2), pp. 307-317, 2016.
`https://doi.org/10.1111/cgf.12833`

2 T. Várady, P. Salvi, M. Vaitkus, Á. Sipos, Multi-sided Bézier surfaces over curved, multi-connected domains. Computer Aided Geometric Design, Vol. 78, 101828, 2020.
`https://doi.org/10.1016/j.cagd.2020.101828`

3 M. Vaitkus, T. Várady, P. Salvi, Á. Sipos, Multi-sided B-spline surfaces over curved, multi-connected domains. Computer Aided Geometric Design, Vol. 89, 102019, 2021.
`https://doi.org/10.1016/j.cagd.2021.102019`

3.37 On Modeling Neural Implicit Surfaces with Detailed Features

Wenping Wang (Texas A&M University – College Station, US)

The neural implicit representation has recently emerged as a compact, powerful means for shape representation. A main challenge in this direction of research is to enable neural networks to accurately represent important shape characteristics, e.g. sharp edges or fine-scale details. Sinusoidal positional encoding (PE) has been proposed in network training to better represent high-frequency details in images or shapes. However, naively applying sinusoidal PE often results in unwanted wavy artifacts on surfaces or even failure of the learned implicit to converge to the target shape. We study how the interplay between the point sample density and the dimension of PE used for network training would affect the expressiveness of a neural implicit representation model. Our finding is that while increasing the dimension of PE is beneficial to modeling fine geometric details, it is critical to also increase the point sample density accordingly in order to avoid unwanted artifact. Specifically, we derive empirical results on the optimal coupling of the point sampling and the dimension of PE. Extensive experiments show that our new training strategy outperforms the other competing SOTA methods for neural implicit surface modeling, in terms of approximation accuracy, shape feature preservation, and training efficiency.

3.38 A Deep Learning Approach for Non-rigid Registration

Juyong Zhang (University of Science & Technology of China – Anhui, CN)

Joint work of Juyong Zhang, Wanquan Feng
Main reference Wanquan Feng, Juyong Zhang, Hongrui Cai, Haofei Xu, Junhui Hou, Hujun Bao: "Recurrent Multi-view Alignment Network for Unsupervised Surface Registration", CoRR, Vol. abs/2011.12104, 2020.
URL https://arxiv.org/abs/2011.12104

Learning non-rigid registration in an end-to-end manner is challenging due to the inherent high degrees of freedom and the lack of labeled training data. In the first work, we resolve these two challenges simultaneously. First, we propose to represent the non-rigid transformation with a point-wise combination of several rigid transformations. This representation not only makes the solution space well-constrained but also enables our method to be solved iteratively with a recurrent framework, which reduces the difficulty of learning. Second, we introduce a differentiable loss function that measures the 3D shape similarity on the projected multi-view 2D depth images so that our full framework can be trained end-to-end without ground truth supervision. In the second work, we propose the differentiable deformation graph based neural non-rigid registration method. Specifically, we design a neural network to predict the correspondence and its reliability confidence rather than the strategies like nearest neighbor search and pair rejection. The model is trained in a self-supervised manner, and thus can be used for arbitrary datasets without ground-truth.

3.39 An Isogeometric Analysis Based Topology Optimization Framework for Additive Manufacturing of 2D Cross-Flow Heat Exchangers

Yongjie Jessica Zhang (Carnegie Mellon University – Pittsburgh, US)

Joint work of Xinghua Liang, Angran Li, Anthony D. Rollett, Yongjie Jessica Zhang
Main reference Xinghua Liang, Angran Li, Anthony D. Rollett, Yongjie Jessica Zhang: "An isogeometric analysis based topology optimization framework for additive manufacturing of 2D cross-flow heat exchangers". Engineering with Computers, under review, 2021

Heat exchangers (HXs) have gained increasing attention due to the intensive demand of performance improving and energy saving for various equipment and machines. As a natural application, topology optimization has been involved in the structural design of HXs aiming at improving heat exchange performance (HXP) and meanwhile controlling pressure drop (PD). In this paper, a novel multiphysics based topology optimization framework is developed to maximize the HXP between two fluids with different temperatures for 2D cross-flow HXs, and concurrently minimize the PD between the fluid inlet and outlet. In particular, an isogeometric analysis (IGA) solver is developed to solve the coupled steady-state Navier-Stokes and heat convection-diffusion equations. Non-body-fitted control mesh is adopted instead of dynamically remeshing the design domain during the evolution of the two-fluid boundary interface. The method of moving morphable voids (MMVs) is employed to represent and track boundary interface between these two different fluids. In addition, various constraints are incorporated to guarantee proper manufacturability of the optimized structures with respect to practical manufacturing process such as additive manufacturing. To implement the iterative optimization process, the method of moving asymptotes (MMA) is employed. Numerical examples show that the HXP of the optimized structure is greatly improved compared with its corresponding initial design, and the PD between the fluid inlet and outlet is minimized concurrently. Moreover, smooth boundary interface between two fluids and improved manufacturability are also obtained for the optimized structures.

4 Working groups

4.1 The Future of CAD

Arturs Berzins (SINTEF – Oslo, NO), Tor Dokken (SINTEF – Oslo, NO), Nira Dyn (Tel Aviv University, IL), Gershon Elber (Technion – Haifa, IL), Konstantinos Gavriil (SINTEF – Oslo, NO), Carlotta Giannelli (University of Firenze, IT), Ron Goldman (Rice University – Houston, US), Hyunsun Alicia Kim (UC – San Diego, US), Jiri Kosinka (University of Groningen, NL), Rimvydas Krasauskas (Vilnius University, LT), Tom Lyche (University of Oslo, NO), Carla Manni (University of Rome "Tor Vergata", IT), Géraldine Morin (IRIT – University of Toulouse, FR), Suraj R. Musuvathy (nTopology – New York, US), Jeff Poskin (The Boeing Company – Seattle, US), Ernst Rank (TU München, DE), Maria Lucia Sampoli (University of Siena, IT), Espen Sande (EPFL – Lausanne, CH), Hendrik Speleers (University of Rome "Tor Vergata", IT), Deepesh Toshniwal (TU Delft, NL), Nelly Villamizar (Swansea University, GB)

4.1.1 Robustness of solid modeling operations

Introduced in the 1970's, boundary representations (B-reps) of solids are today's standard for CAD and engineering applications used in product development including physics simulation and manufacturing. Despite improvements over the past four decades B-rep modeling algorithms (Booleans, offsets, blends, etc.) are still fragile. This is due to the fact that they are built on a fundamentally flawed representation of trimmed surfaces. A B-rep containing several faces will contain many surface-surface intersections, and failure in the computation in any one makes the entire model invalid. Modeling failures usually require tedious manual intervention and rework by experienced CAD users. Therefore design exploration techniques (MDO) and applications like mass customization that require automation of engineering workflows driving parametric solid models are severely limited. New shape synthesis algorithms such as topology optimization create complex organic shapes that challenge B-rep modeling algorithms. It is time to reconsider solid modeling representations and algorithms from a fundamental perspective.

4.1.2 Additive manufacturing induced scalability issues

Additive manufacturing enables fabrication of shapes and structures of unprecedented complexity including for example lattices, geometric surface textures, and organic shapes generated by topology optimization. B-reps require explicit representation of geometry and topology. Consider a hierarchical lattice containing 100 × 100 × 100 beams and each beam being a smaller scale lattice consisting of a 1000 primitives. A B-rep for such a structure requires billions of faces just to represent it. Performance of mainstream commercial CAD systems degrades with objects having more than a hundred thousand faces, leaving them short of designing such complex structures by several orders of magnitude. A fundamentally different approach for representing and performing modeling operations is required in order to leverage the capabilities of additive manufacturing. Other related engineering applications including physics simulation, manufacturing, design exploration, interop standards, etc. will also need to work with new modeling representations in order to support complete product engineering workflows.

4.1.3 Augmentable engineering rich models, variable materials

Modern engineering relies on robust digital models and simulations across separate disciplines to design, produce, maintain, and support products. A key component in integrating disciplines is the creation of a digital thread, a communications framework connecting authoritative sources of information in standard formats throughout the product lifecycle. Current CAD models typically lack sufficient engineering data, e.g. material or structural properties, for downstream models and simulations, e.g. electromagnetic or aerodynamic analysis. In order to support a digital thread, CAD models must incorporate the information required for discipline-specific models and simulations. A promising idea for delivering this information is the extension of geometry models past two or three dependent variables, allowing structural or material data to be delivered with design information.

4.1.4 Adaptive models for digital twins

A digital twin is a virtual representation of a physical system that captures system performance and maintains synchronization with that system through its operational life. The connection between a digital twin and its physical system occurs through sensors gathering data on real-time operations, maintenance and repair reports, etc. The geometric representation of a digital twin differs from traditional design in that the representation must serve an entire engineering process instead of an individual component in the process. In particular, the representation must be adaptable to changes that occur throughout the operational life of its physical counterpart and integrated with multiphysics models that enable accurate predictions of system performance.

4.1.5 V-Rep geometry

The geometric CAD world is employing B-reps or boundary-representations for over half a century, B-reps that are B-spline surfaces based. This representation is no longer sufficient consider the need for representing complex interior geometries (i.e. micro-structures) as well as functionally graded properties. Future fabricated artifacts will be porous as well as heterogeneous, two new degrees of freedom that are enabled by AM. AM not only allows one to create highly complex (porous) geometries but also deposit different materials in different places. Such printers are already out there and yet there is a complete lack of support of such abilities in contemporary CAD software. One possible remedy can be found in an emerging representation that is trivariate based V-reps or volumetric-representation. V-reps fully encompass the geometry as in B-reps but also allows the tight representation of boundary-compatible internal scalar, vector and tensor fields alongside the geometry. The V-reps not only seamlessly support micro-structure representations but also allows one to have multiresolution microstructures, allowing for nanostructures inside microstructures, picostructures in nanostructures, etc. Further, the V-rep representation is fully compatible with IGA and hence makes the connection between design and analysis/optimization much tighter (then in B-reps), as it should be. Finally, V-reps already support 3D printing of heterogeneous (and porous) geometries.

4.1.6 Implicit modeling

Implicit function representations of solids present several benefits. Primitive shapes commonly used in design as well as complex structures like lattices and textures can be represented easily by analytical expressions (e.g., conics, triply periodic minimal surfaces) or procedural formulations (e.g., the modulo function). Modeling operations such as Booleans are simple math expressions (min/max or other R-function variants) irrespective of the geometric and topological complexity of the objects. Therefore modeling operations are robust and computationally fast. Implicit representations naturally lend themselves to parallelizable algorithms and so modern multi-core CPU and GPU architectures can be effectively leveraged. There have been advances in procedural modeling, visualization, physics simulation, topology optimization, and machine learning approaches that work directly with implicits. The theoretical foundations of implicit modeling have been introduced in the literature more than two decades ago, and as implicit function based applications mature, it holds much promise in addressing robustness and scalability challenges of B-rep based design systems. However, several challenges exist in adopting implicit modeling based approaches for engineering product design. Signed distance functions (SDFs) are especially useful for modeling but most implicit functions are not SDFs and the result of modeling operations on SDFs are no longer SDFs. So computing SDFs from arbitrary implicits efficiently remains an open challenge. Given that B-reps are the standard for mainstream product development today, effective and automated interop solutions with B-reps are necessary until entire product engineering workflows can be performed with implicits. Automated construction of a B-rep from an implicit remains an open challenge, especially for objects with sharp features and high genus. It may be desirable to develop other engineering applications such as NC machining directly based on implicits. New interop standards for implicits will also need to be developed.

4.1.7 Topology optimization

The long-standing challenge of topology optimization in the context of CAD is that its outcome is based on the finite element discretization and the piecewise constant representation of geometry. The recent emergence of the level set topology optimization approach has decoupled the geometry update of the optimization operation to the analysis and the implicit level set design representation opens up a new path that can enable a closer integration with CAD and a wider range of computational mechanics methods. This is particularly relevant for complex systems for coupling multiple scales and disciplines, and presents exciting new challenges in interoperable multifidelity mechanics models for topology optimization. The recent trends in topology optimization for complex coupled systems are timely with the rising interests of the digital twin and CAD interoperability. The research is not only from the mathematical and engineering mechanics operations, but also needs to consider the overall design workflow architecture and software modularization, i.e. the geometry is a core element of analysis software interoperability that follows and record the specific physical and mathematical assumptions. There are realistic pathways and the associated challenges now to enable a deployment of topology optimization within CAD to solve complex hierarchical problems and integrate back into the overall systems design.

4.1.8 Local refinement of tensor product spline spaces

There are three current approaches to local refinement of tensor product spline spaces:

- Specification of regions to be refined. Truncated Hierarchical B-splines (THB)
- Refinement by adding new meshline segments: Locally Refine B-splines (LRB)
- Refinement by adding new vertices in the control mesh: Standard T-splines (STS)

The spline spaces of LRB and STS are spanned by B-splines. In the general case the B-splines spanning the space will not form a partition of unity. However, partition of unity is imposed by scaling the B-splines with positive scaling factors of value less than or equal to

1. For THB partition of unity is achieved by truncating B-splines from a rougher level with B-splines from the finer levels. However, for all approach there is a risk that the scaling or truncation of some B-splines is so extreme that it has a direct effect on condition numbers for stiffness and mass matrices. A proposed new direction is to combine the requirement for minimal support B-splines from LRB and truncation from THB to significantly improve these condition numbers. There is also a need to better understand the richness and structure of B-splines of the different approaches and how they distribute the B-splines over the domain. An unwanted spatial distribution can directly influence the result of analysis and other computations.

4.1.9 Other topics

Other topics mentioned during the working group discussion but not developed further are listed here:

- Watertight models
- Integrating analysis results back into CAD
- Additive manufacturing induced scalability issues
- Robustness of solid modeling operations
- Global and sufficiently regular parametrization of CAD models
- CAD interoperability, including proper handling of proprietary data
- Parametric families of non-constant topology models
- Representing geometry to leverage AI techniques
- Rational offset surfaces
- Non-constant geometry models (time-variant, deformable, robotic, etc.)
- New representations (e.g. macro elements, polar splines)
- AI techniques for CAD
- Quantum computing
- Less primitive primitives

4.2 Design Optimization

Konstantinos Gavriil (SINTEF – Oslo, NO), Panagiotis Kaklis (The University of Strathclyde – Glasgow, GB) , Hyunsun Alicia Kim (UC – San Diego, US), Jeff Poskin (The Boeing Company – Seattle, US), Helmut Pottmann (KAUST – Thuwal, SA)

4.2.1 Explainable optimization

A black-box approach or a fully automated optimization process is not always desired. The input and interpretation of an expert designer can lead the optimization process to more desirable results, not possible through rigid automated processes. To facilitate this designer-in-the-loop option, the explainability of the optimization is essential. This can be achieved by several key improvements and features. Clear communication of the design space insights to the designer would eliminate the ambiguity inherent in black-box approaches. Treating the designer's personal preference as a latent optimization objective would set in place a system that leads the solution to the designer's intention. Providing better visualization of the analysis and optimization results will also make the design interaction easier. The possibility of multiple solutions is also a possibility that explainable optimization should be able to handle. This is a more general issue where the compatibility of the problem formulation and the solution methodology is critical and will lead to a better definition of the engineering problem. We find this an important challenge that needs to be addressed by the community.

4.2.2 Exploratory design optimization

We identify several potential future challenges in exploratory design optimization. These are the handling of multiple design optimization solutions during exploration, improved communication and feedback on quantities of interest, and the enhancement of the relation between exploratory design and the digital twin.

Elaborating on the latter part, the relation between design exploration and the digital twin should be bidirectional. The design model should be not only sufficiently flexible to represent the geometry and performance alterations during the product's life cycle, but also be able to incorporate the data collected through the digital twins in a manner that improves design optimization and guides design exploration. This incorporation is an inverse design problem and could lead to augmented simulation, supported by real-life usage data across different scenarios or circumstances, allowing for further adaptation of the design to specific usage needs.

4.2.3 Sensitivities in design optimization

The reliability of the optimization solution is critically dependent on the reliability of the forward solver. The forward solver analyzes the response of the design and provides the critical sensitivity information to the optimizer to search for improving a design or configuration. Therefore, the robustness of the solver convergence is a key enabler. Many forward solvers are sensitive to numerical parameter settings and discretization, which may not converge to an accurate enough solution as the design/configuration changes. As the design changes, the solver or governing equations or assumptions can change and the analysis results are meaningless. As an optimizer searches a wide range of design space going through significant design changes, there remains challenges in ensuring that the forward solvers can guarantee to provide reliable solutions. In addition, the optimization search is critically dependent on the sensitivity therefore, the accuracy, continuity and numerical errors can fundamentally influence the resulting design. The current state of the art solvers are not usually well-equipped to provide reliable sensitivity and have limited understanding of their errors' influence on optimization. There is a need for research in computing reliable sensitivity (which is not necessarily the same as the traditional computational mechanics of predicting a specific response).

4.2.4 Multifidelity and uncertainty management propagation in design optimization

One prominent emerging concept across all engineering disciplines is digital engineering via digital twin. A digital twin is defined as "a virtual representation of a connected asset"[1], with an aim to predict the physical asset's behavior via computational models. Modern engineering systems however, are complex in nature, in which there are many components across a range of scales and their behavior is intrinsically unknown due to the interdependencies and

[1] Digital Twin: Definition and Value, AIAA and AIA position paper, Dec 2020, https://www.aia-aerospace.org/report/digital-twin-paper/

nonlinear interactions. The computational capabilities today are developed to model a specific behavior at one or two scales and the available computational resources are far from being able to model all complex behavior with emergence and nonlinearity across all scales and all governing physics. Indeed, there are many responses and behaviors are we still do not understand and are unable to model. In order to construct and utilize a digital twin to predict unintended consequences, therefore, it is imperative that we have multifidelity models that can integrate and propagate the high order effects and uncertainties across disciplinary and scale boundaries. In the context of design optimization, an accurate prediction of unintended consequences and failure mechanisms is a critical requirement of a digital twin.

4.2.5 Design optimization methods

One fundamental challenge in design optimization arises from design parameterization/representation which defines the design variables. The design variables and their relationship to the governing equations and functional objective/constraints define whether the design space is continuous or discontinuous, convex, multi-modal, ill-posed and whether the existing optimization methods can find an optimum solution. Therefore, research is needed to formulate the geometry and design representation, and to efficiently explore and research the associated design spaces. Today's typical engineering systems are almost always ill-defined and highly complex thus, hence it goes without saying that high performance computing, efficient data structure and parallel algorithms are underlying enablers for design optimization. It is also important that the parameterization and data-structures from design, analysis, manufacturing to operation can be uniquely mapped such that the analyzed and designed and manufactured designs remain consistent. It should be noted that research in efficient design optimization methods would be hugely limited without the simultaneous research in the computational sciences and mechanics research. We recognize that design optimization is not aimed at automating design and taking people out of the design process: Rather, the true purpose is to aid an engineer's design activities. The focus of the design optimization methods research therefore, needs to be on informing an engineer to manage internal requirement conflicts and balance the short- and long-term consequences, offering a tool for investigating the complex design space and the "what-if" scenarios, and providing the necessary data to support engineers' creativity.

4.2.6 Parametric Modelers (PM) in the shape-optimisation loop

Concerning parametric modelers (PM) in the shape optimization loop, we list several topics of interest or desirable features for future research and consideration.

- The capability of a PM to incorporate local and global geometric quantities.
- Parent instances and their impact on the quality of the PM.
- The robustness efficiency of a PM, and specifically estimating the probability of producing non geometrically valid objects.
- The geometric properties of the design space and their influence on its exploration.
- Achieving dimensionality reduction and accelerating Parametric Sensitivity Analysis (PSA) via physics-informed geometric functionals.
- The smooth embeddability and integration of PMs to CFD/FEA solvers.
- Automatic differentiation of QoI (Quantities of Interest) with respect to design parameters.

4.3 Additive Manufacturing

Gershon Elber (Technion – Haifa, IL), Sylvain Lefebvre (LORIA & INRIA – Nancy, FR), Géraldine Morin (IRIT – University of Toulouse, FR), Suraj R. Musuvathy (nTopology – New York, US), Stefanie Hahmann (INRIA Grenoble Rhône-Alpes, FR), Xiaoping Qian (University of Wisconsin – Madison, US), Ernst Rank (TU München, DE), Elissa Ross (Metafold 3D – Toronto, CA), Yongjie Jessica Zhang (Carnegie Mellon University – Pittsburgh, US)

The following list contains talks that took place at the Dagstuhl seminar and relate to the Additive Manufacturing working group (in order of presentation):

1. Stefanie Hahmann. *Geometric construction and fabrication of auxetic metamaterials.*
2. Sylvain Lefebvre. *Generating oriented structures and trajectories within part volumes.*
3. Gershon Elber. *Volumetric Representations: Design, Analysis, Optimization, and Fabrication of Porous/Heterogeneous Artifacts.*
4. Xiaoping Qian. *Topology Optimization for Additive Manufacturing.*
5. Elissa Ross. *3D printed metamaterials in industry.*
6. Yongjie Jessica Zhang. *An Isogeometric Analysis Based Topology Optimization Framework for Additive Manufacturing of 2D Cross-Flow Heat Exchangers.*
7. Ernst Rank. *Interoperability of Geometric Models and Numerical Analysis by Immersed Boundary Methods.*
8. Suraj R. Musuvathy. *Implicit Modeling : Driving A CAD Renaissance.*
9. Géraldine Morin. *Tubular parametric volume objects.*

4.3.1 New opportunities and challenges in AM: overview

AM enables fabrication of shapes of unprecedented complexity, and in particular enables internal structuring of a part interior. This paves the way to the fabrication of volumes embedding microstructures, where small scale geometries directly impact the macro scale physical properties of the part. As 3D printing resolution increases, and as the size of the fabricable objects increases, the micro-structures are becoming akin to a material [3.29], that can be specifically tailored to the object and its future function, including gradients of properties [3.10,3.19,3.6].

This raises novel challenges regarding representation of these geometries [3.6,3.24,3.23] and calls for novel methodologies to analyze and simulate designed parts [3.6,3.26,3.39,3.27]. In particular, design and simulation are merging into a single integrated process, where simulation drives the design to automatically optimize the final parts [3.26,3.39,3.27].

4.3.2 Multiscale modeling, micro-structures as materials

A unique possibility opened by fine scale structuring of a part interior is to enable gradients of internal properties, varying the geometric details such that macro-scale properties change in different locations within the part. Designing such micro-structures requires solving for multiple challenges: the generation techniques have to produce micro-structures triggering the desired behaviors (e.g. anisotropy [3.19], auxeticity [3.10]), have to scale to large volumes, enforce geometric constraints of the target fabrication processes, and offer some degree of control to the designers. Several approaches are considered, such as repeating representative elements, possibly at multiple scales, supported by accurate and efficient volumetric representations [3.6,3.23], that allows, for example, nanostructures inside microstructures, picostructures in nanostructures, etc., limited only by memory and computation abilities. Other techniques rely on implicit definitions to capture highly detailed, unstructured content [3.19,3.24]. A key future challenge is to allow for the efficient simulation of such models, which are diverging from the traditional representations (tetrahedral and hexahedral meshes) used by current state of the art simulation frameworks. Promising directions emerge towards this objective, with simulation approaches that avoid conversions and can work on transient representations of the data [3.27]. Generally, transient representations are created on-the-fly, when needed for display, simulation and production, at the exact resolution they are required. More research is called for at the interaction of these topics, in order to unlock the full potential of these emerging methodologies and develop a complete, novel ecosystem for AM.

4.3.3 Designing and optimizing for final physical properties

Designing with the full potential of AM requires novel workflows supported by computational design. In particular, topology optimization approaches allow optimizing a shape in ways that would be extremely difficult to achieve for a human designer. Such optimization techniques, however, have to be constrained in such a way that they enforce manufacturing constraints, and take them into account to produce parts that are not only optimal for a given function, but also optimal in terms of their fabricability on a target process [3.26,3.39]. Further research is required to further integrate the process in existing topology optimization frameworks, and to make topology optimization amenable to shape representations that are better suited to fabrication and analysis [3.39].

4.3.4 Novel geometric representations and interoperability

As novel representations are developed, based on V-reps [3.6,3.23], implicit functions especially signed distance fields [3.29,3.24] (SDF), or random porous geometries [3.10,3.19], novel challenges appear to allow their interoperability with existing workflows. For instance, conversions from and between such representations often leads to open questions: how to obtain precise SDFs from B-reps or arbitrary implicit formulations and vice-versa, how to obtain quad meshes and surface representations from SDFs, how to obtain volumes from lattice structures [3.23]?

Another question is how to redefine the design engineering processing pipeline of AM (including but not limited to physics simulation and hybrid AM with machining) from these novel representations, or transient representations, avoiding uncontrolled approximations due to conversions. Ultimately, novel industrial standards will have to emerge to support these evolutions.

A major aspect of future fabrication, in addition to porosity, is heterogeneity. AM not only allows one to create highly complex (porous) geometries but also deposit different materials in different places, aka functionally graded materials. Such printers are already out there and yet there is a complete lack of support of such abilities in contemporary CAD software. One possible emerging representation is trivariate based V-reps, that fully encompass the geometry as in B-reps but also allows the tight representation of scalar vector and tensor fields alongside the geometry. Further, this volumetric representation is fully compatible with IGA and hence makes the connection between design and analysis/optimization much tighter, as it should be. Finally, V-reps already support 3D printing of heterogeneous (and porous) geometries. There is also a need for new analysis methods able to deal with graded materials and solve inverse modeling problems.

4.4 Isogeometric Analysis

Carlotta Giannelli (University of Firenze, IT), Panagiotis Kaklis (The University of Strathclyde – Glasgow, GB), Tom Lyche (University of Oslo, NO), Carla Manni (University of Rome "Tor Vergata", IT), Malcolm Sabin (Cambridge, GB), Espen Sande (EPFL – Lausanne, CH), Giancarlo Sangalli (University of Pavia, IT), Hendrik Speleers (University of Rome "Tor Vergata", IT), Deepesh Toshniwal (TU Delft, NL)

In a discussion session on the current status of isogeometric analysis (IGA) and on the interaction between IGA and computer-aided geometric design, we identified future challenges and main areas of interest. This report briefly summarizes this discussion; it is also based on an extended comment by Malcolm Sabin.

At more than 15 years from its inception, IGA is a well-established technology. The field is very active, but it is facing some very difficult challenges and only few industrial inroads have been made.

The original idea of IGA was that the boundary representation (B-rep) of models held in CAD systems could use the basis functions of the NURBS surfaces as the basis for analysis, thus avoiding the need for meshing and remeshing steps along the whole analysis process.

The IGA approach puts B-splines as foundation (basis) for finite element analysis. As a consequence, the accuracy as a function of mesh density is optimal. Moreover, if a model is built primarily for analysis using B-spline elements it will be easy to export to a CAD system for subsequent addition of all the little tweaks which are needed for completion of the design and other downstream reasons.

However, there are some theoretical and practical issues that still need to be pointed out.

- There is a technology gap between 2-variate CAD and 3-variate analysis. Volumetric IGA needs elements fitted to the boundary, which is utterly non-trivial (although there are approaches capable of almost always constructing well-formed hex meshes in a typical B-rep shape). CAD must move to mathematical volumes to properly support 3D IGA. The legacy of existing CAD-models is blocking a move as well.
- In current CAD-systems, representations of real shapes hold rather badly fragmented B-reps. There are a lot more "faces" than the graphics representations on the screen indicate, and they are split in ways which do not really reflect the way that a mesh should flow for good analysis results. Moreover, the mathematics is hidden from the user. This makes bridging analysis and geometry sometimes easier without CAD.
- Finally, and regardless of the aforementioned improvements that should be made to CAD (e.g., 3-variate analysis, tighter integration between mesh ↔ geometry ↔ analysis), another important research direction is the development of IGA approaches for performing analysis on the many existing legacy-CAD-based geometries. These include approaches for watertight reconstructions (i.e., untrimming) as well as IGA in the presence of gaps and overlaps.

As of 2021 the IGA field shows some appreciable progress over the last 4 years.

- There are relevant practical applications where IGA has now proved to be useful as shell modelling, turbulent Navier Stokes or Maxwell's equations (while this was not thought be possible some years ago).
- The IGA approach is embracing more technologies as immersed/embedded methods.

- There is mature software that is publicly available to both the academic and user community. Spline elements are getting into FEM-code (example: LS-DYNA) and there is progress on improving the STEP ISO 10303 standard.

Finally, we report that the wide scientific community is changing its mind about IGA, becoming less skeptical and more positive. In this perspective it is also worth to mention that there are now more (yet still few) joint projects related to IGA funded by industry. Also, in the last 10 years there have been a lot of projects (especially in Europe) training the young generation in IGA before they start working in industry (examples: MSCA ITN networks ARCADES, GRAPES and several ERC projects).

As for future directions we can identify the following items, with the first four issues situate in a closer future while the remaining ones seem to belong to a more distant horizon:

- Many interesting and new spline constructions are available, but it is not yet clear that they would enter the CAD system and help solve the interoperability issue.
- Volume representations (V-reps) still require a lot of investigations and are surely "work in progress".
- Additive manufacturing is a unique opportunity for IGA: a posteriori error estimates can be used in practice to drive refinement and coarsening.
- Address "killer applications" including higher co-dimension contouring, level set methods for topology optimization, simultaneous material/shape optimization, etc.
- Implicit models and embedded IGA have a future but are far away from standards.
- Interaction with machine learning can be profitable: challenges in IGA could be shared with the machine learning community.
- Augmented reality could be a good application of IGA gathering the same basis functions for geometry and analysis.
- The forefront of PDEs on networks could be an area of application for IGA.

Summarizing, all the participants in the focus group agree that IGA can contribute positively to interoperability issues. Besides the need of promotion outside academia, communication of results is critical for interoperability also within academia (between engineers and mathematicians).

4.5 Geometric Machine Learning

Arturs Berzins (SINTEF – Oslo, NO), Ilke Demir (Intel – Hermosa Beach, US), Rene Hiemstra (Leibniz Universität Hannover, DE), Qi-xing Huang (University of Texas – Austin, US), Yang Liu (Microsoft Research – Beijing, CN), Nicolas Mellado (CNRS – Toulouse, FR), Géraldine Morin (IRIT – University of Toulouse, FR), Wenping Wang (Texas A&M University), Juyong Zhang (University of Science & Technology of China – Anhui, CN)

4.5.1 Summary

Machine learning is expected to have a significant impact on the fields of computer aided design, engineering analysis and manufacturing, and the interoperability between these disciplines. In recent years deep structured learning techniques have been wildly successful in several computer vision tasks, speech recognition, and natural language processing, to name a few. Advances in deep learning now progress slowly towards other application fields. Besides enabling new applications within the scope of the product development process, machine learning techniques may augment existing processes, yielding improved interoperability of geometry modeling, analysis, and manufacturing, thereby enabling more efficient design optimization and product development.

Several key challenges are to be resolved in coming years. Application of soft and, in particular, hard constraints into neural network architectures remains a challenging aspect that requires further study. Development of deep learning tools for different geometry representations, including hybrid representations remains an active area of research. The main challenge, however, is to develop one unified theory that encompasses different geometric representations. General and complete theories of geometric machine learning should be developed. Finally, applications where machine learning can boost efficiency and efficacy of the product development process need to be identified. This way more machine learning experts from other fields could start contributing within these areas.

4.5.2 Status of geometric machine learning in 2021

Much progress has been made on deep neural networks, for a range of geometry representations. Explicit representations, including point-clouds and parametric representations such as simplicial meshes, spline surfaces and subdivision surfaces have been considered. Implicit representations, including algebraic surfaces and level-set methods have been investigated. Some of the advantages of explicit and implicit representations have been combined in hybrid representations. Hybrid representations exist in many CAD software tools, addressing limitations of geometric modeling under one representation.

Enforcement of geometric constraints is another area of active research. Soft constraints are well understood and lead to deep learning models with regularization constraints. Imposing hard constraints in current network architectures remains an active area for further research. Examples include equivariant and invariant neural networks. A unified theory that encompasses different geometric representations is still lacking. General and complete theories of geometric machine learning should be developed.

4.5.3 How can machine learning improve the interoperability between geometric modeling, simulation, manufacturing?

Machine learning may enhance geometric modeling, numerical simulation and manufacturing, providing automated procedures that optimize for different end-goals, such as efficiency, accuracy, and durability. An important goal of machine learning is to learn representations from data. With sufficient data machine learning may be used to automate certain manual labor intensive tasks, including geometry clean-up, meshing for engineering analysis, transfer of analysis results back to the geometric modeling process.

Deep learning techniques provide new ways to link classical disciplines, which is hard to achieve using conventional approaches. An example is generation of image descriptions using natural language processing. In the context of geometric modeling, simulation, and manufacturing, deep learning has the potential to establish new links that were previously beyond reach. One example is the use of numerical simulation to generate training data for supervised geometric neural networks. A second example is the development of neural networks that integrate complex user-constraints, e.g. manufacturability constraints, into the geometric modeling process.

We summarize a set of applications where machine learning could be used to augment / improve geometry modeling / isogeometric analysis / additive manufacturing, including the interoperability between these fields:

- Shape registration and analysis
- Reverse engineering and 3D shape generation / scene reconstruction
- Feature learning on surfaces
- Topology optimization
- Generation of quadrature rules for fictitious domain methods
- Learning multiscale models for microstructures.
- Optimization of global quadrilateral parameterizations / meshes
- Deep learning in reduced order modeling
- Learning "cleaning / fixing" of CAD models
- Driving refinement / coarsening of FEA / IGA spaces, domain decomposition
- Machine learning to augment preconditioners and solvers

4.5.4 Challenges in the short-, mid-, and long-term

There are a number of challenges that require attention in the short and mid-term. The goal of machine learning is to learn representations or models from data. Data is application dependent and may be challenging to acquire depending on the application. Availability of open source datasets will benefit the research community. Development of deep learning tools for different geometry representations, including hybrid representations remains an active area of research. For example, recently there has been a lot of progress on neural implicit representations. Some of the new works go beyond traditional geometric modeling tools. The main challenge in the mid- to long-term, however, will be to develop a unified theory of geometric machine learning that encompasses a wide range of geometry representations. Establishing such general theoretical foundations for geometric deep learning, will involve aspects from approximation theory, geometry, topology, optimization, and statistical machine learning.

5 Acknowledgments

Dagstuhl seminars are always a great scientific experience, and it has been a privilege to benefit from this great place and organisation. This Dagstuhl seminar has been very special, because of his hybrid nature and the time where it happened–during the pandemic. The organisers are really grateful to have been given the possibility to have some on-site participants, who were very active in keeping the seminar lively. It did make a great difference of interest with respect to full online conference. We thus want to thank the Dagstuhl staff for welcoming us despite the risks, and for their careful handling of the situation; a special thanks also for the particularly well thought communication in the seminar room that allowed fruitful discussions between online and on-site participant. Also, many thanks to Arturs Berzins and Konstantinos Gavriil, the two junior scientists that made the link to the Dagstuhl technical staff for all practical issues, and also to Konstantinos for handling this report.

Participants

- Arturs Berzins
SINTEF – Oslo, NO
- Tor Dokken
SINTEF – Oslo, NO
- Konstantinos Gavriil
SINTEF – Oslo, NO
- Thomas A. Grandine
Seattle, US
- Stefanie Hahmann
INRIA Grenoble
Rhône-Alpes, FR
- Rene Hiemstra
Leibniz Universität
Hannover, DE
- Bert Jüttler
Johannes Kepler Universität
Linz, AT
- Panagiotis Kaklis
The University of Strathclyde –
Glasgow, GB
- Rimvydas Krasauskas
Vilnius University, LT
- Sylvain Lefebvre
LORIA & INRIA – Nancy, FR
- Tom Lyche
University of Oslo, NO)
- Carla Manni
University of Rome "Tor
Vergata", IT
- Géraldine Morin
IRIT – University of
Toulouse, FR
- Helmut Pottmann
KAUST – Thuwal, SA
- Ulrich Reif
TU Darmstadt, DE
- Espen Sande
EPFL – Lausanne, CH
- Hendrik Speleers
University of Rome "Tor
Vergata", IT
- Deepesh Toshniwal
TU Delft, NL

Remote Participants

- Gudrun Albrecht
Universidad Nacional de
Colombia – Medellin, CO
- Falai Chen
Univ. of Science & Technology of
China – Anhui, CN
- Ilke Demir
Intel – Hermosa Beach, US
- Nira Dyn
Tel Aviv University, IL
- Gershon Elber
Technion – Haifa, IL
- Carlotta Giannelli
University of Firenze, IT
- Ron Goldman
Rice University – Houston, US
- Hans Hagen
TU Kaiserslautern, DE
- Qi-xing Huang
University of Texas –
Austin, US
- Xiaohong Jia
Chinese Academy of Sciences, CN
- H (Alicia) Kim
UC – San Diego, US
- Tae-wan Kim
Seoul National University, KR
- Jiri Kosinka
University of Groningen, NL
- Yang Liu
Microsoft Research – Beijing, CN
- Nicolas Mellado
CNRS – Toulouse, FR
- Suraj R. Musuvathy
nTopology – New York, US
- Francesco Patrizi
MPI für Plasmaphysik –
Garching, DE
- Jörg Peters
University of Florida –
Gainesville, US

- Konrad Polthier
FU Berlin, DE
- Jeff Poskin
The Boeing Company – Seattle, US
- Xiaoping Qian
University of Wisconsin – Madison, US
- Ernst Rank
TU München, DE
- Elissa Ross
Metafold 3D – Toronto, CA
- Malcolm A. Sabin
Cambridge, GB
- Péter Salvi
Budapest University of Technology and Economics, HU
- Maria Lucia Sampoli
University of Siena, IT
- Giancarlo Sangalli
University of Pavia, IT
- Hiromasa Suzuki
University of Tokyo, JP
- Tamas Várady
Budapest University of Technology and Economics, HU
- Nelly Villamizar
Swansea University, GB
- Wenping Wang
Texas A&M University – College Station, US
- Juyong Zhang
Univ. of Science & Technology of China – Anhui, CN
- Yongjie Jessica Zhang
Carnegie Mellon University – Pittsburgh, US

Report from Dagstuhl Seminar 21472

Geometric Logic, Constructivisation, and Automated Theorem Proving

Edited by

Thierry Coquand[1], Hajime Ishihara[2], Sara Negri[3], and Peter M. Schuster[4]

1 **University of Gothenburg, SE,** thierry.coquand@cse.gu.se
2 **JAIST – Ishikawa, JP,** ishihara@jaist.ac.jp
3 **University of Genova, IT,** sara.negri@unige.it
4 **University of Verona, IT,** petermichael.schuster@univr.it

Abstract

At least from a practical and contemporary angle, the time-honoured question about the extent of intuitionistic mathematics rather is to which extent any given proof is effective, which proofs of which theorems can be rendered effective, and whether and how numerical information such as bounds and algorithms can be extracted from proofs. All this is ideally done by manipulating proofs mechanically or by adequate metatheorems, which includes proof translations, automated theorem proving, program extraction from proofs, proof analysis and proof mining. The question should thus be put as: What is the computational content of proofs?

Guided by this central question, the present Dagstuhl seminar puts a special focus on coherent and geometric theories and their generalisations. These are not only widespread in mathematics and non-classical logics such as temporal and modal logics, but also a priori amenable for constructivisation, e.g., by Barr's Theorem, and last but not least particularly suited as a basis for automated theorem proving. Specific topics include categorical semantics for geometric theories, complexity issues of and algorithms for geometrisation of theories including speed-up questions, the use of geometric theories in constructive mathematics including finding algorithms, proof-theoretic presentation of sheaf models and higher toposes, and coherent logic for automatically readable proofs.

Seminar November 21–26, 2021 – http://www.dagstuhl.de/21472

2012 ACM Subject Classification Theory of computation → Constructive mathematics; Theory of computation → Proof theory; Theory of computation → Automated reasoning

Keywords and phrases automated theorem proving, categorical semantics, constructivisation, geometric logic, proof theory

Digital Object Identifier 10.4230/DagRep.11.10.151

Edited in cooperation with Fellin, Giulio

Geometric Logic, Constructivisation, and Automated Theorem Proving, *Dagstuhl Reports*, Vol. 11, Issue 10, pp. 151–172

Editors: Thierry Coquand, Hajime Ishihara, Sara Negri, and Peter M. Schuster

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Thierry Coquand
Hajime Ishihara
Sara Negri
Peter M. Schuster

A central question has remained from the foundational crisis of mathematics about a century ago: What is the extent of intuitionistic mathematics? From a practical angle, the question is to which extent any given proof is effective, which proofs of which theorems can be rendered effective, and whether and how numerical information such as bounds and algorithms can be extracted from proofs. Ideally, all this is treated by manipulating proofs mechanically and/or by adequate proof-theoretic metatheorems (proof translations, automated theorem proving, program extraction from proofs, proof analysis, proof mining, etc.). In this vein, the central question should rather be put as follows: What is the computational content of proofs?

Guided by this form of the central question, the Dagstuhl Seminar 21472 put a special focus on coherent and geometric theories and their generalisations. These indeed are fairly widespread in mathematics and non-classical logics such as temporal and modal logics, a priori amenable for constructivisation in the vein of Barr's Theorem, and particularly suited as a basis for automated theorem proving. Specific topics included categorical semantics for geometric theories, complexity issues of and algorithms for geometrisation of theories with the related speed-up questions, the use of geometric theories in constructive mathematics up to finding algorithms, proof-theoretic presentation of sheaf models and higher toposes, and coherent logic for automated proving.

The Dagstuhl Seminar 21472 attracted researchers and practitioners from all over the world, including participants from various research areas in order to broaden the scope of the seminar and to create connections between communities. The seminar participants presented and discussed their research by means of programmed and ad-hoc talks, and a tutorial on Agda the well developed proof assistant based on dependent type theory – was held over several time slots. Numerous new research directions were developed in small working groups: for example, new perspectives on classifying toposes in algebraic geometry, applications of dynamical methods to quadratic forms, and Zorn induction to capture transfinite methods computationally.

The tireless efforts by Dagstuhl staff notwithstanding, it would not be fair to say that this seminar did not suffer from the pandemic-related travel restrictions by which many invitees were confined to remote participation, which of course made hard if not impossible that they took part at the invaluable informal exchange on-site characteristic for events held at Dagstuhl. Under the given circumstances, however, the seminar was still judged a success by all the participants. Following an unconditional request by many, the organisers intend to propose a follow-up Dagstuhl seminar on a related topic in the near future – if possible, all on-site.

2 Table of Contents

Executive Summary

Thierry Coquand, Hajime Ishihara, Sara Negri, and Peter M. Schuster 152

Overview of Talks

Progress and challenges in program extraction
Ulrich Berger 155

Loop-checking and the uniform word problem for join-semilattices with an inflationary endomorphism
Marc Bezem 155

Bridging the foundational gap: updating algebraic geometry in face of current challenges regarding formalizability, constructivity and predicativity
Ingo Blechschmidt 156

An automated method to reasoning about differentiable functions
Gabriele Buriola 157

Deductive systems and Grothendieck topologies
Olivia Caramello 158

A General Glivenko–Gödel Theorem for Nuclei
Giulio Fellin and Peter M. Schuster 158

Proof mining a nonlinear ergodic theorem for Banach spaces
Anton Freund 158

Conservation theorems on semi-classical arithmetic
Makoto Fujiwara 159

Gluing classifying toposes along open subtoposes
Matthias Hutzler 159

Negative Results in Universal Proof Theory
Rosalie Iemhoff 160

Theorem Proving as Constraint Solving with Coherent Logic
Predrag Janicic and Julien Narboux 161

Proof mining in nonconvex optimization
Ulrich Kohlenbach 161

Geometric theories versus Grothendieck toposes, questions w.r.t. a possible constructive elementary approach
Henri Lombardi 162

Verifiable Solving of Geometric Construction Problems in the Framework of Coherent Logic
Vesna Marinkovic 163

No speedup for geometric theories
Michael Rathjen 164

Constructiveness and lattices in Lorenzen's work
Stefan Neuwirth 164

The distributivity of the category of dependent objects over the Grothendieck category
Iosif Petrakis 166

Supercompactly generated theories
Morgan Rogers 166

Proofs and computation with infinite data
Helmut Schwichtenberg 166

Coherent logic in representation and proving of informal proofs
Sana Stojanovic-Djurdjevic 167

Terminating sequent calculi for a class of intermediate logics
Matteo Tesi 167

Some remarks about Skolem-Noether Theorem
Thierry Coquand 168

Working groups

Tutorial on Agda, the dependently typed proof assistant
Ingo Blechschmidt and Matthias Hutzler 169

Working group on classifying toposes in algebraic geometry
Ingo Blechschmidt, Ulrik Buchholtz, Matthias Hutzler, Henri Lombardi, and Stefan Neuwirth 169

Zorn Induction
Peter M. Schuster and Ulrich Berger 170

Participants 171

Remote Participants 171

3 Overview of Talks

3.1 Progress and challenges in program extraction

Ulrich Berger (Swansea University, GB)

Program extraction from proofs is a technique to exploit the computational content of constructive proofs to extract programs that are provably correct. The technique builds on realizability as introduced by Kleene and Kreisel in the 1940s and 1950s.

We report about recent progress in program extraction, on the one hand regarding the inclusion of limited forms of nonconstructve principles such as the axiom of choice and the law of excluded middle, on the other hand regarding capturing computations that go beyond the usual functional paradigm, namely nondeterminism and concurrency. It turns out that limited form of the law of excluded middle is required to extract concurrent programs. We report on case studies regarding exact real number computation, infinite Gray code and Gaussian elimination for matrices with exact real number entries. This is joint work Hideki Tsuiki, Dieter Spreen, and Monika Seisenberger.

The main current challenge is the general axiom of choice. Raoult gave a reformulation of the axiom of choice as an induction principle (Open Induction). However, this does not seem to be amenable for program extraction (only restricted forms of Open Induction permit program extraction). In joint work with Peter Schuster, we are currently exploring different formulations of the axiom of choice, in its form as Zorn's Lemma, as 'induction-like principles (Zorn Induction), that might permit program extraction.

3.2 Loop-checking and the uniform word problem for join-semilattices with an inflationary endomorphism

Marc Bezem (University of Bergen, NO)

Main reference Marc Bezem, Thierry Coquand: "Loop-checking and the uniform word problem for join-semilattices with an inflationary endomorphism", Theoretical Computer Science, 2022.
URL http://dx.doi.org/10.1016/j.tcs.2022.01.017

We solve in polynomial time two decision problems that occur in type checking when typings depend on universe level constraints.

3.3 Bridging the foundational gap: updating algebraic geometry in face of current challenges regarding formalizability, constructivity and predicativity

Ingo Blechschmidt (Universität Augsburg, DE)

The Lean community recently reached a major milestone in formalizing the definition of schemes, the objects of study in algebraic geometry. However, their development spans more than 10,000 lines of code. A fundamental notion such as that of schemes should not be such demanding to formalize.

We argue that this defect is due to the reliance on transfinite methods in the classical presentation of the foundations of algebraic geometry, which the Lean community decided to follow. Just as they are inappropriate from a constructive and predicative point of view, they don't provide a good basis for formalization. In fact, those three concerns are closely related, perhaps even sides of the same coin.

The talk explores the tension between the foundation of algebraic geometry and these modern challenges, and reports on work in progress recasting the foundation of algebraic geometry to face these challenges, including a constructive and predicative framework for setting up cohomology of quasicoherent sheaves.

References

1 M. Barakat and M. Lange-Hegermann. An axiomatic setup for algorithmic homological algebra and an alternative approach to localization. *J. Algebra Appl.*, 10(2):269–293, 2011.

2 M. Barr. Toposes without points. *J. Pure Appl. Algebra*, 5:265–280, 1974.

3 A. Blass. Injectivity, projectivity, and the axiom of choice. *Trans. Amer. Math. Soc.*, 255:31–59, 1979.

4 I. Blechschmidt. Flabby and injective objects in toposes, 2021.

5 M. Brandenburg. *Tensor categorical foundations of algebraic geometry.* PhD thesis, Universität Münster, 2014.

6 J. Cole. The bicategory of topoi and spectra. *Repr. Theory Appl. Categ.*, (25):1–16, 2016.

7 T. Coquand. Computational content of classical logic. In A. Pitts and P. Dybjer, editors, *Semantics and Logics of Computation*, pages 33–78. Cambridge University Press, 1997.

8 T. Coquand, H. Lombardi, and P. Schuster. The projective spectrum as a distributive lattice. *Cah. Topol. Géom. Différ. Catég.*, 48(3):220–228, 2007.

9 T. Coquand, H. Lombardi, and P. Schuster. Spectral schemes as ringed lattices. *Ann. Math. Artif. Intell.*, 56:339–360, 2009.

10 A. Grothendieck. Introduction to functorial algebraic geometry, part 1: affine algebraic geometry (lecture notes by F. Gaeta), 1973.

11 M. Hakim. *Topos annelés et schémas relatifs*, volume 64 of *Ergeb. Math. Grenzgeb.* Springer, 1972.

12 G. Kempf. Some elementary proofs of basic theorems in the cohomology of quasi-coherent sheaves. *Rocky Mountain J. Math.*, 10(3):637–646, 1980.

13 H. Lombardi and C. Quitté. *Commutative Algebra: Constructive Methods.* Springer, 2015.

14 M. Maietti. Joyal's arithmetic universes as list-arithmetic pretoposes. *Theory Appl. Categ.*, 23(3):39–83, 2010.

15 C. McLarty. What does it take to prove Fermat's last theorem? Grothendieck and the logic of number theory. *Bull. Symbolic Logic*, 16(3):359–377, 2010.

16 C. McLarty. The large structures of Grothendieck founded on finite-order arithmetic. *Rev. Symbolic Logic*, 13(2):296–325, 2020.

17 S. Posur. A constructive approach to Freyd categories. *Appl. Categ. Structures*, 29:171–211, 2021.

18 P. Schuster. Formal zariski topology: positivity and points. *Ann. Pure Appl. Logic*, 137(1):317–359, 2006.

19 M. Tierney. On the spectrum of a ringed topos. In A. Heller and M. Tierney, editors, *Algebra, Topology, and Category Theory. A Collection of Papers in Honor of Samuel Eilenberg*, pages 189–210. Academic Press, 1976.

20 S. Vickers. *Locales and Toposes as Spaces*, pages 429–496. Springer, 2007.

3.4 An automated method to reasoning about differentiable functions

Gabriele Buriola (University of Verona, IT)

Joint work of Gabriele Buriola, Domenico Cantone, Gianluca Cincotti, Eugenio G. Omodeo, Gaetano T. Spartà
Main reference Gabriele Buriola, Domenico Cantone, Gianluca Cincotti, Eugenio G. Omodeo, Gaetano T. Spartà: "A Decidable Theory of Differentiable Functions with Convexities and Concavities on Real Intervals", in Proc. of the 35th Italian Conference on Computational Logic – CILC 2020, Rende, Italy, October 13-15, 2020, CEUR Workshop Proceedings, Vol. 2710, pp. 231–247, CEUR-WS.org, 2020.
URL http://ceur-ws.org/Vol-2710/paper15.pdf

This contribution concerns an enrichment of pre-existing decision algorithms, which in their turn augmented a fragment of Tarski's elementary algebra with one-argument real functions endowed with continuous first derivative. In its present (still quantifier-free) version, our decidable language embodies addition of functions; the issue we address is the one of satisfiability. As regards real numbers, individual variables and constructs designating the basic arithmetic operations are available, along with comparison relators. As regards functions, we have another sort of variables, out of which compound terms are formed by means of constructs designating addition and – outermostly – differentiation. An array of predicates designate various relationships between functions, as well as function properties, that may hold over intervals of the real line; those are: function comparisons, strict and nonstrict monotonicity / convexity / concavity, comparisons between the derivative of a function and a real term. Our decision method consists in preprocessing the given formula into an equi-satisfiable quantifier-free formula of the elementary algebra of real numbers, whose satisfiability can then be checked by means of Tarski's decision method. No direct reference to functions will appear in the target formula, each function variable having been superseded by a collection of stub real variables; hence, in order to prove that the proposed translation is satisfiability-preserving, we must figure out a flexible-enough family of interpolating C^1 functions that can accommodate a model for the source formula whenever the target formula turns out to be satisfiable.

3.5 Deductive systems and Grothendieck topologies

Olivia Caramello (University of Insubria – Como, IT)

Main reference Olivia Caramello: "Theories, Sites, Toposes: Relating and studying mathematical theories through topos-theoretic 'bridges' " Oxford University Press, 2017.
URL https://doi.org/10.1093/oso/9780198758914.001.0001

I will show that the classical proof system of geometric logic over a given geometric theory is equivalent to new proof systems based on the notion of Grothendieck topology. These equivalences result from a proof-theoretic interpretation of the duality between the quotients of a given geometric theory and the subtoposes of its classifying topos. Interestingly, these alternative proof systems turn out to be computationally better-behaved than the classical one for many purposes, as I will illustrate by discussing a few selected applications.

3.6 A General Glivenko–Gödel Theorem for Nuclei

Giulio Fellin (University of Verona, IT) and Peter M. Schuster (University of Verona, IT)

Main reference Giulio Fellin, Peter Schuster: "A General Glivenko-Gödel Theorem for Nuclei", in Proc. of the Proceedings 37th Conference on Mathematical Foundations of Programming Semantics, MFPS 2021, Hybrid: Salzburg, Austria and Online, 30th August – 2nd September, 2021, EPTCS, Vol. 351, pp. 51–66, 2021.
URL http://dx.doi.org/10.4204/EPTCS.351.4

Glivenko's theorem says that, in propositional logic, classical provability of a formula entails intuitionistic provability of double negation of that formula. We generalise Glivenko's theorem from double negation to an arbitrary nucleus, from provability in a calculus to an inductively generated abstract consequence relation, and from propositional logic to any set of objects whatsoever. The resulting conservation theorem comes with precise criteria for its validity, which allow us to instantly include Gödel's counterpart for first-order predicate logic of Glivenko's theorem. The open nucleus gives us a form of the deduction theorem for positive logic, and the closed nucleus prompts a variant of the reduction from intuitionistic to minimal logic going back to Johansson.

The present study was carried out within the projects "A New Dawn of Intuitionism: Mathematical and Philosophical Advances" (John Templeton Foundation, ID 60842) and "Reducing complexity in algebra, logic, combinatorics – REDCOM" ("Ricerca Scientifica di Eccellenza 2018", Fondazione Cariverona); and within GNSAGA of INdAM.

3.7 Proof mining a nonlinear ergodic theorem for Banach spaces

Anton Freund (TU Darmstadt, DE)

Joint work of Anton Freund, Ulrich Kohlenbach
Main reference Anton Freund, Ulrich Kohlenbach: "Bounds for a nonlinear ergodic theorem for Banach spaces. Ergodic Theory and Dynamical Systems", CoRR, Vol, abs/2108.08555, 2021
URL https://arxiv.org/abs/2108.08555

Proof mining uses tools from logic to extract quantitative (and sometimes new qualitative) results from seemingly noneffective proofs in core mathematics (see the textbook by Ulrich Kohlenbach [3]). This talk presents joint work of Kohlenbach and the speaker [1], which is concerned with nonexpansive maps on Banach spaces: by analysing a proof due to Kazuo Kobayasi and Isao Miyadera [2], we obtain a rate of metastability for the strong convergence of Cesàro means. In the talk, we focus on one particular step of the analysis, which deals with a seemingly noneffective use of a limit inferior. This focus allows us to explain fundamental ideas of proof mining by means of a concrete mathematical example.

Both Anton Freund and Ulrich Kohlenbach were supported by the "Deutsche Forschungsgemeinschaft" (DFG, German Research Foundation) – Projects 460597863, DFG KO 1737/6-1 and DFG KO 1737/6-2.

References

1 Anton Freund and Ulrich Kohlenbach, *Bounds for a nonlinear ergodic theorem for Banach spaces*, Ergodic Theory and Dynamical Systems, to appear. Preprint available as `arXiv:2108.08555`.

2 Kazuo Kobayasi and Isao Miyadera, *On the strong convergence of the Césaro means of contractions in Banach spaces*, Proc. Japan Acad. 56 (1980) 245-249.

3 Ulrich Kohlenbach, *Applied Proof Theory: Proof Interpretations and their Use in Mathematics*, Springer Monographs in Mathematics, Springer, Berlin and Heidelberg, 2008.

3.8 Conservation theorems on semi-classical arithmetic

Makoto Fujiwara (Meiji University – Kawasaki, JP)

Joint work of Makoto Fujiwara, Taishi Kurahashi
Main reference Makoto Fujiwara, Taishi Kurahashi: "Conservation theorems on semi-classical arithmetic", CoRR, Vol. abs/2107.11356, 2021
URL https://arxiv.org/abs/2107.11356

It is well-known that classical arithmetic PA is Π_2-conservative over intuitionistic arithmetic HA. Using a generalized negative translation, we relativize this result with respect to theories of semi-classical arithmetic, which lie in-between PA and HA. In particular, it follows from our main result that PA is Π_{k+2}-conservative over HA + Σ_k-LEM where Σ_k-LEM is the low-of-excluded-middle scheme for formulas of Σ_k form.

3.9 Gluing classifying toposes along open subtoposes

Matthias Hutzler (Universität Augsburg, DE)

A geometric theory classified by some Grothendieck topos can be regarded as a syntactic presentation of the theory. In this talk, we consider the question how to construct such a syntactic presentation for a topos from syntactic presentations of a covering family of open subtoposes, and how to capture appropriate additional gluing data in a syntactic way.

Here, extensions (or expansions) of geometric theories, which can add new sorts, symbols and axioms, and which can be regarded as syntactic presentations of geometric morphisms, play an important role. As an instructive example, we construct a geometric theory classified by the big Zariski topos of the projective line, which is covered by two copies of the big Zariski topos of the affine line, both classifying local algebras with one distinguished element.

References

1 O. Caramello. *Theories, Sites, Toposes: Relating and studying mathematical theories through topos-theoretic 'bridges'.* Oxford University Press, 2017.

2 M. Hakim. *Topos annelés et schémas relatifs*, volume 64 of *Ergeb. Math. Grenzgeb.* Springer, 1972.

3 M. Hutzler. Internal language and classified theories of toposes in algebraic geometry. Master's thesis, University of Augsburg, 2018.

4 M. Hutzler. *Syntactic presentations for glued toposes and for crystalline toposes.* PhD thesis, University of Augsburg, 2021.

5 D. Tsementzis. A syntactic characterization of Morita equivalence. *J. Symbolic Logic*, 82(4):1181–1198, 2017.

6 G. Wraith. Generic galois theory of local rings. In M. Fourman, C. Mulvey, and D. Scott, editors, *Applications of sheaves*, volume 753 of *Lecture Notes in Math.*, pages 739–767. Springer, 1979.

3.10 Negative Results in Universal Proof Theory

Rosalie Iemhoff (Utrecht University, NL)

In this talk I explain how a property of logics, such as uniform interpolation, can be used to establish that a logic does not have proof systems of a certain kind, in this case sequent calculi with good structural properties and other desirable qualities. This connection between the properties of a logic and its proof systems is based on a proof method for uniform interpolation that applies to any intermediate, substructural, modal or intuitionistic modal logic that has a sequent calculus of that kind.

Some of the relevant references:

References

1 A. Akbar Tabatabai, R. Iemhoff, and R. Jalali. Uniform Lyndon Interpolation for Basic Non-normal Modal Logics. In: Silva A., Wassermann R., de Queiroz R. (eds) Logic, Language, Information, and Computation. WoLLIC 2021. Lecture Notes in Computer Science, vol 13038, Springer, 2021.

2 I. van der Giessen and R. Iemhoff. Sequent Calculi for Intuitionistic Gödel-Löb Logic, *Notre Dame Journal of Formal Logic* 62(2): 221–246 (May 2021). DOI: 10.1215/00294527-2021-0011

3 R. Iemhoff. Uniform interpolation and the existence of sequent calculi, *Annals of Pure and Applied Logic* 170 (11), 2019, p. 1–37.

3.11 Theorem Proving as Constraint Solving with Coherent Logic

Predrag Janicic (University of Belgrade, RS) and Julien Narboux (University of Strasbourg, FR)

Main reference Predrag Janičić, Julien Narboux, Theorem Proving as Constraint Solving with Coherent Logic, submitted, 2021

We think coherent logic is well suited framework for automatic generation of readable proofs. In contrast to common automated theorem proving approaches, in which the search space is a set of some formulae and what is sought is again a (goal) formula, we propose an approach based on searching for a proof (of a given length) as a whole. Namely, a proof of a formula in a fixed logical setting can be encoded as a sequence of natural numbers meeting some conditions and a suitable constraint solver can find such sequence. The sequence can then be decoded giving a proof in the original theory language. This approach leads to several unique features, for instance, it can provide shortest proofs. We use SAT and SMT solvers for solving sets of constraints. We implemented the proposed method and we present its features, perspectives and performance.

3.12 Proof mining in nonconvex optimization

Ulrich Kohlenbach (TU Darmstadt, DE)

Proof mining uses so-called proof interpretations, such as suitable forms of Gödel's functional interpretation, to extract explicit computational information from given prima facie noneffective proofs in mathematics. In recent years this has been successfully applied in convex optimization with the extraction of effective rates of asymptotic regularity for cyclic projection methods ([3]) and effective rates of metastability for Proximal Point Type algorithms such as PPA and HPPA which approximate zeros of maximally monotone operators ([2, 4, 5]). In order to be able to treat also nonconvex/nonconcave optimization problems one has to generalize the concept of monotone operator. Recently, Bauschke et al. [1] studied as such a generalization so-called comonotone operators. In the case studies [4, 5] of applying proof mining to PPA and HPPA it becomes apparent that the monotonicity of A is used only in a restricted form which makes it easily possible to adopt the extracted bounds as well as the underlying qualitative convergence theorems also to comonotone operators ([6]). This illustrates how proof mining also facilitates the generalization of proofs.

This research is supported by the 'Deutsche Forschungsgemeinschaft' (Project DFG KO 1737/6-2).

References

1 Bauschke, H.H., Moursi, W.A, Wang, X., Generalized monotone operators and their averaged resolvents. Math. Programming 189, pp. 55-74 (2021).
2 Dinis, B., Pinto, P., Quantitative Results on the Multi-Parameters Proximal Point Algorithm. J. Convex Anal. 28, pp. 729-750 (2021)
3 Kohlenbach, U., A polynomial rate of asymptotic regularity for compositions of projections in Hilbert space. Foundations of Computational Mathematics 19, pp. 83-99 (2019).

4 Kohlenbach, U., Quantitative analysis of a Halpern-type proximal point algorithm for accretive operators in Banach spaces. J. Nonlin. Convex Anal. 9, pp. 2125-2138 (2020).

5 Kohlenbach, U., Quantitative results on the proximal point algorithm in uniformly convex Banach spaces. J. Convex Anal. **28**, pp. 11-18 (2021).

6 Kohlenbach, U., On the Proximal Point Algorithm and its Halpern-type variant for generalized monotone operators in Hilbert space. To appear in: Optimization Letters.

3.13 Geometric theories versus Grothendieck toposes, questions w.r.t. a possible constructive elementary approach

Henri Lombardi (University of Franche-Comté – Besancon, FR)

We use the terminology and notations of dynamical theories. See [1, 2, 4, 8, 12, 13, 14].

Dynamical theories, introduced in [8], are a version without logic, purely computational, of geometric theories. See also the paper [1] describing some advantages of this approach, and pioneering articles [19, Sections 1.5 and 4.2], [18] and [11].

Dynamical algebraic structures are explicite in [12, 14] and implicite in [8], where they are described through their presentations. They are also implicite in [13] and, last but not least, in [9, D5], which was a main source: it is possible to compute inside the algebraic closure of a discrete field, even if it is impossible to construct the structure. So it suffices to consider the algebraic closure as a dynamical algebraic structure à la D5 rather than a usual algebraic structure: *lazy evaluation à la D5 gives a constructive semantic for the algebraic closure of a discrete field.*

Since geometric theories, which are concrete objects are closely related to Grothendieck toposes, our aim is to describe, using a constructive external mathematical world à la Bishop ([3]) all the work on toposes in terms of geometric theories and dynamical theories. See related work in [5, 6, 7, 15, 17] and a preliminary draft in [16].

References

1 Bezem, M. and Coquand, T. (2005). Automating coherent logic. In *Logic for programming, artificial intelligence, and reasoning. 12th international conference, LPAR 2005, Montego Bay, Jamaica, December 2–6, 2005. Proceedings*, pages 246–260. Berlin: Springer.

2 Bezem, M. and Coquand, T. (2019). Skolem's theorem in coherent logic. *Fundam. Inform.*

3 Bishop, E. (1967). *Foundations of constructive analysis.* McGraw-Hill, New York.

4 Coquand, T. (2005). A completeness proof for geometrical logic. In *Logic, methodology and philosophy of science. Proceedings of the 12th international congress, Oviedo, Spain, August 2003*, pages 79–89. London: King's College Publications.

5 Coquand, T. and Lombardi, H. (2006). A logical approach to abstract algebra. *Math. Structures Comput. Sci.*, 16(5):885–900.

6 Coquand, T. and Lombardi, H. (2016). Anneaux à diviseurs et anneaux de Krull (une approche constructive). *Comm. Algebra*, 44:515–567.

7 Coquand, T., Lombardi, H., and Quitté, C. (2022). Dimension de Heitmann des treillis distributifs et des anneaux commutatifs. In *Publications Mathématiques de l'Université de Franche-Comté Besançon. Algèbre et théorie des nombres. Années 2003–2006.* Besançon: Laboratoire de Mathématiques de Besançon, 2006, p. 57–100, version corrigée.

8 Coste, M., Lombardi, H., and Roy, M.-F. (2001). Dynamical method in algebra: effective Nullstellensätze. *Ann. Pure Appl. Logic*, 111(3):203–256.

9 Della Dora, J., Dicrescenzo, C., and Duval, D. (1985). About a new method for computing in algebraic number fields. In *EUROCAL '85. Lecture Notes in Computer Science no. 204, (Ed. Caviness B.F.)*, pages 289–290. Springer, Berlin.

10 Kemper, G. and Yengui, I. (2020). Valuative dimension and monomial orders. *J. Algebra*, 557:278–288.

11 Lifschitz, V. (1980). Semantical completeness theorems in logic and algebra. *Proc. Amer. Math. Soc.*, 79(1):89–96.

12 Lombardi, H. (1998). Relecture constructive de la théorie d'Artin-Schreier. *Ann. Pure Appl. Logic*, 91(1):59–92.

13 Lombardi, H. (2002). Dimension de Krull, Nullstellensätze et évaluation dynamique. *Math. Z.*, 242(1):23–46.

14 Lombardi, H. (2006). Structures algébriques dynamiques, espaces topologiques sans points et programme de Hilbert. *Ann. Pure Appl. Logic*, 137(1-3):256–290.

15 Lombardi, H. (2020). Spectral spaces versus distributive lattices: a dictionary. In *Advances in rings, modules and factorizations. Selected papers based on the presentations at the international conference on rings and factorizations, Graz, Austria, February 19–23, 2018*, pages 223–245. Cham: Springer.

16 Lombardi, H. (2021). Théories géométriques pour l'algèbre constructive. `http://hlombardi.free.fr/Theories-geometriques.pdf`.

17 Lombardi, H. and Quitté, C. (2015). *Commutative algebra: constructive methods. Finite projective modules.* Algebra and applications, 20. Springer, Dordrecht. Translated from the French (Calvage & Mounet, Paris, 2011, revised and extended by the authors) by Tania K. Roblot.

18 Matijasevič, J. V. (1975). A metamathematical approach to proving theorems in discrete mathematics. *Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, 49:31–50, 177. Theoretical applications of the methods of mathematical logic, I.

19 Prawitz, D. (1971). Ideas and results in proof theory. In *Proceedings of the Second Scandinavian Logic Symposium (Univ. Oslo, Oslo, 1970)*, pages 235–307. Studies in Logic and the Foundations of Mathematics, Vol. 63. North-Holland, Amsterdam.

3.14 Verifiable Solving of Geometric Construction Problems in the Framework of Coherent Logic

Vesna Marinkovic (University of Belgrade, RS)

Geometry construction problems are one of the longest studied problems in mathematical education. Solving construction problem corresponds to proving constructively a theorem in a coherent logic form. Automated solving of construction problems has been studied, however rarely in rigorous logical terms.

In this talk I will present a formal logical framework describing a traditional four phases process of solving construction problems and a mechanism for automated generation of solutions, both formalized and human-readable. For this purpose a solver for construction problems ArgoTriCS and a prover for coherent logic ArgoCLP (both developed in our research group) are used. It turns out that coherent logic is a natural framework for carrying out two of four phases. In order to obtain proofs as close as possible to ones generated by humans, automatically generated proofs in coherent logic form are simplified using a simplification procedure integrated within ArgoCLP.

3.15 No speedup for geometric theories

Michael Rathjen

Main reference Michael Rathjen: "No speedup for geometric theories", CoRR, Vol. abs/2105.04661v1, 2021, also to appear in: M. Benini, O. Beyersdorff, M. Rathjen, P. Schuster (eds.), Mathematics for Computation, World Scientific Publishing.
URL https://arxiv.org/abs/2105.04661v1
Main reference Michael Rathjen: "Remarks on Barr's Theorem: Proofs in Geometric Theories", pp. 347–374, De Gruyter, 2016.
URL http://dx.doi.org/doi:10.1515/9781501502620-019

Geometric theories based on classical logic are conservative over their intuitionistic counterparts for geometric implications. In the talk I plan to look at two aspects of geometric theories.

The first will be concerned with the cost of transforming a classical proof in a geometric theory into an intuitionistic one. The latter result (sometimes referred to as Barr's theorem) is squarely a consequence of Gentzen's Hauptsatz. Prima facie though, cut elimination can result in superexponentially longer proofs, posing the question of whether this transformation can be achieved in feasibly many steps.

There is also an infinitary version of geometric theories formulated in the logics $L_{\infty\omega}$ with arbitrary infinite (of any set size) disjunctions and conjunctions. These logics are very expressive. I'd like to discuss the constructivity of the proof that classical $L_{\infty\omega}$-proofs of infinite geometrical implications can be turned into intuitionistic proofs. Can this proof be carried out in CZF? An even more basic question presents itself: What is the proper notion of infinite proof? The latter question is also relevant if one works in classical set theory without AC (recall Barwise's completeness theorem). Also the choice of the proof system is relevant, for instance, if one wants to show that if a proof of phi from T exists in a forcing extension then there is also one in the ground model (assuming phi and T are in the latter).

3.16 Constructiveness and lattices in Lorenzen's work

Stefan Neuwirth (University of Franche-Comté – Besancon, FR)

Joint work of Thierry Coquand, Henri Lombardi, Stefan Neuwirth
Main reference "The free pseudocomplemented semilattice generated by a preordered set", work in progress.
URL http://lmb.univ-fcomte.fr/IMG/pdf/the_free_pseudocomplemented_semilattice_generated_by_a_preordered_set.pdf

This is joint work with Thierry Coquand and Henri Lombardi.

Let $(M, \leq_M)$ be a preordered set.

Let us define the free meet-semilattice over M. Let us consider the set H of unordered lists of elements of M, denoted by $a = \alpha_1 \wedge \cdots \wedge \alpha_n$; we shall define a relation $\leq_H$ on H by the following deduction rules:

(1) if $\alpha \leq_M \beta$ then $\alpha \leq_H \beta$;
(2) if $a \leq_H c$ then $a \wedge b \leq_H c$;
(3) if $c \leq_H a$ and $c \leq_H b$ then $c \leq_H a \wedge b$.

It is easy to prove that the converse holds in (3) and that $\leq_H$ is transitive by showing the admissibility of the corresponding deduction rules. Furthermore, as (2) and (3) introduce relations only between elements one of which is a list of at least two elements of M, the converse holds as well in (1): this is a conservativity result.

A meet-semilattice with least element 0 is *pseudocomplemented* if for every b there is c such that $a \wedge b \leq 0$ if and only if $a \leq c$; the element c is denoted by $\bar{b}$. Let us define the free pseudocomplemented meet-semilattice over M. Let us consider the set H generated inductively from M as the set of unordered lists of elements of H or of formal pseudocomplements of elements of H; we shall define a relation $\leq_H$ on H by the deduction rules (1)–(5) with:

(4) if $a \wedge b \leq_H 0$ then $a \leq_H \bar{b}$;
(5) if $a \leq_H b$ then $a \wedge \bar{b} \leq_H c$.

It is easy to prove that the converse holds in (1), (3), (4); but it is quite difficult to prove that $\leq_H$ is transitive.

A meet-semilattice is *σ-complete* if for every sequence $(a_1, a_2, \dots)$ there is a meet $\bigwedge(a_1, a_2, \dots)$. Let us define the free σ-complete pseudocomplemented meet-semilattice over M. Let us consider the set H generated as before but with the additional inductive clause of containing the sequences $(a_1, a_2, \dots)$ of elements of H written as formal meets $\bigwedge(a_1, a_2, \dots)$; we shall define a relation $\leq_H$ on H by the deduction rules (1)–(8) with:

(6) if $a_k \wedge b \leq_H c$ then $\bigwedge(a_1, a_2, \dots) \wedge b \leq_H c$;
(7) if $c \leq_H a_1$, $c \leq_H a_2$, $\dots$, then $c \leq_H \bigwedge(a_1, a_2, \dots)$;
(8) if $a \wedge a \wedge b \leq_H c$ then $a \wedge b \leq_H c$.

It is easy to prove that the converse holds in (1), (3), (4), (7); the proof of the transitivity of $\leq_H$ is much easier here because of the inclusion of the contraction rule (8) among the deduction rules.

The first two constructions appear in Paul Lorenzen's "Algebraische und logistische Untersuchungen über freie Verbände" (1951). The third one appears in his manuscript "Ein halbordnungstheoretischer Widerspruchsfreiheitsbeweis" (1944), in which he explains why this construction is the semilattice counterpart to the proof of consistency of elementary number theory: this theory may be viewed as contained in the free σ-complete pseudocomplemented meet-semilattice over the set M of numerical propositions preordered by material implication. "The fact that the logic calculuses are semilattices or lattices permits a simple logistic application of free lattices" (Lorenzen 1951).

This talk is also an invitation to reflect upon mathematical objects (like the semilattices here) as given dynamically by rules instead of being considered statically as completed totalities.

References

1 Thierry Coquand and Stefan Neuwirth. Lorenzen's proof of consistency for elementary number theory. *Hist. Philos. Logic*, 41(3), 281–290, 2020. arXiv:2006.08996.

2 Paul Lorenzen. Algebraische und logistische Untersuchungen über freie Verbände. *J. Symb. Log.*, 16(2), 81–106, 1951. `http://www.jstor.org/stable/2266681`. Translation by S. Neuwirth: 'Algebraic and logistic investigations on free lattices', 2017, arXiv:1710.08138.

3 Paul Lorenzen. Ein halbordnungstheoretischer Widerspruchsfreiheitsbeweis [A proof of freedom from contradiction within the theory of partial order]. *Hist. Philos. Logic*, 41(3), 265–280, 2020. arXiv:2006.08996. Dual German-English text, edited and translated by Stefan Neuwirth.

3.17 The distributivity of the category of dependent objects over the Grothendieck category

Iosif Petrakis (LMU München, DE)

In [1] and [2] the type-theoretic axiom of choice, or the distributivity of the Π-type over the Σ-type, is translated into Bishop set theory (BST) as the distributivity of the Π-set over the Σ-set. We present this distributivity categorically, as the distributivity of the category of dependent objects over the Grothendieck category. Similarly to the fact that the category of dependent objects is defined through the Grothendieck category and the functor category, in BST the Π-set can be defined through the Σ-set and the function set.

References

1 I. Petrakis: Dependent sums and Dependent Products in Bishop's Set Theory, in P. Dybjer et. al. (Eds) TYPES 2018, LIPIcs, Vol. 130, Article No. 3, 2019.
2 I. Petrakis: *Families of Sets in Bishop Set Theory*, Habilitationsschrift, LMU, Munich, 2020.

3.18 Supercompactly generated theories

Morgan Rogers (University of Insubria – Como, IT)

There are a few standard ways to identify theories classified by a given topos. I discussed how to do so starting from a theory of presheaf type, in the special case of a topos obtained from a principal site, which is to say a site whose covering families are generated by a class of individual covering morphisms, based on the fourth chapter of my forthcoming thesis. I only got as far as presenting the case of topologies on the simplex category, but I illustrated the background principles involved.

3.19 Proofs and computation with infinite data

Helmut Schwichtenberg (LMU München, DE)

Joint work of Helmut Schwichtenberg, Nils Köpp

It is natural to represent real numbers in $[-1, 1]$ by streams of signed digits $-1, 0, 1$. Algorithms operating on such streams can be extracted from formal proofs involving a unary coinductive predicate CoI on (standard) real numbers x: a realizer of $\mathrm{CoI}(x)$ is a stream representing x. We address the question how to obtain bounds for the lookahead of such algorithms: how far do we have to look into the input streams to compute the first n digits of the output stream? We present a proof-theoretic method how this can be done. The idea is to replace the coinductive predicate $\mathrm{CoI}(x)$ by an inductive predicate $\mathrm{I}(x, n)$ with the intended meaning that we know the first n digits of a stream representing x. Then from a formal proof of $\mathrm{I}(x, n+1) \to \mathrm{I}(y, n+1) \to I(1/2(x+y), n)$ we can extract an algorithm for the average function whose lookahead is $n+1$ for both arguments.

3.20 Coherent logic in representation and proving of informal proofs

Sana Stojanovic-Djurdjevic (University of Belgrade, RS)

Main reference Sana Stojanovic Durdevic: "From informal to formal proofs in Euclidean geometry", Ann. Math. Artif. Intell., Vol. 85(2-4), pp. 89–117, 2019.
URL http://dx.doi.org/10.1007/s10472-018-9597-7

There are several different approaches to verification of proofs from mathematical textbooks. I will discuss one idea for using coherent logic for representation of semi-formal textbook proofs. Also, coherent logic vernacular can be used for automatic generation of more detailed proof objects, and eventually generate formal proofs in language of different interactive theorem provers. This approach is tested on two sets of theorem proofs using classical axiomatic system for Euclidean geometry created by David Hilbert, and a modern axiomatic system E created by Jeremy Avigad, Edward Dean, and John Mumma.

References

1 Stojanovic-Djurdjevic, S., From Informal to Formal proofs in Euclidean Geometry, Annals of Mathematics and Artificial Intelligence, Volume 85, pp 89-117, 2019

3.21 Terminating sequent calculi for a class of intermediate logics

Matteo Tesi (Scuola Normale Superiore – Pisa, IT)

Syntactic decision procedures for propositional intuitionistic logic usually exploit a suitably formulated sequent calculus. There are various approaches known in the literature, the reader can see [3] for an extended survey. These systems fail to satisfy one of the following four *desiderata*: 1. a simple termination procedure which does not require a loop-checking, 2. the invertibility of every rule of the calculus which eliminates the need for backtraking, 3. the extraction of a finite countermodel out of a failed proof search and 4. modularity, i.e. the possibility to extend the general methodology to various extension of intuitionistic logic.

We offer a new method based on labelled sequent calculi [2] which meets the *desiderata* listed above. To start with, we propose a variant with respect to the usual Kripke semantics for intuitionistic logic. In particular, we introduce *strict* Kripke models, i.e. models based on finite transitive and irreflexive orders.

The standard truth condition for the implication is replaced by the following. $x \Vdash A \to B$ if and only if the two conditions:

1. If $x \Vdash A$, then $x \Vdash B$
2. For all y (if $x < y$ and $y \Vdash A$, then $y \Vdash B$).

hold. The two semantics are shown to be equivalent and thus intuitionistic propositional logic proves sound and complete with respect to the strict semantics. This is shown using the finite model property for intuitionistic propositional logic [1] and by providing an easy transformation of finite partial orders into finite strict orders and vice versa. We introduce the following abbreviation:

$$x \Vdash A > B \equiv \text{for all } y \text{ (if } x < y \text{ and } y \Vdash A\text{, then } y \Vdash B)$$

and we show that in every strict intuitionistic model condition 2. is equivalent to:

$2'$. For all y (if $x < y$ and $y \Vdash A$ and $y \Vdash A > B$, then $y \Vdash B$)

The new semantics is employed to obtain a labelled sequent calculus $\mathbf{G3I}_<$ in which the rules for the implication → are obtained through those for the new connective $>$. The following rules govern the implication connective:

$$\dfrac{x : A > B, \Gamma \Rightarrow \Delta, x : A \qquad x : B, x : A > B, \Gamma \Rightarrow \Delta}{x : A \to B, \Gamma \Rightarrow \Delta}\, L\to \qquad \dfrac{\Gamma \Rightarrow \Delta, x : A > B \qquad x : A, \Gamma \Rightarrow \Delta, x : B}{\Gamma \Rightarrow \Delta, x : A \to B}\, R\to$$

$$\dfrac{x < y, x : A > B, \Gamma \Rightarrow \Delta, y : A \qquad y : B, x < y, x : A > B, \Gamma \Rightarrow \Delta}{x < y, x : A > B, \Gamma \Rightarrow \Delta}\, L > \qquad \dfrac{x < y, y : A > B, y : A, \Gamma \Rightarrow \Delta, y : B}{\Gamma \Rightarrow \Delta, x : A > B}\, R >, y \text{ fresh}$$

The termination of the calculus $\mathbf{G3I}_<$ is proved by showing that every proof search ends and yields either a proof or a strict countermodel. This gives a completeness result and a decision procedure for intuitionistic logic. The termination depends on the formulation of the rule R> prevents the formation of loops.

Finally, the sequent calculus $\mathbf{G3I}_<$ can be extended with relational rules which preserve the properties of the base system. We focus on the extensions for intermediate logics characterized by classes of frames with a condition of the form $\forall \overline{x} \varphi$ where φ is a quantifier-free formula. The termination strategy encompasses all these systems and so we obtain terminating calculi for intermediate logics with a universal frame condition and the finite model property.

References

1 Chagrov, A., Zakharyaschev, M., *Modal Logic*, Oxford University Press, 1997.

2 Dyckhoff, R., Negri, S., *Proof analysis in intermediate logics*, Archive for Mathematical Logic 51, pp. 71-92, 2012.

3 Dyckhoff, R., *Intuitionistic decision procedures since Gentzen*, in Kahle R., Strahm T., Studer T. (eds) Advances in Proof Theory. Progress in Computer Science and Applied Logic, vol 28. Birkhäuser, Cham., 2016.

4 Negri, S., *Proof analysis in modal logic*, Journal of Philosophical Logic 34, 507, 2005.

3.22 Some remarks about Skolem-Noether Theorem

Thierry Coquand

We discuss a constructive proof of Skolem-Noether Theorem. In particular, the original proof of Skolem was an early example of the technique of Galois descent. This is part of a general constructive study of the theory of central simple algebra.

References

1 Coquand, T., Lombardi, H. & Neuwirth, S. Constructive basic theory of central simple algebras. (2021)

4 Working groups

4.1 Tutorial on Agda, the dependently typed proof assistant

Ingo Blechschmidt (Universität Augsburg, DE) and Matthias Hutzler (Universität Augsburg, DE)

We give an introduction to Agda, a dependently typed proof assistant, loosely following a tutorial by Martín Escardó given at Proof and Computation 2018 in Fischbachau.

References

1 Escardó, M. Introduction to Univalent Foundations of Mathematics with Agda. (2021), https://www.cs.bham.ac.uk/ mhe/HoTT-UF-in-Agda-Lecture-Notes/
2 Wadler, P., Kokke, W. & Siek, J. Programming Language Foundations in Agda. (2020), https://plfa.inf.ed.ac.uk/20.07/

4.2 Working group on classifying toposes in algebraic geometry

Ingo Blechschmidt (Universität Augsburg, DE), Ulrik Buchholtz (TU Darmstadt, DE), Matthias Hutzler (Universität Augsburg, DE), Henri Lombardi (University of Franche-Comté – Besancon, FR), and Stefan Neuwirth (University of Franche-Comté – Besancon, FR)

A logical way to present a Grothendieck topos is to give a geometric theory which is classified by the topos. This point of view originated from Monique Hakim's PhD thesis, in which she determined such syntactic presentations of two important toposes in algebraic geometry, the Zariski topos and the étale topos.

However, for many related toposes in algebraic geometry, similar syntactic presentations are still lacking. This state of affairs only started to change in recent years, when the theories corresponding to the fppf and the surjective topologies and when theories presenting the infinitesimal and the crystalline topos have been determined.

In this working group, we studied several of the remaining toposes, and made progress on several such, namely the cl, cdf, cdp and the f toposes.

References

1 M. Anel. Grothendieck topologies from unique factorisation systems. 2009.
2 I. Blechschmidt. *Using the internal language of toposes in algebraic geometry.* PhD thesis, University of Augsburg, 2017.
3 O. Gabber and S. Kelly. Points in algebraic geometry. 2014.
4 M. Hakim. *Topos annelés et schémas relatifs*, volume 64 of *Ergeb. Math. Grenzgeb.* Springer, 1972.
5 M. Hutzler. Internal language and classified theories of toposes in algebraic geometry. Master's thesis, University of Augsburg, 2018.
6 M. Hutzler. *Syntactic presentations for glued toposes and for crystalline toposes.* PhD thesis, University of Augsburg, 2021.
7 S. Schröer. Points in the fppf topology. 2014.

8 G. Wraith. Generic galois theory of local rings. In M. Fourman, C. Mulvey, and D. Scott, editors, *Applications of sheaves*, volume 753 of *Lecture Notes in Math.*, pages 739–767. Springer, 1979.

4.3 Zorn Induction

Peter M. Schuster (University of Verona, IT) and Ulrich Berger (Swansea University, GB)

We put forward Zorn Induction as a competitor of Raoult's Open Induction [3] in the undertaking to rephrase as classically equivalent but computationally interesting induction principles the minimal (or maximal) element principle known as Zorn's Lemma [7]. As compared to Open Induction, Zorn Induction works with chains rather than directed subsets, and refers to a strict partial order. We expect Zorn Induction to be of use for computation just as is Open Induction [1, 2], also in abstract algebra [5, 4]. A challenge will be how to capture the nondeterministic consequent of Zorn Induction, as this does not fit directly the setting of least fixed points [6].

By a *(strict) partial order* we understand a pair $(X, <)$ where X is a set and $<$ is a transitive and irreflexive relation on X. An element $x \in X$ is a *lower bound* of a subset Y of X if $x < y$ for all $y \in Y$. By $\mathrm{lb}(Y)$ we denote the set of lower bounds of Y. Let $\mathcal{P}$ be a property of subsets of X. We say that a subset A of X is *$\mathcal{P}$-progressive* if, for all subsets Y of X having property $\mathcal{P}$, if A contains all lower bounds of Y, then A contains an element of Y, i.e.,

$$\forall Y \subseteq X(\mathcal{P}(Y) \wedge \mathrm{lb}(Y) \subseteq A \to Y \cap A \neq \emptyset)\,.$$

We now can formulate *Zorn Induction* as the principle

$$\forall A \subseteq X(A \text{ chain-progressive } \to X \subseteq A).$$

With classical logic, Zorn Induction is equivalent to Zorn's Lemma in the form

$$\forall B \subseteq X(B \text{ inductive and nonempty } \to \ B \text{ has a minimal element})$$

where B is *inductive* if every chain contained in B has a lower bound in B, and b is a *minimal element* in B if $b \in B$ and there is no $y \in B$ with $y < b$.

References

1 Berger, U. A computational interpretation of open induction. *Proceedings – Symposium On Logic In Computer Science.* **19** pp. 326 – 334 (2004,8)
2 Coquand, T. A Note on the Open Induction Principle. (1997) `www.cse.chalmers.se/~coquand/open.ps`
3 Raoult, J. Proving Open Properties by Induction. *Inf. Process. Lett.*. **29** pp. 19-23 (1988)
4 Rinaldi, D. & Schuster, P. A universal Krull–Lindenbaum theorem. *Journal Of Pure And Applied Algebra.* **220**, 3207-3232 (2016)
5 Schuster, P. Induction in Algebra: a First Case Study. *Logical Methods In Computer Science.* **9** (2013,9)
6 Tarski, A. A lattice-theoretical fixpoint theorem and its applications.. *Pacific Journal Of Mathematics.* **5**, 285 – 309 (1955)
7 Zorn, M. A remark on method in transfinite algebra. *Bulletin Of The American Mathematical Society.* **41**, 667 – 670 (1935)

Participants

- Karim Johannes Becher
University of Antwerp, BE
- Arnold Beckmann
Swansea University, GB
- Ulrich Berger
Swansea University, GB
- Ulrik Buchholtz
TU Darmstadt, DE
- Gabriele Buriola
University of Verona, IT
- Giulio Fellin
University of Verona, IT
- Anton Freund
TU Darmstadt, DE
- Matthias Hutzler
Universität Augsburg, DE
- Rosalie Iemhoff
Utrecht University, NL
- Ulrich Kohlenbach
TU Darmstadt, DE
- Henri Lombardi
University of Franche-Comté – Besancon, FR
- Julien Narboux
University of Strasbourg, FR
- Stefan Neuwirth
University of Franche-Comté – Besancon, FR
- Eugenio Orlandelli
University of Bologna, IT
- Iosif Petrakis
LMU München, DE
- Morgan Rogers
University of Insubria – Como, IT
- Peter M. Schuster
University of Verona, IT
- Matteo Tesi
Scuola Normale Superiore – Pisa, IT

Remote Participants

- Jan Belle
LMU München, DE
- Marc Bezem
University of Bergen, NO
- Pierre Boutry
INRIA – Sophia Antipolis, FR
- Olivia Caramello
University of Insubria – Como, IT
- Liron Cohen
Ben Gurion University – Beer Sheva, IL
- Thierry Coquand
University of Gothenburg, SE
- Laura Crosilla
University of Oslo, NO
- Tiziano Dalmonte
University of Turin, IT
- Makoto Fujiwara
Meiji University – Kawasaki, JP
- Hajime Ishihara
JAIST – Ishikawa, JP
- Predrag Janicic
University of Belgrade, RS
- Tatsuji Kawai
JAIST – Nomi, JP
- Vesna Marinkovic
University of Belgrade, RS
- Kenju Miyamoto
LMU München, DE
- Sara Negri
University of Genova, IT
- Takako Nemoto
Hiroshima Institute of Technology, JP
- Satoru Niki
Ruhr-Universität Bochum, DE
- Edi Pavlovic
LMU München, DE
- Cosimo Perini Brogi
University of Genova, IT
- Thomas Powell
University of Bath, GB

- Michael Rathjen
University of Leeds, GB
- Helmut Schwichtenberg
LMU München, DE
- Monika Seisenberger
Swansea University, GB
- Sana Stojanovic-Djurdjevic
University of Belgrade, RS
- Daniel Wessel
LMU München, DE
- Chuangjie Xu
fortiss GmbH – München, DE

Report from Dagstuhl Seminar 21481

Secure Compilation

Edited by

David Chisnall[1], Deepak Garg[2], Catalin Hritcu[3], Mathias Payer[4]

1 **Microsoft Research – Cambridge, UK,** david.chisnall@microsoft.com
2 **MPI-SWS – Saarbrücken, DE,** dg@mpi-sws.org
3 **MPI-SP – Bochum, DE,** catalin.hritcu@mpi-sp.org
4 **EPFL – Lausanne, CH,** mathias.payer@nebelwelt.net

Abstract

Secure compilation is an emerging field that puts together advances in security, programming languages, compilers, verification, systems, and hardware architectures in order to devise more secure compilation chains that eliminate many of today's security vulnerabilities and that allow sound reasoning about security properties in the source language. For a concrete example, all modern languages provide a notion of structured control flow and an invoked procedure is expected to return to the right place. However, today's compilation chains (compilers, linkers, loaders, runtime systems, hardware) cannot efficiently enforce this abstraction against linked low-level code, which can call and return to arbitrary instructions or smash the stack, blatantly violating the high-level abstraction. Other problems arise because today's languages fail to specify security policies, such as data confidentiality, and the compilation chains thus fail to enforce them, especially against powerful side-channel attacks. The emerging secure compilation community aims to address such problems by identifying precise security goals and attacker models, designing more secure languages, devising efficient enforcement and mitigation mechanisms, and developing effective verification techniques for secure compilation chains.

This seminar strived to take a broad and inclusive view of secure compilation and to provide a forum for discussion on the topic. The goal was to identify interesting research directions and open challenges by bringing together people working on building secure compilation chains, on designing security enforcement and attack-mitigation mechanisms in both software and hardware, and on developing formal verification techniques for secure compilation.

Seminar November 28–December 3, 2021 – http://www.dagstuhl.de/21481
2012 ACM Subject Classification Security and privacy → Formal security models
Keywords and phrases secure compilation, low-level attacks, source-level reasoning, attacker models, full abstraction, hyperproperties, enforcement mechanisms, compartmentalization, security architectures, side-channels
Digital Object Identifier 10.4230/DagRep.11.10.173
Edited in cooperation with Roberto Blanco (MPI-SP – Bochum, DE)

Secure Compilation, *Dagstuhl Reports*, Vol. 11, Issue 10, pp. 173–204
Editors: David Chisnall, Deepak Garg, Catalin Hritcu, and Mathias Payer
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

David Chisnall (Microsoft Research – Cambridge, UK)
Deepak Garg (MPI-SWS – Saarbrücken, DE)
Catalin Hritcu (MPI-SP – Bochum, DE)
Mathias Payer (EPFL – Lausanne, CH)

Secure compilation is an emerging field that puts together advances in security, programming languages, compilers, systems, verification, and hardware architectures to devise compilation chains that eliminate security vulnerabilities, and allow sound reasoning about security properties in the source language. For example, all modern languages define valid control flows, e.g., calls must always return to the instruction after the calling point, and many security-critical analyses such as data flow analysis rely on programs adhering to these valid control flows. However, today's compilation chains (compilers, linkers, loaders, runtime systems, hardware) cannot efficiently prevent violations of source-level control flows by co-linked low-level code, which can call and return to arbitrary instructions or smash the stack, blatantly violating the high-level abstraction. Other problems arise because languages fail to specify security policies, such as data confidentiality, and the compilation chains thus fail to enforce them, especially against powerful attacks such as those based on side channels. Yet other problems arise because enforcing source-level abstractions requires runtime checks with noticeable overhead, so compilation chains often forego security properties in favor of efficient code. The emerging field of secure compilation aims to address such problems by:

1. **Identifying precise security goals and attacker models.**
   Since there are many interesting security goals and many different kind of attacks to defend against, secure compilation is very diverse. Secure compilation chains may focus on providing (some degree of) type and memory safety for unsafe low-level languages like C and C++, or on providing mitigations that make exploiting security vulnerabilities more difficult. Other secure compilation chains use compartmentalization to limit the damage of an attack to only those components that encounter undefined behavior, or to enforce secure interoperability between code written in a safer language (like Java, C#, ML, Haskell, or Rust) and the malicious or compromised code it links against. Yet another kind of secure compilation tries to ensure that compilation and execution on a real machine does not introduce side-channel attacks.

2. **Designing secure languages.**
   Better designed programming languages and new language features can enable secure compilation in various ways. New languages can provide safer semantics, and updates to the semantics of old unsafe languages can turn some undefined behaviors into guaranteed errors. Components or modules in the source language can be used as units of compartmentalization in the compilation chain. The source language can also make it easier to specify the intended security properties. For instance, explicitly annotating secret data that external observers or other components should not be able to obtain (maybe indirectly through side channels) may give the compilation chain the freedom to more efficiently handle any data that it can deduce is not influenced by secrets.

3. **Devising efficient enforcement and mitigation mechanisms.**
   An important reason for the insecurity of today's compilation chains is that enforcing security can incur prohibitive overhead or significant compatibility issues. To overcome these problems, the secure compilation community is investigating various efficient security

enforcement mechanisms such as statically checking low-level code, compiler optimizations, software rewriting (e.g. software fault isolation), dynamic monitoring, and randomization. Another key enabler is the emergence of new hardware features that enable efficient security enforcement: access checks on pointer dereferencing (e.g. Intel MPX, Hardbound, WatchdogLite, Oracle SSM, SPARC ADI, or HWASAN), protected enclaves (e.g. Intel SGX, ARM TrustZone, Sanctum, or Sancus), capability machines (e.g. CHERI, Arm Morello), or micro-policy machines (e.g. Draper PUMP, Dover CoreGuard). The question is how such features can enable various security features in source languages efficiently, i.e., how hardware extensions can provide enforcement mechanisms for security properties.

4. **Developing effective verification techniques for secure compilation chains.** Criteria for secure compilation are generally harder to prove than compiler correctness. As an example, showing full abstraction, a common criterion for secure compilation, requires translating any low-level context attacking the compiled code to an equivalent high-level context that can attack the original source code. Another example is preservation of secret independent timing even in the presence of side-channels, as required for "constant-time" cryptographic implementations, which can require more complex simulation proofs than for compiler correctness. Finally, scaling such proofs up to even a simple compilation chain for a realistic language is a serious challenge that requires serious proof engineering in a proof assistant.

The Secure Compilation Dagstuhl Seminar 21481 attracted a large number of excellent researchers with diverse backgrounds. The 42 participants (12 on site, 30 remote) represented the programming languages, formal verification, compilers, security, systems, and hardware communities, which led to many interesting points of view and enriching discussions. Due to COVID-19 pandemic-related travel restrictions and uncertainties, many of the participants had to participate remotely using a combination of video conferencing, instant messaging, and ad-hoc gatherings. Despite this mixed environment, discussions thrived. Some of these conversations were ignited by the 5 plenary discussions and the 28 talks contributed by the participants. The contributed talks spanned a very broad range of topics: formalizing ISA security guarantees, hardware-software contracts, detection and mitigation of (micro-architectural) side-channel attacks, securing trusted execution environments, memory safety, hardware-assisted testing, sampled bug detection, formal verification techniques for low-level languages and secure compilation chains, machine-checked proofs, stack safety, integrating hardware-safety guarantees, effective compartmentalization and its enforcement, cross-language attacks, security challenges of software supply chains, capability machines, (over-)aggressive compiler optimizations, concurrency, new programming language abstractions, compositional correct/secure compilation, component safety, compositional verification, contextual and secure refinement, hardening WebAssembly, secure interoperability, (not) forking compilers, interrupts, hardware design, and many more. Talks were interspersed with lively discussions since, by default, each speaker could only use half of the time for presenting and had to use the other half for answering questions and engaging with the audience. Given the high interest spurred by this second edition and the positive feedback received afterwards, we believe that this Dagstuhl Seminar should be repeated in the future, when hopefully all the participants will be able to attend onsite. One important aspect that could still be improved in future editions is spurring more participation from the systems and hardware communities, especially people working at the intersection of these areas and security or formal verification.

2 Table of Contents

Executive Summary

David Chisnall, Deepak Garg, Catalin Hritcu, and Mathias Payer 174

Plenary Discussions

Real-world deployment and remaining frontiers for secure compilation research
Mathias Payer 178

Microarchitectural and side-channel attacks
Marco Guarnieri 179

Designing New Security Architectures and Verifying their Properties
Shweta Shinde 181

Verification techniques for secure compilation
Dominique Devriese 185

Secure interoperability and compartmentalization
David Chisnall 186

Overview of Talks

Enforcement and compiler preservation of fine-grained constant-time policies
Gilles Barthe 190

Formalizing Stack Safety as a Security Property
Roberto Blanco 190

Are Compiler Optimizations Doing it Wrong? An Investigation of Array Bounds Checking Elimination
Stefan Brunthaler 191

Cross-Language Attacks
Nathan Burow 191

Securing Interruptible Enclaved Execution on Small Microprocessors
Matteo Busi 192

Project Verona: An abstract machine allowing partial verification
David Chisnall 192

On information flow preserving refinement
Mads Dam 193

Formalizing ISA security guarantees in the form of universal contracts
Dominique Devriese 193

Proof techniques for secure compilation with memory sharing
Akram El-Korashy 194

Preserving Memory Safety from C to MSWasm
Anitha Gollamudi 194

Contract-aware secure compilation: a foundation for side-channel resistant compilers – Challenges and open questions
Marco Guarnieri 195

Formally verifying a secure compilation chain for unsafe C components
Catalin Hritcu . . . 195

Conditional Contextual Refinement
Chung-Kil Hur . . . 196

CompCertO: Compiling Certified Open C Components
Jérémie Koenig . . . 196

Changing Compilation without Changing the Compiler
Per Larsen . . . 197

WebAssembly as an intermediate language for safe interoperability
Zoe Paraskevopoulou . . . 197

Compositional Secure Compilation against Spectre
Marco Patrignani . . . 197

Automatic inference of effective compartmentalization policies
Mathias Payer . . . 198

Hardware-Software Contracts and Secure Programming
Jan Reineke . . . 199

Hardware-assisted testing in production
Kostya Serebriany . . . 199

Morello status and verification
Peter Sewell . . . 200

A Wishlist for the Next Generation of Trusted Execution Environments
Shweta Shinde . . . 200

Swivel: Hardening WebAssembly against Spectre
Deian Stefan . . . 200

Compiler-based Side Channel Detection and Mitigation
Gang Tan . . . 201

Cerise: Program Verification on a Capability Machine in the Presence of Untrusted Code
Thomas Van Strydonck . . . 201

Software Supply Chains: Challenges and Opportunities
Nikos Vasilakis . . . 201

How we design hardware and what is costs?
Ingrid Verbauwhede . . . 202

Verifying Speculation Security of Processor Implementations
Drew Zagieboylo . . . 202

Participants . . . 203

Remote Participants . . . 203

3 Plenary Discussions

3.1 Real-world deployment and remaining frontiers for secure compilation research

Discussion led by Mathias Payer (EPFL – Lausanne, CH)

In this session we focused on two key topics: real-world deployment and remaining frontiers for secure compilation research. Both topics are challenging, the former focusing on how we can introduce formal methods into the compilation toolchain, making developers more aware of the different advantages, lowering the barrier to entry for using our tools, and addressing practical deployment concerns. The latter focuses on where to go next such as targeting different compilers, better SAT/SMT solving, scalability issues, targeting large code bases, as well as combining formal methods with other techniques.

3.1.1 Topic 1: deployment

We started the discussion by focusing on issues that keep secure compilation techniques from being applied in practice. The discussion focused around software testing but mostly focused on mitigations. Mitigations are defenses that make exploitation of remaining bugs harder. There is an inherent trade-off between the incurred overhead and the effectiveness of the mitigations. Generally speaking, today it is extremely challenging to deploy new mitigations in practice.

Maintaining mitigations comes inherently at higher overheads on engineering and performance. Any mitigation that is added becomes part of the TCB. After being deployed, mitigations will have to be maintained and will increase engineering complexity by, e.g., making debugging and development more challenging. We don't understand the process of how mitigations transition into practice fully yet but they are often implemented by major players such as Linux kernel developers and companies like Microsoft and Google. A recent example of a mitigation that was changed (apart from CFI being deployed broadly) are stack canaries slowly being deprecated in favor of shadow stacks on the Android operating system. This process is showing how challenging sunsetting mitigations actually is.

The true cost of security is not just the overhead of mitigations or the cost of patching but the real cost is the downtime of important systems. For confidentiality (in addition to integrity), this is a big issue, so improvements will be extremely important. Confidentiality attacks can be passive, making them harder and slower to find. The Heartbleed bug was one such confidentiality attack. On one hand, we looked at network captures from the past and did not find any exploitation that happened. On the other hand, people did not update their keys after compromise or updated them wrongly (with weaker keys).

As a community, we need to go beyond either or approaches and work on finding metrics to evaluate the benefits of a mitigation along with coming up with models on how to maintain them as well as deploying them in a hybrid manner to anticipate how they will be sunsetted whenever a stronger and better mitigation comes around (or a bug class disappears).

3.1.2 Topic 2: research frontiers

After extensively talking about deployment problems we started discussing research frontiers on where and how secure compilation can help. The first seminar on secure compilation was right after the disclosure of Spectre and Meltdown. This resulted in speculative execution being a major ad-hoc topic and, three years later it has evolved into a key focus of secure compilation, with a large number of researchers.

The underlying issue is that hardware may not always do what it is supposed to do and even though it gives us guarantees, these guarantees may not hold. One option is to move towards fault tolerant computation where we adjust our assumptions in our models that the hardware may fail. Including this assumption will allow us to give guarantees despite hardware failures. When looking at the different stacks, so far security is dependent on each individual layer without cross-cutting concerns. As a way forward, we need to drive the argument across the different layers and handle security between layers of abstractions.

In addition to side channels, we also looked at integrity violations such as RowHammer. In the past, hardware was modeled for performance. Architects exclusively optimized for performance at lower cost. Without clear benefits, users are unwilling to pay for security, we need to justify why and how this is necessary. Focusing on fault tolerance, redundancy could solve the issue at some constant cost factor. The underlying challenge is what the cost of mitigating the issue would be at the hardware level – is it really 2x or could it be lower? As a follow up question, we wondered if and how we can add security to our performance model (i.e., not just focusing on throughput, latency, power). So far security is neither present nor added and this needs to change.

Another research frontier is specifications at all levels of abstractions. While ARM released some specifications, they can only partially be turned into guarantees and not all of them are in a directly usable state. Evidently, people are not very good at writing specifications (or even writing functional tests). We wondered if we need weaker specifications that nevertheless remain useful for verification or better composition techniques. Deriving better and tighter/more precise specifications will be an interesting research frontier, especially when moving towards side channels and hardware faults.

3.2 Microarchitectural and side-channel attacks

Discussion led by Marco Guarnieri (IMDEA Software – Madrid, ES)

In this session, we focused on microarchitectural and side-channel attacks. We started by discussing "How can we design principled countermeasures and mitigations against these attacks?" The discussion on this point highlighted that a clear attacker model and a precise description of the security-relevant hardware/software interface are needed to design principled mitigations. Next, we discussed the hardware/software interface for security and, in particular, the security guarantees that hardware should provide to software and how to express them. We concluded by discussing whether secure compilation techniques can help in securing the hardware/software interface. We now present a short summary of each discussion point.

3.2.1 How can we design principled countermeasures and mitigations?

We started the discussion by observing that software has only limited control on microarchitectural aspects and side-effects (e.g., using memory fences to limit the scope of speculation or dedicated commands to flush internal caches and buffers). Even with these commands, however, it is often difficult to build *principled* mitigations due to a lack of detailed microarchitectural models.

One possible way forward consists in extending the ISA with dedicated commands for controlling part of the microarchitectural state. The hardware implementation, then, will be in charge of correctly implementing these commands in a secure manner. We discussed two alternative approaches for this:

- A coarse-grained approach where the processor provides two execution modes: a secure execution mode and an insecure execution mode. The former provides strong security against side-channel and microarchitectural attacks (e.g., by disabling several processor optimizations and reducing resource sharing) at the price of a performance overhead, whereas the latter provides no security guarantees. In this case, programmers need to precisely identify how to partition programs into a secure and insecure parts.
- A fine-grained approach where programmers can specify isolation programmatically down to the microarchitectural level. In principle, with more control on the microarchitectural state, programmers could implement secure code with better performance. Secure compilers could also help in correctly enforcing the desired security properties. Programmatic partitioning at microarchitectural level might also have performance benefits. However, properly using microarchitectural partitioning at software-level might be very challenging (opening the door to potential vulnerabilities).

Another important aspect in mitigating microarchitectural attacks is that countermeasures often come with some performance overhead. As a result, designers need to consider the trade-off between security and performance. For this, we need quantitative measures for security. For instance, in hardware power-based side-channel attacks, security is quantified using the the number of samples needed to obtain some information, with higher numbers being better for security. Work on quantitative information flow can provide the theoretical foundations for building these microarchitectural quantitative measures.

3.2.2 What security guarantees should hardware provide to software?

Most of the recent microarchitectural attacks break the *intuitive* assumptions about the security guarantees that hardware should provide to software. For instance, Spectre attacks break the assumption that code is executed following a program's control flow, whereas Rowhammer attacks break assumptions about memory integrity.

A precise specification about the security guarantees that hardware provides to software is needed as a starting point for building secure systems. Such a specification establishes a *contract* between hardware and software.

Recently, there have been several proposals for formalizing such hardware/software contracts. We discussed several aspects of these specifications:

- Existing proposals are rather narrow and they mostly focus on timing-based attacks. It is unclear whether and how these proposals can be extended to other classes of attacks such as power-based side-channel attacks.
- A key point in defining such a specification is identifying the right level of abstraction for microarchiectural components and side-effects. A contract that comes with a detailed microarchitectural model imposes more constraints on hardware designers. At the same time, a more detailed contract can provide information, e.g., a specific cache replacement policy, that programmers can use to build systems with better performance.
- Such specifications need to be easy to use by software-level tools like program analyses and compilers.
- These hardware/software security specifications need to distribute “security obligations” between hardware and software. Additionally, we need ways to change such specifications as hardware and software systems evolve.

3.2.3 How can secure compilation help to secure the hardware/software interface?

We discussed several ways in which secure compilation could help in securing the hardware/software interface.

From a foundational perspective, secure compilation criteria like full abstraction and preservation of specific properties (such as non-interference) can provide inspiration for the formalization of hardware/software contracts for microarchitectural security.

From a practical perspective, compilers can assist in building secure systems since they can inform the hardware about which information is sensitive and should be protected. For instance, in the coarse-grained model mentioned before compilers can help in partitioning the code, whereas in the fine-grained model compilers can help in correctly inserting instructions for partitioning the microarchitectural state.

3.3 Designing New Security Architectures and Verifying their Properties

Discussion led by Shweta Shinde (ETH Zürich, CH)

We have seen a rise in academic and industry-led efforts for building new architectures. These advances were motivated by several reasons such as diminishing returns of Moore's law, emphasis on energy-efficient designs, and opportunities presented by unified memory architectures. In addition to performance improvements, this shift has led to an opening where security-centric thinking can either be tightly coupled or orthogonality added to these new architectures. In this session, we discussed the ramifications of this shift, particularly to assess the possibility of building secure designs with strong security guarantees that are amenable to formal verification.

3.3.1 Are security architectures actually on the rise, and why?

The discussion started by questioning the veracity of the claim that security architectures are on the rise and within the scope of real-world deployments. The hardware design and verification flows have adopted agile deployment pipelines. This allows researchers to come up with new designs and prototype them quickly. First, there are extensions (e.g., CHERI) that can be tested out quickly because they do not disrupt the rest of the flow (e.g., perform operations only while accessing memory). Second, several primitives can be applied by raising the ISA abstraction and adding new instructions without changing the critical parts of the architecture. Such proposals offer easy adoption paths for manufacturers, who are willing to enable security features.

When it comes to deciding if some primitives are inherently suited for hardware or software, there is a risk of pushing the responsibility. In this case, an extreme example, can be viewing the hardware only as a way to achieve high performance; whereas functional security and isolation are solely a responsibility of software. A counter view is to treat hardware as a component that provides a common abstraction such that the functionality can be appropriately leveraged in software. If the software layer has to simulate a different abstraction because the underlying hardware does not natively support it, this can imply that there is indeed a semantic gap that needs to be bridged with better abstractions. The speculative side-channel attacks are a good example of such a mismatch.

3.3.2 Need for First-class Security Primitives

There are two aspects when it comes to designing robust and effective security primitives. First, one can be ambitious and provide elaborate primitives that are indeed useful to solve large classes of attacks. However, if they are not easy to use, adapt, and apply then such primitives are not immediately practical (e.g., homomorphic encryption, oblivious RAM). Second, given an intuitive and useful primitive, should it be implemented and enforced in hardware or software. This is a question of optimizing non-security aspects such as performance and compatibility. So, the answer usually depends on the specific primitive. For example, doing physical memory isolation can be easier at the hardware level because it is simple enough. However, ensuring non-interference might be challenging to achieve purely in hardware.

3.3.3 Risks and Benefits of Adding New Hardware Primitives

It is crucial to be selective with the hardware primitives. Otherwise, we run a risk of having too many primitives in hardware. Such a proliferation in the best case causes fragmentation and in the worst case can lead to poor security of the overall system. Even if a primitive can be implemented in hardware, practical limitations impose several constraints (e.g., power, area, memory latency). Further, hardware development cycles are much longer. A rough estimate is 5 years for prototyping one generation; at least 2-3 generations before the performance characteristics of a new primitive are acceptable and up to the designers' expectations. Lastly, removing features from hardware is non-trivial and expensive. It breaks compatibility, incurs large changes to software and tools, and may severely annoy customers who are vested in using the hardware features. On the upside, hardware has relatively fewer complex interactions and is a good vantage point for certain enforcement. Thus, at least for simple primitives, implementing them in hardware gives one a better chance of getting it right with a high potential impact. To bring in the best of both worlds, the optimal strategy might be to let the hardware provide only bare minimum security functionality. The software can handle the complex aspects with the potential of several cheap design iterations. The only downside of this approach is that one has to then trust the software to use the primitives correctly and guarantee the overall security without leaving gaps that the attacker can exploit. Perhaps, this is where secure compilation can be vital.

3.3.4 Secure Compilation

In the case of secure compilation, one always considers that the attacker operates at the lowest level. But is this a realistic assumption, and if so, is it practical to protect against such an adversary? For example, researchers have demonstrated the practicality of severe attacks (e.g., cache side-channels, fault injection) that are not addressable purely in software. Unfortunately, the layers at which fixes and mitigations could be introduced (i.e., hardware design) are very removed from the layers that suffer directly from poor security (i.e., software). Moreover, in practice, one has to consider what threats are important to the end-user. If the customers express a strong desire for a given feature, then designers can carve a way forward to an industrial evaluation.

3.3.5 Practical Roadblocks for Hardware-enforced Security

Besides the economic incentive, there are several reasons for investigating hardware features for strengthening security. If a small slowdown in hardware could result in speedups in software, such a slowdown would still be beneficial. However, there are several barriers to making these strides. First, quantifying and associating performance gains to design decisions is not straightforward, especially in complex architectures. Quantifying the costs of a given overhead, not only directly, but how they propagate across pipelines and scale to large numbers of computers is non-trivial. Without concrete and clear root-cause analysis, it is difficult to convince the importance of such changes to the stakeholders. There is no immediate clear path forward from these challenges.

Many performance studies traditionally start with simulations. For example, academic researchers might first build and prototype the smallest steps to validate their ideas and then rely on industry partners to build complete products. Despite the appeal of this approach, the error margin of simulator studies is large enough that statistical noise is in the same order of magnitude, or bigger, than the effects that one intends to measure. Conversely, reasonably accurate simulators are woefully slow. Thus, meaningfully extrapolating insights from prototypes is challenging. In this, as in many other matters, it is clear that early collaborations between academia and industry is crucial.

There are a few lessons that we have learned by building multiple hardware primitives. First, incentives play an important role in the practical adoption of hardware. A clear and drastic reduction in security risks is still a good incentive. However, OS and mobile systems developers do not wish to take responsibility for hardware problems. If we identify a security mechanism or part thereof, that needs to be implemented in hardware, we have to consider several constraints. These include performance, throughput, area, power, and energy. Estimating the exact impact of a mechanism of these factors is an open problem. However, collaborations can help identify non-negotiable constraints early on.

3.3.6 Role of Formal Verification

Formal verification has been traditionally used in security settings to strengthen reasoning. However, verification is also useful for the functional validation of hardware features and reliably projecting costs such as area and timing delays. If one applies rigorous formal modeling, it can open up avenues for simplifying the designs and to make them easy for verification. In addition, such efforts pay off even in the hardware verification phase. There have been several success stories of such a cycle, including virtual memory, protection bits, capability machines, and more recently CHERI. A lot of these success stories have roots in early hardware abstractions for scalability. However, in the last decade or so, there has been a steady rise in the number of purely security-centric hardware extensions. Other than the publicly available extensions that have seen adoption, manufacturers are more receptive to such ideas. It is worth noting that recent extensions are driven by software – techniques that have stood the test of time in software and compiler enforced security are now being moved to hardware for efficiency (e.g., pointer authentication). Moving forward, there are a few key principles that can help accelerate such adoption success. Designers can start with clear specifications and then build the primitives. This opens up an easy path for ISA-level verification at a later stage. It further allows us to cross-check if the hardware implementation indeed adheres to the specification.

3.3.7 Potential Avenues for New Security Primitives

Given the wide-scale hardware support for trusted execution environments (e.g., Intel SGX) and the success of non-TEE primitives such as CHERI, is there any scope for building on these wins? For example, could one re-purpose CHERI to support TEE primitives? If so, would it address problems beyond the scope of TEEs (e.g., side channels)? One has to be careful when it comes to confidentiality, since it requires addressing all possible side-channel attacks. Tangentially, modern cloud deployments use virtual machine and container abstractions. Is there potential for new primitives that are closer to and well-tailored for these abstractions? We have seen that hardware designers are keen on supporting fast virtualization (e.g., Intel VT, ARM CCA) as well as language-specific extensions (e.g., for JavaScript). Are there low-hanging fruits that are within the scope of hardware adoption either based on programming models (e.g., secure language virtual machines) or system abstractions (e.g., library/object isolation)? This requires considerations about developer efforts and ease of using the abstractions correctly. The interface exposed to the developers should not be drastically in contrast to the traditional programming model and should be easy to infer and/or encapsulate at the programming abstraction level (e.g., libraries).

3.3.8 Ramifications for Verification Efforts

Intuitively, adding clean abstractions in hardware does help in overall reasoning. For example, introducing explicit hardware-software contracts allows one to reason about specific types of speculative side-channel attacks. On the other hand, the added complexity of the ISA may impact the proof efforts. For example, in proofs of compartmentalization, the attacker model is arbitrary assembly code. Thus, the security proof has to reason about all possible instructions and their side effects. There are well-known mechanisms to circumvent these challenges. To scale security proofs, designers use simplified models (e.g., infinite memory, unbound number of enclaves). However, one has to be cautious in assessing and closing any gaps between model and reality. One way to do this is by improving the model and checking the assumptions. We have seen that these efforts do pay off (e.g., static analyses are now applied to increasingly larger systems). On the other extreme, it has been shown that substantive formalization and sufficient efforts of end-to-end systems are feasible (e.g., the verified light bulb). With regards to prospects of hardware primitives, free and open-source models (e.g., RISC-V) gives researchers an advantage to easily explore design options without the need to circumvent complex legacy systems or wait for hardware manufacturers to adopt the designs. One opportunity is to build secure compilers from the ground up, in a similar spirit to CompCert, but now for hardware.

3.3.9 Summary

There have been several success stories of introducing new hardware primitives that either indirectly aid or are directly beneficial for improving security. Although this requires substantial efforts from both industry and academia, the overall barrier to entry has and continues to be reduced. The next step toward sustainable security is to make formal verification an integral part of the process. Looking at the evolution of adopting security as a first-class concern, we anticipate a similar journey for formal verification.

3.4 Verification techniques for secure compilation

Discussion led by Dominique Devriese (KU Leuven, BE)

The discussion was centered on three major themes:

(1) what to prove about a secure compiler,
(2) how to prove secure compilation properties, and
(3) machine-checked proofs of secure compilation.

3.4.1 What to prove about a secure compiler?

There are generally two types of properties that one can formally prove about a secure compiler. First: robust preservation of contextual equivalence (full abstraction) or the generalization to robust preservation of hyperproperties on traces of interaction with the outside world. These are robust in the sense that they consider security properties that hold in the presence of an active attacker represented by a program context. Second: in some settings, secure compilation is proven in a non-robust form as preservation of properties on a trace. In such properties, a passive attacker is considered that only interacts with a program through observing or providing inputs on the trace. This is the case, for example, for constant-time preservation in constant-time compcert and the Jasmin compiler. Generally, there is a consensus that it is good that many properties have been proposed so that we can choose the best suited property for a particular system.

It was pointed out that some properties take a kind of hybrid perspective, combining both a passive and active attacker. This is the case, for example, for the typical interpretation of robust non-interference preservation, where the attacker is present as a context, but his/her goal is to learn secrets from the trace of interactions with the outside world. It is not clear whether the attacker should be able to manipulate traces in this model and how secrets may enter a system. Some people think only the context should represent the active attacker and traces and hyperproperties should represent only a success criterion for the attacker. It is not generally clear how contexts (active attackers) can be given all the capabilities of traces.

3.4.2 How to prove secure compilation properties?

We have talked about the fact that weaker secure compilation properties can be easier to prove. For example, back-translations may depend on more information, when proving robust preservation of more restricted classes of hyperproperties. This has been used in some results, but how large the advantage is, can depend on the compiler at hand. For preservation of non-robust properties, one can do without a back-translation. For example, constant-time preservation can be proven using specific forms of simulation cubes and related techniques. Overapproximating an intended security property can also help to simplify proofs (for example a policy on accessing information as an overapproximation for information flow).

A general challenge is the reuse of existing proofs when proving secure compilation. We would like to be able to construct secure compilation proofs in such a way that they can be reused in compiler chains. However, there is also interest in reusing a compiler correctness proof when reasoning about a secure compiler. Some people point out that it can be beneficial to decompose secure compiler (passes) into several smaller secure compiler passes. Combining the results of secure compilation passes may not be obvious when they prove (robust) preservation of different (hyper)properties. A framework like that of CompCertM may help to formulate intended secure compilation results in the presence of different calling conventions. In addition to vertical composability, horizontal composability is an important challenge, where it is not yet entirely clear how this should work.

3.4.3 Machine-checked secure compilation proofs

The final topic of discussion was the machine verification of secure compilation proofs. There are very few papers that go all the way through this. These proofs are inherently hard and it simply follows that mechanizing them is difficult.

It was asked whether sufficient value is attached to machine-verified proofs. Generally, machine verification is regarded as an important extra quality for a paper (or an extra contribution in a journal version), but not a result in itself, even though the amount of effort can be similar to the amount of effort for the paper itself. A mechanization can sometimes simply increase confidence in a result without adding much extra insights, but in some cases, they have been known to uncover important problems that require additional insights to solve.

Given the difficulty of machine-verified secure compilation proofs, some people suggest that we should attempt to build libraries of reusable components and proofs. In the verified compilation community, we have recently started seeing many results building on and improving existing large formalizations (particularly CompCert) rather than starting from scratch. It is not clear whether the field of secure compilation is already sufficiently mature for a similar evolution to take place. To facilitate this further, it would be good if we have reusable proofs of security primitives, common languages to express properties and traces, shared languages for interacting with the outside world, contracts for side-channel leakage etc.

3.5 Secure interoperability and compartmentalization

Discussion led by David Chisnall (Microsoft Research – Cambridge, UK)

Modern programs are typically written in more than one language. There is a growing trend towards writing as much as possible in safe languages to benefit from their extra language-enforced guarantees. These guarantees hold only for code that enforces them. For example, a C♯ or Java program linking to a C library must trust that the C code does not contain any memory-safety bugs because a single memory-safety error can invalidate the invariants in the safe part of the program.

Rewriting all of the existing code in a safe language is usually not economically feasible even in the cases where it is technically possible. Waiting until everything in a program is written in a safe language before being able to claim security properties arising from the safe language is therefore not a valid option. Instead, we would like to explore options for using unsafe code in safe ways. This problem requires exploring multiple levels of the stack, including the safe languages' abstract machines and the OS and hardware mechanisms used for isolation.

This discussion was driven by several high-level questions, covered in the following subsections.

3.5.1 What would you like to see in the abstract machines of new languages to facilitate interoperability?

This discussion was partly related to previous discussion on verification techniques. Linear memory and linear capabilities provide a clear transfer-of-ownership model that is easy to reason about. Once memory has been transferred to an untrusted domain, the trusted portion of a system does not have to enforce any guarantees for it. Transferring memory back is more complicated and requires enforcement mechanisms to prevent the untrusted code being able to access it.

Garbage collection poses some challenges for interoperability. Unsafe code must not be able to materialise pointers to garbage-collected memory and the garbage collector must be able to see and invalidate all pointers in untrusted memory. In the context of a model like that of WebAssembly, is it possible to provide garbage collector as a service and prove (and then rely on) properties of it?

Within the context of language abstractions there is a large open question: what properties should be statically or dynamically enforced and in which contexts? For example, a common compilation target (along the lines of Java or CLR bytecode) that has a strong type system can guarantee a lot of properties (at the expense of requiring that these properties hold for every source language) without the need for dynamic checks. Some dynamic checks can be offloaded to hardware, for example CHERI can dynamically enforce spatial safety even on uncooperative and untrusted machine code. Statically typed assembly languages have a long history and allow interoperating code to make stronger assumptions, should we be advocating for their use in all compilation chains, including those for unsafe languages?

A purely static approach to memory safety would require a very sophisticated type system, which might not even be feasible in the general case. A strong type system might be useful for optimization, where a language that statically guarantees certain properties can elide dynamic checks. For example, a language without linear types would need indirection and dynamic checks to guarantee that it enforced linearity at the boundary. There are still open questions, even assuming the existence of such a type system, that existing code could have dynamic checks inserted to be able to enforce useful properties at the interfaces.

WebAssembly, for example is typed (but its type system is simple) and therefore still needs dynamic checks. This provides some evidence for the amount of typing that compilers from C-like languages are willing to insert. Even simple properties, such as existential types, are probably difficult to drive to universal adoption. Implementing a garbage collector on top of WebAssembly is almost impossible because the type system does not differentiate between pointers and integers. This would become feasible with something like MSWasm, though some experiments implementing garbage collectors on top of CHERI suggest that most C/C++ code is not correct in the presence of copying garbage collection and so one would be restricted to non-moving collectors.

3.5.2 What guarantees in existing languages and abstract machines would you like to be able to protect when composing languages?

For safe composition, the source language needs some kind of many-worlds abstraction. Java has this at the abstract-machine level. The JNI defines an interface that allows Java code to call native code and for the native code to interact with the JVM, but in typical implementations there is no isolation enforced at the boundary. CHERI can be used to retrofit strong enforcement at this kind of boundary but a two-world abstraction doesn't provide any useful fault isolation in the unsafe world: any native code can compromise any other native code.

Memory safety was identified as the key building block for safe composition. Even in the absence of stronger guarantees, ensuring that a C module cannot access any memory that is not explicitly passed to it allows a safe language to make strong guarantees about the damage that a bug in the C code can do. Linearity, though not essential, was the top of the nice-to-have list, giving a simple mechanism for ensuring temporal non-interference between safe and unsafe components.

3.5.3 What features of existing languages would you like to be able to expose in other languages?

As before in this discussion, linearity was top of the list for many participants. Stepping back to think less specifically about source-language properties, there was one high-level guarantee that the participants agreed was critical, the Vegas Principle: What happens in an unsafe language, stays in the unsafe language. More specifically **no sequence of operations in one language may alter state shared between multiple languages in a way that is not possible in all of the sharing languages**.

Implementing secure interoperability is difficult if the abstraction is a *Foreign Function Interface* (FFI), because there is no associated notion of state ownership with a function-call abstraction. Ideally, we would have **Foreign Library Interface** (FLI) where one language could instantiate components in another language and then invoke functions within the scope of that instantiation. This is the model that the picoprocess abstraction, from the Bytecode Alliance, and Project Verona, from Microsoft Research, are attempting to adopt.

A shared-nothing design, such as that used by Erlang's to provide actors written in other languages, is the simplest to secure. All communication requires a copy (at least at the abstract-machine level) and so provides an explicit interception point. This also makes it plausible to compose memory-management policies. Some components can have garbage-collected private memory, others can have manual memory management, reference counting, or linear types. There is no need for a global garbage collector unless a language with automatic memory management can observe all memory.

The CHERI project has done a number of experiments on compartmentalization overheads. In many cases on existing systems, the overhead comes from defensive copies. Hardware acceleration for read-only sharing or ownership transfer would eliminate a lot of these.

3.5.4 What would language designers and implementers like hardware designers to provide?

The participants at this seminar spanned hardware and software and so this discussion provided an opportunity for software-facing researchers and practitioners to present a wishlist of features and for those on the hardware side to provide feedback on their feasibility.

As before, linear types were a popular request. There have been several proposals for linear capabilities on top of CHERI since around 2013. At the hardware level, these are not very difficult. The hardware would clear the tag if a linear capability were loaded, zero the source register on store, and require atomic exchange operations to load a usable linear capability. The problems typically arise in the operating system and compiler. Even in languages with linear types, compilers assume that they can duplicate register values, for example spilling a pointer of a linear type to the stack, using it, and then clobbering it in a subsequent instruction. Similarly, context-switch code, signal, and similar abstractions in the operating system assume that it's safe to load and store register values. Supporting hardware-enforced linear capabilities would require invasive changes to the entire software stack.

There have also been various proposals for restricting information flow in CHERI. The current proposal has a 2-bit information-flow policy, with a local/global split and a store-local permission such that local capabilities can be stored only through a capability with store-local permission. C has thread-local, stack, heap, and global objects with different lifetimes and so this policy is not sufficient for enforcing even this ordering (assuming C code that doesn't store stack pointers on the heap and so on). Capabilities for uninitialized memory that require every memory location to be stored through would help with some categories of vulnerability. These would allow a post-increment store on the address, but not loads until the address reached the end of the object.

There has also been a proposal for store-once capabilities. These contain a one-bit counter that is decremented on store. Attempting to store with the counter cleared would lose the tag bit. This scheme would prevent some exploit techniques. For example, a spilled return capability would be loadable multiple times but could not be stored to overwrite another spilled return capability with a valid capability.

WebAssembly was designed to be efficient to implement on current commodity hardware. The 32-bit memory model is partly accidental: it is easy to enforce on current hardware. Proposals to extend WebAssembly to support a 64-bit address space have suffered from a lack of efficient implementation techniques. Using multiple WebAssembly linear memories has suffered similarly.

In general, it is important to remember one fundamental rule for safe interoperability: **Isolation is easy, (safe) sharing is hard**. Full isolation can be achieved by complete physical partitioning of resources, but useful interoperability requires close communication, which typically implies data sharing (at least at the implementation level, even if the abstract machine describes copies). WebAssembly with WASI, for example, could be trivially implemented with native compilation and running the result in a process using Capsicum sandboxing. This would result in very fast WebAssembly execution but would require a full IPC and context switch for communication with the embedding environment, which would be too slow.

3.5.5 What do people see as the biggest open problems in language interoperability today?

From the formal side of the question, there are still a lot of open areas. If we want to define a formal model and proof of safe interoperability between two languages, is there a general understanding of what the proof structure should look like? Is there anything beyond multi-language semantics, for example capturing properties of the mechanisms used to enforce isolation?

On the practical side, there are very few examples of safe interoperability and most of them are research prototypes. For example, Robusta, Arabica, and CHERI-JNI all provided safe interoperability between Java and native code, but did not escape the lab. The examples that do exist are effectively different syntaxes on the same language. For example, Java and Kotlin both expose the JVM abstract machine, C♯ and F♯ the .NET abstract machine, and TypeScript compiles directly to JavaScript and so supports all of the target's semantics. Are there core abstractions that are both efficient to implement in hardware and provide useful guarantees for software engineering? Are sandboxing solutions part of this problem, providing coarse-grained isolation with explicit sharing?

The whole problem domain of secure compilation can be seen as an extreme case of safe interoperability, between a high-level language and a restricted subset of the target's functionality such that nothing in the output of a compiled program can violate the invariants of the source semantics.

4 Overview of Talks

4.1 Enforcement and compiler preservation of fine-grained constant-time policies

Gilles Barthe (MPI-SP – Bochum, DE)

Joint work of Gilles Barthe, Basavesh Ammanaghatta Shivakumar, Benjamin Gregoire, Vincent Laport, Swarn Priya

The constant-time (CT) policy is an information flow policy used by crypto libraries as a protection against cache-based side-channel attacks. At the core of the CT policy is a baseline leakage model, which assumes that only memory accesses and control flow are leaked. While this leakage model is adequate for analyzing many attacks from the literature: (a) it does not account for time-variable instructions, whose execution time depends on its operands; (b) it excludes real-world code, which uses a weaker leakage model and consequently achieves higher performance. We introduce a general class of fine-grained constant-time policies that supports both weaker and stronger leakage models and their combination. Then, we propose a two-step approach for enforcing fine-grained constant-time policies: first, prove that source programs are constant-time w.r.t. a fine-grained policy using relational Hoare logic, and then prove that compilation preserves constant-time w.r.t. a fine-grained policy. We implement the approach in the Jasmin framework for high-assurance cryptography. We use the framework to verify real-world cryptographic code that was out of the scope of previous approaches.

4.2 Formalizing Stack Safety as a Security Property

Roberto Blanco (MPI-SP – Bochum, DE)

Joint work of Roberto Blanco, Sean Anderson, Leonidas Lampropoulos, Benjamin Pierce, Andrew Tolmach
Main reference Sean Noble Anderson, Leonidas Lampropoulos, Roberto Blanco, Benjamin C. Pierce, Andrew Tolmach: "Security Properties for Stack Safety", CoRR, Vol. abs/2105.00417, 2021.
URL https://arxiv.org/abs/2105.00417

What does "stack safety" mean, exactly? The phrase is associated with a variety of compiler, run-time, and hardware mechanisms for protecting stack memory, but these mechanisms typically lack precise specifications, relying instead on informal descriptions and examples of the bad behaviors that they prevent.

We propose a generic, formal characterization of stack safety based on concepts from language-based security: a combination of an integrity property ("the private state in each caller's stack frame is held invariant by the callee"), and a confidentiality property ("the callee's behavior is insensitive to the caller's private state"), which can optionally be extended with a control flow property.

We use these properties to validate the stack-safety micro-policies proposed by Roessler and DeHon. Specifically, we check (with property-based random testing) that their "eager" micro-policy, which catches violations as early as possible, enforces a simple "stepwise" variant of our properties, and that (a repaired version of) their more performant "lazy" micro-policy enforces a slightly weaker and more extensional observational property. Meanwhile our testing successfully detects violations in several broken variants, including Roessler and DeHon's original lazy policy.

4.3 Are Compiler Optimizations Doing it Wrong? An Investigation of Array Bounds Checking Elimination

Stefan Brunthaler (Universität der Bundeswehr – München, DE)

At the 2018 Secure Compilation meeting, I had several interesting discussions, which eventually lead to the realization that compiler optimizations usually have no stated threat model and are thus assuming overly benign operating conditions that do not withstand scrutiny at a closer look. In this talk, I present my preliminary analysis of a popular array bounds check elimination algorithm, ABCD.

4.4 Cross-Language Attacks

Nathan Burow (MIT Lincoln Laboratory – Lexington, US)

Joint work of Nathan Burow, Hamed Okhravi, Samuel Mergendahl
Main reference Samuel Mergendahl, Nathan Burow, Hamed Okhravi: “Cross-Language Attacks”, Proceedings of the Network and Distributed System Security Symposium (NDSS’22), San Diego, CA, 2022

Memory corruption attacks against unsafe pro- gramming languages like C/C++ have been a major threat to computer systems for multiple decades. Various sanitizers and runtime exploit mitigation techniques have been shown to only provide partial protection at best. Recently developed “safe” programming languages such as Rust and Go hold the promise to change this paradigm by preventing memory corruption bugs using a strong type system and proper compile-time and runtime checks. Gradual deployment of these languages has been touted as a way of improving the security of existing applications before entire applications can be developed in safe languages. This is notable in popular applications such as Firefox and Tor. In this paper, we systematically analyze the security of multi-language applications. We show that because language safety checks in safe languages and exploit mitigation techniques applied to unsafe languages (e.g., Control-Flow Integrity) break different stages of an exploit to prevent control hijacking attacks, an attacker can carefully maneuver between the languages to mount a successful attack. In essence, we illustrate that the incompatible set of assumptions made in various languages enables attacks that are not possible in each language alone. We study different variants of these attacks and analyze Firefox to illustrate the feasibility and extent of this problem. Our findings show that gradual deployment of safe programming languages, if not done with extreme care, can indeed be detrimental to security.

4.5 Securing Interruptible Enclaved Execution on Small Microprocessors

Matteo Busi (University of Pisa, IT)

Joint work of Matteo Busi, Job Noorman, Jo Van Bulck, Letterio Galletta, Pierpaolo Degano, Jan Tobias Mühlberg, Frank Piessens
Main reference Matteo Busi, Job Noorman, Jo Van Bulck, Letterio Galletta, Pierpaolo Degano, Jan Tobias Mühlberg, Frank Piessens: "Securing Interruptible Enclaved Execution on Small Microprocessors", ACM Trans. Program. Lang. Syst., Vol. 43(3), pp. 12:1–12:77, 2021.
URL https://doi.org/10.1145/3470534

Computer systems often provide hardware support for isolation mechanisms like privilege levels, virtual memory, or enclaved execution. Over the past years, several successful software-based side-channel attacks have been developed that break, or at least significantly weaken the isolation that these mechanisms offer. Extending a processor with new architectural or micro-architectural features brings a risk of introducing new software-based side-channel attacks.

In this talk we show how we extended a processor with new features without weakening the security of the isolation mechanisms that the processor offers. Our solution is heavily based on techniques from research on programming languages. More specifically, we propose to use the programming language concept of full abstraction as a general formal criterion for the security of a processor extension. We instantiate the proposed criterion to the concrete case of extending a microprocessor that supports enclaved execution with secure interruptibility. This is a very relevant instantiation as several recent papers have shown that interruptibility of enclaves leads to a variety of software-based side-channel attacks. We propose a design for interruptible enclaves, prove that it satisfies our security criterion and explain how such design drove the actual implementation of an enclave-enabled microprocessor.

4.6 Project Verona: An abstract machine allowing partial verification

David Chisnall (Microsoft Research – Cambridge, UK)

URL https://github.com/Microsoft/Verona

Project Verona is a project by MSR in collaboration with various academic partners to build a new secure programming language for large-scale infrastructure. Verona aims to eliminate any "unsafe" escape hatches so that the type safety and concurrency safety guarantees exist even in the presence of existing C/C++ libraries.

Verona has a "many worlds" abstract machine, providing isolation at the type-system level for units of concurrent execution and for data structures that can be transferred between units of execution. For pure Verona code, we aim to enforce all of these guarantees statically in the compiler, rejecting programs that would violate them. For programs using foreign libraries, we aim to use the type system to define where we must add dynamic checks.

Verona intends to expose instances of foreign libraries as regions, with an isolated memory space (containing the library's heap, stack, globals, and so on) for each instance. This can be enforced with various techniques, such as processes, other MMU-based isolation, SFI, CHERI. The guarantees exposed from Verona to the foreign code are similarly strong: No concurrent access, no out-of-bounds memory accesses. Verona is currently an early-stage research project, this talk aims to provide a taste of the guarantees that we expect to be able to provide to people who will be able to use them as a building block for partially verified systems.

4.7 On information flow preserving refinement

Mads Dam (KTH Royal Institute of Technology – Stockholm, SE)

Joint work of Mads Dam, Christoph Baumann, Hamed nemati, Roberto Guanciale
Main reference Christoph Baumann, Mads Dam, Roberto Guanciale, Hamed Nemati: "On Compositional Information Flow Aware Refinement", in Proc. of the 34th IEEE Computer Security Foundations Symposium, CSF 2021, Dubrovnik, Croatia, June 21-25, 2021, pp. 1–16, IEEE, 2021.
URL https://doi.org/10.1109/CSF51468.2021.00010

Information flow security and refinement have had a troublesome relationship since many years. Refinement injects implementation decisions that in general will cause information content to increase and, as a consequence, can cause information flow properties to be violated. How to address this in a way that supports the many use cases of refinement (changes in data representation, reduction of nondeterminism/underspecification, addition of new observation variables, for instance to reflect low-level features such as caches) has remained open for many years. Building on initial work by Morgan we propose a new approach based on ignorance preservation: A refinement step should be viewed as information flow preserving, if it does not cause observers ignorance to be reduced, for instance by revealing some secret bit of information. In the talk we present the basic epistemic set-up, give some examples, and discuss different proof methods and complications related, in particular, to compositionality.

4.8 Formalizing ISA security guarantees in the form of universal contracts

Dominique Devriese (KU Leuven, BE)

Joint work of Dominique Devriese, Sander Huyghebaert, Steven Keuchel

Where ISA specifications used to be defined in long prose documents, we have recently seen progress on formal and executable ISA specifications. However, for now, formal specifications provide only a functional specification of the ISA, without specifying the ISA's security guarantees. In this paper, we present a novel, general approach to specify an ISA's security guarantee in a way that (1) can be semi-automatically validated against the ISA semantics, producing a mechanically verifiable proof, (2) supports informal and formal reasoning about security-critical software in the presence of adversarial code. Our approach is based on the use of universal contracts: software contracts that express bounds on the authority of arbitrary untrusted code on the ISA. We semi-automatically verify these contracts against existing ISA semantics implemented in Sail using our Katamaran tool: a verified, semi-automatic separation logic verifier for Sail. For now, in this paper, we will illustrate our approach for MinimalCaps: a simplified custom-built capability machine ISA. However, we believe our approach has the potential to redefine the formalization of ISA security guarantees and we will sketch our vision and plans.

4.9 Proof techniques for secure compilation with memory sharing

Akram El-Korashy (MPI-SWS – Saarbrücken, DE)

Joint work of Akram El-Korashy, Adrien Durier, Roberto Blanco, Deepak Garg, Catalin Hritcu, Jeremy Thibault
Main reference Akram El-Korashy, Roberto Blanco, Jérémy Thibault, Adrien Durier, Deepak Garg, Catalin Hritcu: "SecurePtrs: Proving Secure Compilation with Data-Flow Back-Translation and Turn-Taking Simulation", CoRR, Vol. abs/2110.01439, 2021.
URL https://arxiv.org/abs/2110.01439

In two recent pieces of work, we studied techniques for proving two theorems about secure compilation of partial programs (namely, a compiler full abstraction and a preservation of robust safety theorem). Secure compilation of partial programs aims to defend against adversarial contexts (e.g. untrusted libraries). We focus on settings in which the compiled partial program is allowed to share – at run-time – parts of its memory with the context by pointer passing (and the context is also allowed the same).

Proving secure compilation of partial programs typically requires back-translating a target attack against the compiled program to an attack against the source program. To prove this back-translation step, we propose a new technique called data-flow back-translation that is simple, handles unstructured control flow and memory sharing well, and we have proved it correct in Coq.

Our proof techniques work without relying on any assumption about the behavior of the context, but they do rely on target-language support that enforces spatial memory safety. I will present the proof techniques and explain how they allow reusing whole-program compiler correctness proofs. Such reuse is novel, especially for settings with memory sharing, and it is practically desirable in order to avoid redoing laborious proofs should a compiler correctness theorem already exist.

4.10 Preserving Memory Safety from C to MSWasm

Anitha Gollamudi (Yale University – New Haven, US)

Joint work of Anitha Gollamudi, Marco Vassena, Marco Patrignani, Deian Stefan, Craig Disselkoen, Alexandra Michael, Aidan Denlinger, Bryan Parno, Jay Bosamiya, Conrad Watt

WebAssembly (Wasm) has gained traction as the new portable compilation target language for deploying on the web applications written in high-level source languages like C, C++, and Rust. Memory safety is key to the isolation mechanism of the sandboxed execution environment: well-typed programs cannot corrupt the memory outside the sandbox (e.g., the Javascript virtual machine). Unfortunately, Wasm is still insecure: buffer overflows and use-after-free can still corrupt the memory of a program within the sandbox, opening the door to attacks like cross-site scripting and remote code execution.

In this talk we present Memory-Safe WebAssembly (MSWasm), an extension of Wasm with built-in spatial memory safety. That is, any well-typed program in MSWasm is proven to attain spatial memory safety robustly, i.e., even in the presence of arbitrary code the program links against. Additionally, we show that MSWasm can be used as a compilation target for C programs.

Our MSWasm development is built with solid formal foundations: we provide a formal model of MSWasm which we use to prove robust spatial memory safety; we formalise our compiler from (a subset of) C to MSWasm and prove that the compiler is not just correct, but it preserves memory safety of C programs into their MSWasm counterparts; and provide an implementation for C to MSWasm as well as benchmark its efficiency.

4.11 Contract-aware secure compilation: a foundation for side-channel resistant compilers – Challenges and open questions

Marco Guarnieri (IMDEA Software – Madrid, ES)

In the talk, I discussed how compilers can help in building systems that are secure against side-channel and microarchitectural attacks. For this, I presented an overview of hardware-software contracts: an abstraction that captures a processor's security guarantees in a simple, mechanism-independent manner by specifying which program executions a microarchitectural attacker can distinguish. Next, I introduced the idea of contract-aware secure compilation (CASCO). Contract-aware compilers leverage the guarantees expressed in a given contract to generate code that is free from microarchitectural leaks. This enables decoupling program-level security (e.g., ensuring that a password is not leaked under the program semantics), which is the programmer's responsibility, from microarchitectural security (e.g., ensuring that a password is not leaked due to microarchitectural side-effects), which is automatically enforced by compilers. I concluded by discussing challenges and open questions that needs to be solved for building CASCO compilers.

4.12 Formally verifying a secure compilation chain for unsafe C components

Catalin Hritcu (MPI-SP – Bochum, DE)

Joint work of Catalin Hritcu, Arthur Azevedo de Amorim, Roberto Blanco, Akram El-Korashy, Deepak Garg, Marco Patrignani, Jeremy Thibault, Carmine Abate, Ştefan Ciobâcă, Adrien Durier, Boris Eng, Ana Nora Evans, Guglielmo Fachini, Théo Laurent, Benjamin C. Pierce, Marco Stronati, Éric Tanter, Andrew Tolmach

Main reference Guglielmo Fachini, Catalin Hritcu, Marco Stronati, Arthur Azevedo de Amorim, Ana Nora Evans, Carmine Abate, Roberto Blanco, Théo Laurent, Benjamin C. Pierce, Andrew Tolmach: "When Good Components Go Bad: Formally Secure Compilation Despite Dynamic Compromise", CoRR, Vol. abs/1802.00588, 2018.

URL http://arxiv.org/abs/1802.00588

Undefined behavior is widespread in the C language and leads to devastating security vulnerabilities. We study how compartmentalization can mitigate this problem by restricting the scope of undefined behavior both (1) spatially to just the components that encounter undefined behavior and (2) temporally by still providing protection to each component up to the point in time when it encounters undefined behavior and becomes compromised. In this talk, we report on a project that has been ongoing for over 5 years on building a formally verified secure compilation chain for unsafe C components based on a variant of the CompCert compiler and various low-level enforcement mechanisms.

We discuss how far did we get and what were the main challenges we had to overcome: from defining formally what it means for a compilation chain to be secure in this setting, to devising more scalable proof techniques that also allow sharing memory dynamically by passing pointers between components, from mechanizing our proofs in the Coq proof assistant, to supporting multiple enforcement mechanisms such as SFI and a programmable tagged architecture. We conclude with future work specific to our project as well as more general open challenges.

4.13 Conditional Contextual Refinement

Chung-Kil Hur (Seoul National University – Seoul, KR)

Joint work of Chung-Kil Hur, Youngju Song, Minki Cho, Dongjae Lee

Contextual refinement (CR) is one of the standard notions of specifying open programs. CR has two main advantages: (i) (horizontal and vertical) compositionality that allows us to decompose a large contextual refinement into many smaller ones enabling modular and incremental verification, and (ii) no restriction on programming features thereby allowing, e.g., mutual recursive, pointer-value passing, and higher-order functions. However, CR has a downside that it cannot impose conditions on the context since it quantifies over all contexts, which indeed plays a key role in support of full compositionality and programming features.

In this work, we address the problem of finding a notion of refinement that satisfies all three requirements: support of full compositionality, full (sequential) programming features, and rich conditions on the context. As a solution, we propose a new theory of refinement, called CCR (Conditional Contextual Refinement), and develop a verification framework based on it, which allows us to modularly and incrementally verify a concrete module against an abstract module under separation-logic-style pre and post conditions about external modules. It is fully formalized in Coq and provides a proof mode that combines (i) simulation reasoning about preservation of side effects such as IO events and termination and (ii) propositional reasoning about pre and post conditions. Also, the verification results are combined with CompCert, so that we formally establish behavioral refinement from top-level abstract programs, all the way down to their assembly code.

4.14 CompCertO: Compiling Certified Open C Components

Jérémie Koenig (Yale University – New Haven, US)

Joint work of Jérémie Koenig, Zhong Shao
Main reference Jérémie Koenig, Zhong Shao: "CompCertO: compiling certified open C components", in Proc. of the PLDI '21: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, Virtual Event, Canada, June 20-25, 2021, pp. 1095–1109, ACM, 2021.
URL https://doi.org/10.1145/3453483.3454097

Since the introduction of CompCert, researchers have been refining its language semantics and correctness theorem, and used them as components in software verification efforts. Meanwhile, artifacts ranging from CPU designs to network protocols have been successfully verified, and there is interest in making them interoperable to tackle end-to-end verification at an even larger scale. Recent work shows that a synthesis of game semantics, refinement-based methods, and abstraction layers has the potential to serve as a common theory of certified components. Integrating certified compilers to such a theory is a critical goal. However, none of the existing variants of CompCert meets the requirements we have identified for this task. CompCertO extends the correctness theorem of CompCert to characterize compiled program components directly in terms of their interaction with each other. Through a careful and compositional treatment of calling conventions, this is achieved with minimal effort.

4.15 Changing Compilation without Changing the Compiler

Per Larsen (Immunant – Irvine, US)

Joint work of Per Larsen, Andrei Homescu, Stephen Crane

Lots of security research requires changing how compilation is done. For prototyping purposes, this is usually done by downloading the source code of the compiler, etc. If the underlying techniques were ever to be put into practice, we run into the problem that different folks use different compilers. This talk covers a few projects that ran into this challenge and what one can do to avoid the need to customize the compiler. Specifically, I will cover a code randomization project that wraps the linker to rewrite the output of the compiler and a compartmentalization project that rewrites C/C++ headers to avoid modifying the compiler.

4.16 WebAssembly as an intermediate language for safe interoperability

Zoe Paraskevopoulou (Northeastern University – Boston, US)

Joint work of Zoe Paraskevopoulou, Amal Ahmed, Michael Fitzgibbons

In this talk I discussed ongoing work on WebAssembly that focuses on enhancing WebAssembly with capabilities (static and dynamic) in order to facilitate interoperability between languages with different features.

4.17 Compositional Secure Compilation against Spectre

Marco Patrignani (CISPA – Saarbrücken, DE)

Joint work of Marco Patrignani, Marco Guarnieri, Xaver Fabian, Matthis Kruse
Main reference Marco Guarnieri, Marco Patrignani: "Exorcising Spectres with Secure Compilers", CoRR, Vol. abs/1910.08607, 2019.
URL http://arxiv.org/abs/1910.08607

I reported on the CCS'21 paper on the secure compilation against spectre v1 and then talk about how we want to scale these results to v2, v4, v5 and their compositions (e.g., proving that a compiler is secure against v1+v4 simultaneously). Doing this requires reasoning compositionally about robust compilation and what properties we want to preserve, which is an interesting extension of the robust compilation line of work. I also spoke about this compositionality issue in more general terms, trying to generalise these results beyond preservation of Spectre security to the preservation of other security properties.

4.18 Automatic inference of effective compartmentalization policies

Mathias Payer (EPFL – Lausanne, CH)

Severe vulnerabilities are continuously discovered in low level code. Those vulnerabilities threaten the confidentiality and integrity of our systems. Compartmentalization enforces isolation between components and allows breaking up large complex systems into small trust compartments that contain any faults.

While different efficient compartmentalization mechanisms exist, developing effective policies is challenging and generally remains a manual process. In a study on the Linux kernel, we evaluate the feasibility of simple directory-based policies and develop a framework for reasoning about memory accesses during program execution [1].

Shifting towards generating policies, we propose two approaches that leverage the language environment to implement effective compartmentalization. First, HAKCs [2] targets the Linux kernel and allows developers to specify ownership for data along with passing said data between strictly enforced compartments. This explicit ownership model enables efficient checks but revoking privileges for aliased pointers remains challenging. In a prototype targeting the IPv6 module, we demonstrate how such compartmentalization is feasible. Second, in programming environments that heavily rely on third-party libraries, trusting said libraries remains challenging and this trust may be broken by arbitrary updates that are outside of the control of the software developer. With Enclosures [3], we enable flexible closures that bind calls across library boundaries to dynamically created compartments with access to limited system calls. The address space of the process is shared, in part, with the compartment so that data exchange effectively functions.

Finally, we discuss extensions to existing architectures that enable efficient sharing between compartments. Our RISC V prototype leverages a metadata table that defines compartments inside an address space (similar, at a high level, to flexible segments with an inter-segment switching policy), introducing unprivileged instructions that securely switch between compartments.

The discussion centered around effective implementations, creation of strict policies, and limitations of different hardware extensions. Compartmentalization is an active research area that profits from advances in secure compilation in several ways, namely through policy generation, inference of compartment boundaries, and integration as well as activation of the different hardware extensions. Many challenging open questions remain in this active research area.

References

1 uSCOPE: A Methodology for Analyzing Least-Privilege Compartmentalization in Large Software Artifacts. Nick Roessler, Lucas Atayde, Imani Palmer, Derrick McKee, Jai Pandey, Vasileios P. Kemerlis, Mathias Payer, Adam Bates, Andre DeHon, Jonathan M. Smith, and Nathan Dautenhahn. In RAID'21: Recent Advances in Intrusion Detection, 2021

2 Preventing Kernel Hacks with HAKCs. Derrick McKee, Yianni Giannaris, Carolina Ortega, Howard Shrobe, Mathias Payer, Hamed Okhravi, and Nathan Burow. In NDSS'22: Network and Distributed System Security Symposium, 2022

3 Enclosure: language-based restriction of untrusted libraries. Adrien Ghosn, Marios Kogias, Mathias Payer, James R. Larus, and Edouard Bugnion. In ASPLOS'21: International Conference on Architectural Support for Programming Languages and Operating Systems, 2021

4.19 Hardware-Software Contracts and Secure Programming

Jan Reineke (Universität des Saarlandes – Saarbrücken, DE)

Joint work of Jan Reineke, Marco Guarnieri, Boris Köpf, Pepe Vila
Main reference Marco Guarnieri, Boris Köpf, Jan Reineke, Pepe Vila: "Hardware-Software Contracts for Secure Speculation", in Proc. of the 2021 IEEE Symposium on Security and Privacy (SP), pp. 1868–1883, 2021.
URL https://doi.org/10.1109/SP40001.2021.00036

Cache attacks and more recently transient-execution attacks demonstrate that microarchitectural components may leak information in unintended and surprising ways. I will discuss recent work on formally capturing microarchitectural leakage using hardware-software contracts with the goal of enabling secure programming.

4.20 Hardware-assisted testing in production

Kostya Serebriany (Google – Mountain View, US)

Every software vendor is trying to "shift left", i.e. to move bug detection to earlier stages of software development. This is an important goal, which we are unlikely to ever achieve 100%, and thus we need to keep finding bugs when the software is already released. In this talk we will discuss three testing mechanisms that use special hardware features to enable sampled bug detection with near-zero overhead in production:

- GWP-ASan, detects heap corruption with hardware page protection.
- Per-allocation sampling with Arm Memory Tagging Extension.
- GWP-TSan, detects data races using hardware watchpoints.

4.21 Morello status and verification

Peter Sewell (University of Cambridge, UK)

Main reference Thomas Bauereiss, Brian Campbell, Thomas Sewell, Alasdair Armstrong, Lawrence Esswood, Ian Stark, Graeme Barnes, Robert N. M. Watson, Peter Sewell: "Verified security for the Morello capability-enhanced prototype Arm architecture", 2021.
URL https://doi.org/10.48456/tr-959

I gave an update on the state of Morello, the Arm prototype architecture and processor incorporating CHERI hardware capability support, and on our work to verify fundamental properties of the full 62k LoS architecture specification.

4.22 A Wishlist for the Next Generation of Trusted Execution Environments

Shweta Shinde (ETH Zürich, CH)

I highlighted the ongoing initiatives and research directions for building the next generation of trusted execution environments with the primary goal of supporting verified secure software.

4.23 Swivel: Hardening WebAssembly against Spectre

Deian Stefan (University of California – San Diego, US)

Joint work of Deian Stefan, Shravan Narayan, Craig Disselkoen, Daniel Moghimi, Sunjay Cauligi, Evan Johnson, Zhao Gang, Anjo Vahldiek-Oberwagner, Ravi Sahita, Hovav Shacham, Dean Tullsen
Main reference Shravan Narayan, Craig Disselkoen, Daniel Moghimi, Sunjay Cauligi, Evan Johnson, Zhao Gang, Anjo Vahldiek-Oberwagner, Ravi Sahita, Hovav Shacham, Dean Tullsen, Deian Stefan: "Swivel: Hardening WebAssembly against Spectre", in Proc. of the 30th USENIX Security Symposium (USENIX Security 21), pp. 1433–1450, USENIX Association, 2021.
URL https://www.usenix.org/conference/usenixsecurity21/presentation/narayan

We describe Swivel, a new compiler framework for hardening WebAssembly (Wasm) against Spectre attacks. Outside the browser, Wasm has become a popular lightweight, in-process sandbox and is, for example, used in production to isolate different clients on edge clouds and function-as-a-service platforms. Unfortunately, Spectre attacks can bypass Wasm's isolation guarantees. Swivel hardens Wasm against this class of attacks by ensuring that potentially malicious code can neither use Spectre attacks to break out of the Wasm sandbox nor coerce victim code-another Wasm client or the embedding process-to leak secret data. We describe two Swivel designs, a software-only approach that can be used on existing CPUs, and a hardware-assisted approach that uses extension available in Intel 11th generation CPUs. For both, we evaluate a randomized approach that mitigates Spectre and a deterministic approach that eliminates Spectre altogether. Our randomized implementations impose under 10.3% overhead on the Wasm-compatible subset of SPEC 2006, while our deterministic implementations impose overheads between 3.3% and 240.2%. Though high on some benchmarks, Swivel's overhead is still between 9x and 36.3x smaller than existing defenses that rely on pipeline fences.

4.24 Compiler-based Side Channel Detection and Mitigation

Gang Tan (Pennsylvania State University – University Park, US)

Main reference Robert Brotzman, Shen Liu, Danfeng Zhang, Gang Tan, Mahmut Kandemir: "CaSym: Cache Aware Symbolic Execution for Side Channel Detection and Mitigation", in Proc. of the 2019 IEEE Symposium on Security and Privacy (SP), pp. 505–521, 2019.
URL https://doi.org/10.1109/SP.2019.00022
Main reference Robert Brotzman, Danfeng Zhang, Mahmut Taylan Kandemir, Gang Tan: "SpecSafe: detecting cache side channels in a speculative world", Proc. ACM Program. Lang., Vol. 5(OOPSLA), pp. 1–28, 2021.
URL https://doi.org/10.1145/3485506

We describe two systems (CaSym and SpecSafe), which use symbolic execution to detect and mitigate cache-based side channels in software or verify their absence. We will also discuss what components are needed to achieve secure compilation in the presence of side channels caused by conventional or speculative execution.

4.25 Cerise: Program Verification on a Capability Machine in the Presence of Untrusted Code

Thomas Van Strydonck (KU Leuven, BE)

Joint work of Thomas Van Strydonck, Lars Birkedal, Dominique Devriese, Armaël Guéneau, Aïna Linn Georges, Amin Timany, Alix Trieu
URL https://github.com/logsem/cerise

A capability machine is a type of CPU allowing fine-grained privilege separation using capabilities, machine words that represent certain kinds of authority. We present Cerise, a mathematical model and accompanying proof methods that can be used for formal verification of functional correctness of programs running on a capability machine, even when they invoke and are invoked by unknown (and possibly malicious) code. Our work has been entirely mechanized in the Coq proof assistant using the Iris program logic framework. The methodology we present underlies recent work of the authors on formal reasoning about capability machines, but was left somewhat implicit in those publications. This presentation exposes in further details a pedagogical introduction to the methodology, in a simple setting (no exotic capabilities), and starting from minimal examples.

4.26 Software Supply Chains: Challenges and Opportunities

Nikos Vasilakis (MIT – Cambridge, US)

To lower the time and cost of engineering software, developers today use software supply chains of unprecedented scale: It is not uncommon for a modern application to use hundreds or even thousands of third-party dependencies developed by many developers with varying needs, skill levels, care, and intentions. In this talk, I will outline some of the security challenges associated with third-party dependencies, and show how key characteristics of these dependencies enable program-transformation techniques to overcome these challenges while keeping developer burden manageable.

4.27 How we design hardware and what is costs?

Ingrid Verbauwhede (KU Leuven, BE)

I gave basics on what are the hardware design constraints: how we measure performance, throughput, latency.

4.28 Verifying Speculation Security of Processor Implementations

Drew Zagieboylo (Cornell University – Ithaca, US)

Joint work of Drew Zagieboylo, Edward Suh, Andrew Myers

We discuss existing tools for verifying security properties in RTL designs and their applicability to speculation-aware contracts. In particular, we highlight the difficulty of verifying speculation security since it is often dependent upon verifying functional correctness. Existing tools either require significant manual input which cannot easily be re-purposed across designs, or they involve assumptions which are difficult to trust or reason about in complex designs.

We propose integrating domain knowledge of speculative processor design directly into a higher-level hardware description language to simplify correctness and speculative-security reasoning. We hope to limit difficult RTL verification tasks to modularized components that abstract common microarchitectural optimizations such as bypassing, speculation, and instruction re-ordering.

Participants

- Roberto Blanco
MPI-SP – Bochum, DE
- Stefan Brunthaler
Universität der Bundeswehr – München, DE
- Matteo Busi
University of Pisa, IT
- Dominique Devriese
KU Leuven, BE
- Akram El-Korashy
MPI-SWS – Saarbrücken, DE
- Deepak Garg
MPI-SWS – Saarbrücken, DE
- Anitha Gollamudi
Yale University – New Haven, US
- Marco Guarnieri
IMDEA Software – Madrid, ES
- Catalin Hritcu
MPI-SP – Bochum, DE
- Marco Patrignani
CISPA – Saarbrücken, DE
- Jan Reineke
Universität des Saarlandes – Saarbrücken, DE
- Shweta Shinde
ETH Zürich, CH
- Jeremy Thibault
MPI-SP – Bochum, DE
- Thomas Van Strydonck
KU Leuven, BE
- Ingrid Verbauwhede
KU Leuven, BE

Remote Participants

- Amal Ahmed
Northeastern University – Boston, US
- Arthur Azevedo de Amorim
Boston University, US
- Gilles Barthe
MPI-SP – Bochum, DE
- Joseph Bialek
Microsoft – Redmond, US
- Sandrine Blazy
University & IRISA – Rennes, FR
- Nathan Burow
MIT Lincoln Laboratory – Lexington, US
- David Chisnall
Microsoft Research – Cambridge, GB
- Mads Dam
KTH Royal Institute of Technology – Stockholm, SE
- Ergys Dona
EPFL Lausanne, CH
- Cédric Fournet
Microsoft Research – Cambridge, GB
- Tal Garfinkel
Corepoint Systems – Penn Valley, US
- Chung-Kil Hur
Seoul National University, KR
- Jérémie Koenig
Yale University – New Haven, US
- Per Larsen
Immunant – Irvine, US
- Amit Levy
Princeton University, US
- Toby Murray
The University of Melbourne, AU
- Andrew Myers
Cornell University – Ithaca, US
- Santosh Nagarakatte
Rutgers University – Piscataway, US
- Elisabeth Oswald
Alpen-Adria-Universität Klagenfurt, AT
- Zoe Paraskevopoulou
Northeastern University – Boston, US
- Mathias Payer
EPFL – Lausanne, CH

- Andreas Rossberg
  Dfinity – Zürich, CH
- Kostya Serebryany
  Google – Mountain View, US
- Peter Sewell
  University of Cambridge, GB
- Zhong Shao
  Yale University – New Haven, US
- Deian Stefan
  University of California – San Diego, US
- Gang Tan
  Pennsylvania State University – University Park, US
- Nikos Vasilakis
  MIT – Cambridge, US
- Marco Vassena
  CISPA – Saarbrücken, DE
- Drew Zagieboylo
  Cornell University – Ithaca, US