DAGSTUHL
REPORTS

Volume 5, Issue 3, March 2015

Dagstuhl Reports, Vol. 5, Issue 3 ISSN 2192-5283

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
September, 2015

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.5.3.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 15101

Bridging Information Visualization with Machine Learning

Edited by

Daniel A. Keim[1], Tamara Munzner[2], Fabrice Rossi[3], and Michel Verleysen[4]

1 **Universität Konstanz, DE,** daniel.keim@uni-konstanz.de
2 **University of British Columbia – Vancouver, CA,** tmm@cs.ubc.ca
3 **Université Paris I, FR,** Fabrice.Rossi@univ-paris1.fr
4 **Université Catholique de Louvain, BE,** michel.verleysen@uclouvain.be

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15101 "Bridging Information Visualization with Machine Learning". This seminar is a successor to Dagstuhl seminar 12081 "Information Visualization, Visual Data Mining and Machine Learning" held in 2012. The main goal of this second seminar was to identify important challenges to overcome in order to build systems that integrate machine learning and information visualization.

Seminar March 1–6, 2015 – http://www.dagstuhl.de/15101

1998 ACM Subject Classification H.5 Information Interfaces and Presentations, I.3 Computer Graphics, I.5.4 Computer Vision, H.2.8 Database Applications, H.3.3 Information Search and Retrieval, I.2.6 Learning

Keywords and phrases Information visualization, Machine learning, Visual data mining, Exploratory data analysis

Digital Object Identifier 10.4230/DagRep.5.3.1

1 Executive Summary

Daniel A. Keim
Tamara Munzner
Fabrice Rossi
Michel Verleysen

Motivations and context of the seminar

Following the success of Dagstuhl seminar 12081 "Information Visualization, Visual Data Mining and Machine Learning" [1, 2], which provided to the participants from the IV and ML communities the ground for understanding each other, this Dagstuhl seminar aimed at bringing once again the visualization and machine learning communities together.

Information visualization and visual data mining leverage the human visual system to provide insight and understanding of unorganized data. Visualizing data in a way that is appropriate for the user's needs proves essential in a number of situations: getting insights about data before a further more quantitative analysis (e.g., for expert selection of a number of clusters in a data set), presenting data to a user through well-chosen table, graph or other structured representations, relying on the cognitive skills of humans to show them extended information in a compact way, etc.

The scalability of visualization methods is an issue: human vision is intrinsically limited to between two and three dimensions, and the human preattentive system cannot handle more than a few combined features. In addition the computational burden of many visualization methods is too large for real time interactive use with large datasets. In order to address these scalability issues and to enable visual data mining of massive sets of high dimensional data (or so-called "big data"), simplification methods are needed, so as to select and/or summarize important dimensions and/or objects.

Traditionally, two scientific communities developed tools to address these problems: the machine learning (ML) and information visualization (IV) communities. On the one hand, ML provides a collection of automated data summarizing/compression solutions. Clustering algorithms summarize a set of objects with a smaller set of prototypes, while projection algorithms reduce the dimensionality of objects described by high-dimensional vectors. On the other hand, the IV community has developed user-centric and interactive methods to handle the human vision scalability issue.

Building upon seminar 12081, the present seminar aimed at understanding key challenges such as interactivity, quality assessment, platforms and software, and others.

Organization

The seminar was organized in order to maximize discussion time and in a way that avoided a conference like program with classical scheduled talks. After some lightning introduction by each participant, the seminar began with two tutorial talks one about machine learning (focused on visualization related topics) followed by another one about information visualization. Indeed, while some attendants of the present seminar participated to seminar 12081, most of the participants did not. The tutorials helped establishing some common vocabulary and giving an idea of ongoing research in ML and IV.

After those talks, the seminar was organized in parallel working groups with periodic plenary meeting and discussions, as described below.

Topics and groups

After the two tutorials, the participants spend some time identifying topics they would like to discuss during the seminar. Twenty one emerged:

1. Definition and analysis of quantitative evaluation measures for dimensionality reduction (DR) methods (and for other methods);
2. In the context of dimensionality reduction: visualization of quality measures and of the sensitivity of some results to user inputs;
3. What IV tasks (in addition to DR related tasks) could benefit from ML? What ML tasks could benefit from IV?
4. Reproducible/stable methods and the link of those aspects to sensitivity and consensus results;
5. Understanding the role of the user in mixed systems (which include both a ML and an IV component);
6. Interactive steerable ML methods (relation to intermediate results);
7. Methods from both fields for dynamic multivariate networks;
8. ML methods that can scale up to IV demands (especially in terms of interactivity);

9. Interpretable/transparent decisions;
10. Uncertainty;
11. Matching vocabularies/taxonomies between ML and IV;
12. Limits to ML;
13. Causality;
14. User guidance: precalculating results, understanding user intentions;
15. Mixing user and data driven evaluation (leveraging a ROC curve, for instance);
16. Privacy;
17. Applications and use cases;
18. Prior knowledge integration;
19. Formalizing task definition;
20. Usability;
21. Larger scope ML.

After some clustering and voting those topics were merged into six popular broader subjects which were discussed in working groups through the rest of the week:

1. Dynamic networks
2. Quality
3. Emerging tasks
4. Role of the user
5. Reproducibility and interpretability
6. New techniques for Big Data

The rest of the seminar was organized as a series of meeting in working groups interleaved with plenary meetings which allowed working groups to report on their joint work, to steer the global process, etc.

Conclusion

As reported in the rest of this document, the working groups were very productive as was the whole week. In particular, the participants have identified a number of issues that mostly revolve around complex systems that are being built for visual analytics. Those systems need to be scalable, they need to support rich interaction, steering, objective evaluation, etc. The results must be stable and interpretable, but the system must also be able to include uncertainty into the process (in addition to prior knowledge). Position papers and roadmaps have been written as a concrete output of the discussions on those complex visual analytics systems.

The productivity of the week has confirmed that researchers from information visualization and from machine learning share some common medium to long term research goals. It appeared also clearly that there is still a strong need for a better understanding between the two communities. As such, it was decided to work on joint tutorial proposals for upcoming IV and ML conferences. In order to facilitate the exchange between the communities outside of the perfect conditions provided by Dagstuhl, the blog "Visualization meets Machine Learning[1]" was initiated.

It should be noted finally that the seminar was very appreciated by the participants as reported by the survey. Because of the practical organization of the seminar, participants did not know each other fields very well and it might have been better to allows slightly more

[1] http://vismeetsml.b.uib.no/

time for personal introduction. Some open research questions from each field that seems interesting to the other fields could also have been presented. But the positive consequences of avoiding a conference like schedule was very appreciated. The participants were pleased by the ample time for discussions, the balance between the two communities and the quality of the discussions. Those aspects are quite unique to Dagstuhl.

References

1 Daniel A. Keim, Fabrice Rossi, Thomas Seidl, Michel Verleysen, and Stefan Wrobel. Dagstuhl Manifesto: Information Visualization, Visual Data Mining and Machine Learning (Dagstuhl Seminar 12081). *Informatik-Spektrum*, 35:58–83, 8 2012.

2 Daniel A. Keim, Fabrice Rossi, Thomas Seidl, Michel Verleysen, and Stefan Wrobel, (editors). *Information Visualization, Visual Data Mining and Machine Learning (Dagstuhl Seminar 12081)*, Dagstuhl Reports, 2(2):58–83, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012. http://dx.doi.org/10.4230/DagRep.2.2.58

2 Table of Contents

Executive Summary

Overview of Tutorial Talks

Working Groups

3 Overview of Tutorial Talks

3.1 Machine Learning and Visualisation

Ian Nabney (Aston University – Birmingham, GB)

This talk describes two principal modes of data projection (or dimensionality reduction): topographic mappings and latent variable models. Principal Component Analysis is defined and it shown how it can be generalised to a non-linear projection based on distance preservation (topographic mapping exemplified by Neuroscale) or as a density model for the data (latent variable model exemplified by Generative Topographic Mapping – GTM). We then discuss how GTM can be extended to deal with missing values, discrete and mixed data types, hierarchies and feature selection. Illustrations from real applications are provided throughout.

3.2 Visualization Analysis and Design

Tamara Munzner (University of British Columbia – Vancouver, CA)

Computer-based visualization (vis) systems provide visual representations of datasets designed to help people carry out tasks more effectively. Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods. The design space of possible vis idioms is huge, and includes the considerations of both how to create and how to interact with visual representations. Vis design is full of trade-offs, and most possibilities in the design space are ineffective for a particular task, so validating the effectiveness of a design is both necessary and difficult. Vis designers must take into account three very different kinds of resource limitations: those of computers, of humans, and of displays. Vis usage can be analyzed in terms of why the user needs it, what data is shown, and how the idiom is designed. I will discuss this framework for analyzing the design of visualization systems.

4 Working Groups

4.1 Dynamic Networks

Tamara Munzner (University of British Columbia – Vancouver, CA), Stephen North (Infovisible – Oldwick, US), Eli Parviainen (Aalto University, FI), Daniel Weiskopf (Universität Stuttgart, DE), and Jarke van Wijk (TU Eindhoven, NL)

Networks are ubiquitous. Telecom networks, biological networks, software call graphs, citation graphs, sensor networks, financial transactions, social networks are some examples. In all these cases, it is not only the network structure that is relevant. Nodes and edges have associated multivariate data, and also, they are often dynamic. Attributes change, and also, in many cases networks are derived from streams of events (messages, communications, transactions), where each event has at least a time stamp, and two nodes as associated data.

Such large and complex networks are notoriously hard to visualize and understand. Just showing the structure of networks with a few hundred nodes already gives rise to the so-called hairball images, dynamics and associated data are yet another dimension of complexity. In the visualization community, novel representations and interaction techniques are proposed, but the problem is far from solved. Hence, the generic question is what machine learning can offer to provide more insight in such networks. Typical tasks are the identification of outliers and anomalous behavior, partitioning a sequence of time steps into clusters, identification of trends and discontinuities, and finding dynamic clusters of nodes.

A lively discussion gave rise to three possible approaches. As model for the data we used a simple sequence of networks $G_i, i = 1, \ldots, N$. The first approach concerns the use of a predictive model. Given such a model, one can predict for each time step a graph G'_i, given the other graphs $G_j, j \neq i$. Next, the difference between prediction and actual data can be shown, to reveal how the given data differs from expectation. A second idea is to use dynamic clustering: derive clusters across multiple graphs, such that emerging and disappearing clusters can be shown. Finally, one approach could be to translate each network into some feature vector, and next apply machine learning on these feature vectors.

Conceptually, all these approaches seem plausible and promising, however, also many questions remain. First, all these require models and metrics, for instance to make predictions, to cluster, and to select features; second, a question is if one should strive for generic solutions, or that questions on network data are strongly application dependent and require custom solutions.

The participants of the workshop were excited about the topic and the possible approaches. However, the group lacked expertise to make further steps. Therefore, we decided not to continue and join other working groups.

4.2 Machine Learning Meets Visualization: A Roadmap for Scalable Data Analytics

Daniel Archambault (Swansea University, GB), Kerstin Bunte (UC Louvain, BE), Miguel Á. Carreira-Perpiñán (University of California – Merced, US), David Ebert (Purdue University – West Lafayette, US), Thomas Ertl (Universität Stuttgart, DE), and Blaz Zupan (University of Ljubljana, SI)

4.2.1 Introduction

The big data problem requires the development of novel analytic tools for knowledge discovery and data interpretation (for example [1, 2]). The fields of visualization and machine learning have been addressing this problem from different perspectives and advances in both communities need to be leveraged in order to make progress. Machine learning has proposed algorithms that can address and represent large volumes of data enabling visualizations to scale. Conversely, visualization provides can leverage the human perceptual system to interpret and uncover hidden patterns in these data sets.

In this short report we identify areas where machine learning can assist the process of data visualization and areas where visualization can drive machine learning processes. These areas are summarized in Figure 1.

4.2.2 Visualization benefits from Machine Learning

Traditional uses of machine learning for visualization have included exploratory procedures such as feature selection, dimensionality reduction and clustering. Here we describe additional machine learning concepts that may be of benefit for visualization research.

Binary hashing. Binary hashing has emerged in recent years as an efficient way to speed up information retrieval of high-dimensional data, such as images or documents. Given, say, a query image, searching in a large database of images for the nearest images to the query is a high-dimensional nearest neighbor finding problem whose exact solution is computationally very expensive. For example, representing each image with a 300-dimensional vector of SIFT features would take over one terabyte for one billion images. In binary hashing, one maps every image to a compact binary vector so that Hamming distances in binary space approximately preserve distances in image space. Searching for neighbors in binary space is much faster because 1) the dimensionality of the binary vector is much smaller than the dimensionality of the image, 2) Hamming distance computations can be done very efficiently with hardware support for binary arithmetic, and 3) the size of the binary-vector database is small enough that it can even fit in RAM memory rather than disk. In the earlier example, using 32 bits per image the database would take 4 GB. The success of binary hashing depends on being able to learn a good hash function, which maps images to bit vectors so that distances are approximately preserved. Initial algorithms learned a dimensionality reduction mapping and simply truncated it to output binary values [3], while recent efforts try to optimize the function directly respecting the binary nature of its outputs [4, 5].

As an example application, consider visualizing a stream of Tweets. Given a new Tweet, we can turn it into a high-dimensional vector using a bag-of-words representation and then map it to binary space using the binary hash function. Searching in a binary database of Tweets quickly retrieves a selection of approximate neighboring Tweets, which can be refined to keep only true neighbors by computing distances between the retrieved bag-of-words vectors and the query.

Coresets. Besides the dimensionality which leads to high computational costs and memory requirements also the number of samples influences the efficiency of many applications whenever very large amounts are collected as in Astronomy, Photography, streaming and so on. Random sampling, feature extraction and ϵ-samples are often used strategies to deal with this problem. This leads to a general concept combining these ideas referred to as coresets [6, 7]. The aim is to find a small (weighted) subset of the data, which guarantees, that a training procedure based on this subset provides comparable good results also for the original set. The effectiveness has been shown for several objectives, ranging from for example dimension reduction, clustering and Gaussian Mixture Models and surprisingly also coresets with size independent from the size of the data set have been proven. Moreover, efficient parallel and distributed strategies to find coresets are proposed, which makes them perfectly suitable for big data analysis and streaming settings. Information visualization can directly benefit from this concept, since it usually depends on pairwise similarities or distances resulting in quadratic complexity with respect to the number of data items.

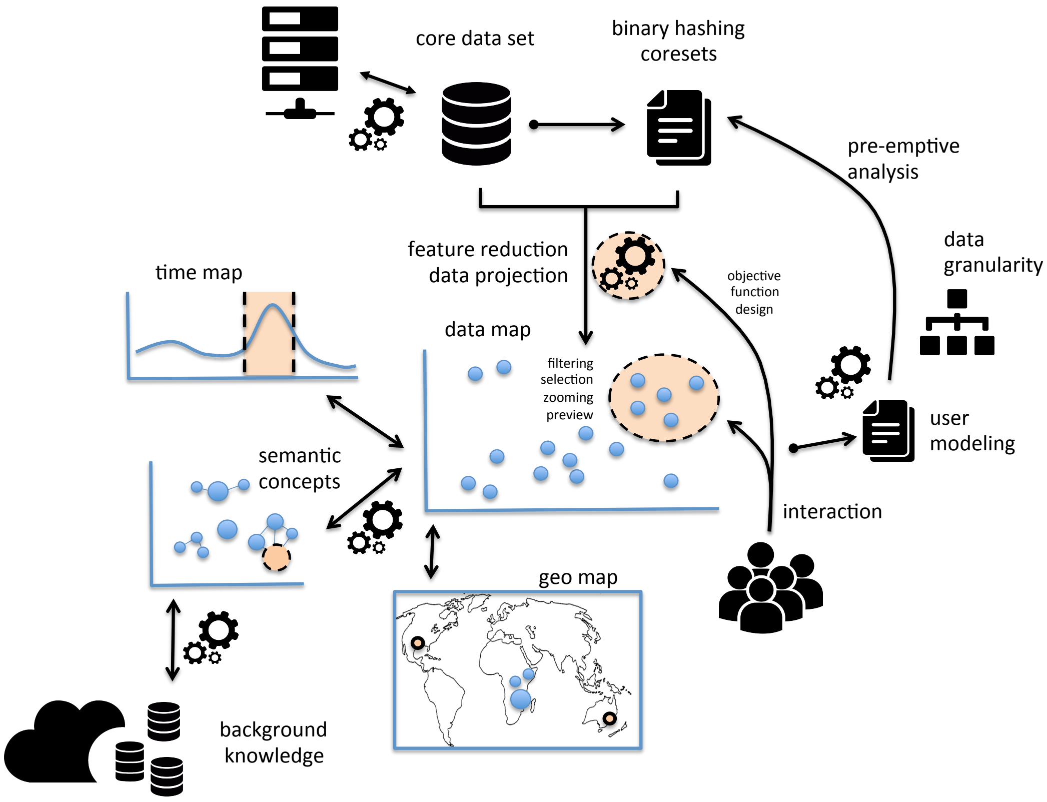


Figure 1 Possible interplay between machine learning and data visualization. The core data set (top), possibly storing the information from the data stream, is preprocessed for binary hashing and coresets discovery. Preprocessing enables index-based data retrieval, selection of the representative data instances, and fast distance computation. Multi-view visualization initially displays data in the coreset, but also supports user in digging deeper and retrieving data from neighborhood, time, location or concept-specific spaces. Data-related semantic concepts are retrieved from related data bases and organized in ontology or network. Visualizations are interlinked: any change in selection in one view updates the information in all other views. Machine learning algorithms for clustering, assessment of concept enrichment, outlier detection and classification of uncharacterized data instances are triggered on the fly. User's interactions are recorded and modeled, and provide means of predicting them and executing the most likely data-intensive operations that the user can trigger in the future before they are actually needed. User can change the attributes or position of data instances in any visualization, thus visually changing the objective function that is optimized in the visualizations. Change of objective function is followed by repositioning of data elements in the visualizations.

Inclusion of background knowledge. Besides the core data which we are trying to analyze, there may be additional information available that may shed light on the interaction between data entities, or additionally explain the discovered data patterns. Background knowledge may be incorporated at various stages of data analysis. For machine learning, it may serve as a prior that constrains the hypothesis space and steers the optimization towards models that are consistent both with data and additional information. For visualization, background knowledge may provide information that support interpretation. What characteristics outside of the data space are common to a set of co-clustered data points? What is the match between the visualized data and the concepts that are related to the problem investigated but were not included in the original data set? Crucial to exploration of the interplay between the data any additional information are graphical user interfaces to access and explore such interaction, and quantification of relations between data instances and concepts to draw statistically founded conclusions. An example of the later are enrichment analysis techniques from bioinformatics [8], which ranks the data annotation terms according to their association with a selected group of data entities.

Visualization of classifiers. Recent approaches accommodate for the growing demand of interpretable models, which lead to visualizations, not only showing the data, but also an inferred classification model [9, 10]. This enables the use of the human perceptual qualities to detect: 1) potential mis-labelings which might emerge as outliers, 2) noisy regions which are difficult to classify, 3) the modality of each class and 4) overfitting effects of the model for example.

Visualization of machine learning processes. Recent work in both the machine learning and human computer interaction communities has focused on how to use visualization in order to improve how we tune machine learning approaches. Specifically, the approaches have been applied to the problem of network alarm triage [11] and optimizing machine learning approaches for given performance constraints [12, 13, 14]. This work provides a way to optimize machine learning processes for given tasks, instead of treating the approach as a black box.

Steerability, semantic zoom and user constraints. One of the most promising applications for information visualization to machine learning is steerability. Steerable approaches in the field of visualization allow for the user to interactively guide large computations towards areas of interest. Such approaches, when applied in conjunction with machine learning can be very powerful, allowing heavy weight computations to be targeted to areas of interest in a very large data set. Steerable approaches first emerged in the field of scientific visualization [15] and have been subsequently been applied to the process of visualizing graphs [16, 17].

Moreover, user constraints, like for example walls in maps or must-link and cannot-link constraints for clustering, can be accumulated by interactions with a display. Any machine learning algorithms suitable for constraint-based optimization as satisfiability optimization can benefit from such interactive solutions. First steps to directly incorporate user constraints into the optimization process of visualizations has been taken for example in [18]. These constraints can be imposed through user interaction and the resultant computation could be used in conjunction with a semantic zoom.

4.2.3 Way Forward

Visual design of objective functions. Recently, some methods have been proposed to make model parametrization and data exploration more intuitive without requiring deep methodological knowledge of the data expert. Those approaches provide for example a simplex where a point in the area corresponds to a parametrization of the underlying model comparable to multidimensional sliders. Other tools facilitate an interactive data exploration, by visually combining modules implementing different data processing steps, which could be combined by the user. However, the parametrization is a very high level design mechanism and limited in its impact on the final model. To change the fundamental design and assumptions of the model one would need to interact on much lower levels such as the mathematical formulation. It would be interesting when a user could visually combine mathematical atoms to form new objectives as for instance using graphical models in Bayesian formulations, which are inferred automatically.

Modelling of user interactions. Machine learning should not only be used for summarizing data. One approach is to use machine learning to learn user actions and predict the likely future ones. The area of adaptive user interfaces and intelligent user interfaces could be applied to the field of information visualization to determine likely future interactions with the system to give it a *head start* on heavyweight computational processes in a steerable environment.

Data fusion. In making quality decisions, us, humans, tend to use all available information that is directly or only indirectly related to the problem. In machine learning, the notion of wide-range data integration has been explored by data methods of fusion. So far, data fusion has primarily focused on development of predictive models by combining different data sources through, say, through kernel-based methods [19] or collective matrix factorization [20]. The research in this field is important to big data, as it addresses the variety and span of data sources. To bring the resulting models to the data analyst, however, data fusion would need to be combined with data visualization using the approaches that have yet to be conceptualized and developed.

References

1 Junghoon Chae, Dennis Thom, Harald Bosch, Yun Jang, Ross Maciejewski, David S Ebert, and Thomas Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 143–152. IEEE, 2012.

2 Harald Bosch, Dennis Thom, Florian Heimerl, Edwin Puttmann, Steffen Koch, Robert Kruger, Michael Worner, and Thomas Ertl. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Trans. on Visualization and Computer Graphics*, 19(12):2022–2031, 2013.

3 Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In Daphne Koller, Yoshua Bengio, Dale Schuurmans, Leon Bottou, and Aron Culotta, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 21, pages 1753–1760, 2009.

4 Miguel Á. Carreira-Perpiñán and Ramin Raziperchikolaei. Hashing with binary autoencoders. In *Proc. of the 2015 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, June 7–12 2015.

5 Ramin Raziperchikolaei and Miguel Á. Carreira-Perpiñán. Learning hashing with affinity-based loss functions using auxiliary coordinates. arXiv:1501.05352, January 21 2015.

6 Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. In *Combinatorial and Computational Geometry, MSRI*, pages 1–30. University Press, 2005.

7 Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC'11, pages 569–578, New York, NY, USA, 2011. ACM.

8 Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, 13(3):281–291, 2012.

9 A. Schulz, A. Gisbrecht, K. Bunte, and B. Hammer. How to visualize a classifier? In B. Hammer and T. Villmann, editors, *New Challenges in Neural Computation (NC2), ser. Workshop of the GI-Fachgruppe Neuronale Netze and the German Neural Networks Society in connection to DAGM 2012*, Graz, Austria, August 2012. LNCS.

10 Alexander Schulz, Andrej Gisbrecht, and Barbara Hammer. Using nonlinear dimensionality reduction to visualize classifiers. volume 7902 of *IWANN(1)*, pages 59–68. Springer, 2013.

11 S. Amershi, B. Lee, A Kapoor, R. Mahajan, and B. Christian. Cuet: Human-guided fast and accurate network alarm triage. In *Proc. CHI 2011*, pages 157–166, 2011.

12 A. Kapoor, B. Lee, D. Tan, and E. Horvitz. Interactive optimization for steering machine classification. In *Proc. of CHI 2010*, pages 1343–1352, 2010.

13 A. Kapoor, B. Lee, D. Tan, and E. Horvitz. Performance and preferences: Interactive refinement of machine learning procedures. In *Proc. of AAAI 2012*, 2012.

14 S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Simard, and Jina Suh. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proc. CHI 2015*, 2015.

15 S.G. Parker and C.R. Johnson. SCIrun: A scientific programming environment for computational steering. In *Proc. of Supercomputing*, 1995.

16 D. Archambault, T. Munzner, and D. Auber. GrouseFlocks: Steerable exploration of graph hierarchy space. *IEEE Trans. on Visualization and Computer Graphics*, 14(4):900–913, 2008.

17 D. Archambault, H. C. Purchase, and B. Pinaud. The readability of path-preserving clusterings of graphs. *Computer Graphics Forum*, 29(3):1173–1182, 2010.

18 Kerstin Bunte, Matti Järvisalo, Jeremias Berg, Petri Myllymäki, Jaakko Peltonen, and Samuel Kaski. Optimal neighborhood preserving visualization by maximum satisfiability. In *Proceedings of AAAI-14, The Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

19 Shi Yu, Léon-Charles Tranchevent, Bart De Moor, and Yves Moreau. *Kernel-based Data Fusion for Machine Learning.* Springer-Verlag, Berlin, Heidelberg, 2011.

20 Marinka Zitnik and Blaz Zupan. Data fusion by matrix factorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37:41–53, 2014.

4.3 User and Machine Learning Dialogue for Visual Analytics

Francois Blayo (Ipseite SA – Lausanne, CH), Ignacio Díaz Blanco (University of Oviedo, ES), Alex Endert (Georgia Institute of Technology, US), Ian Nabney (Aston University – Birmingham, GB), William Ribarsky (University of North Carolina – Charlotte, US), Fabrice Rossi (Université Paris I, FR), Cagatay Turkay (City University – London, GB), and B. L. William Wong (Middlesex University, GB)

Thomas and Cook (2005) presented the visual analytics community with a challenge to create visualization technologies that work interactively and smoothly with computational algorithms. They describe such a dialog as analytic discourse. They described this as "…visually-based methods to support the entire analytic reasoning process", including the analysis of data as well as structured reasoning techniques such as the construction of arguments, convergent- divergent investigation, and evaluation of alternatives. These methods must support not only the analytical process itself but also the progress tracking and analytical review processes.

The merger of machine learning and visual analytics presents many potential opportunities for visual data analysis. Visual analytics leverages the cognitive and perceptual abilities of humans to enable them to explore, reason, and discover data features visually. Machine learning leverages computational abilities of computers to perform complex data-intensive calculations to produce results for specific questions or tasks. Currently, visual analytic techniques exist that make use of select machine learning models or algorithms (often, dimension reduction techniques). However, there are additional techniques that can apply to the broader visual data analysis process. Doing so reveals opportunities for how to couple user tasks and activities with such models.

The discussion at this Dagstuhl seminar focuses on the role of the user in this process of integrating machine learning into visual analytics. We discussed challenges and difficulties of designing a system that would enable analytic discourse. How should specific machine learning techniques be incorporated into the visual data exploration process? We present a discussion of the role of user interaction in such a dialog between machine learning techniques, interactive visualisation and cognitive processes, and provide a scenario to illustrate these concepts. What would be or should be the role of the user when we combine machine learning with interactive visualization in ways that would enable users to steer and drive the computational algorithms?

We claim that user interactions are an important aspect of such a combination. In visual analytics, user interactions have been designed and implemented as mechanisms by which users can augment the visualization parameters, filter data, and other direct changes to the application. In machine learning, user interaction has been used as directed feedback on results of computation (e.g., classification models, predictive models, etc.). However, we challenge these two communities to consider an additional lens through which user interaction can be viewed. We posit that every user interaction encodes some (potentially small) part of analytical reasoning and insight. The challenge posed to the community is how to adequately leverage these bits of analytical reasoning and integrate them into the holistic visual analytics system.

4.4 Bridging the Analytics Gap: Human-centered Machine Learning

Michael Sedlmair (Universität Wien, AT), Leishi Zhang (Middlesex University, GB), Dominik Sacha (Universität Konstanz, DE), John Aldo Lee (UC Louvain, BE), Daniel Weiskopf (Universität Stuttgart, DE), Bassam Mokbel (Universität Bielefeld, DE), Stephen North (Infovisible – Oldwick, US), Thomas Villmann (Hochschule Mittweida, DE), and Daniel Keim (Universität Konstanz, DE)

The goal of visual analytics systems is to solve complex problems by integrating automated data analysis methods with interactive visualizations. While numerous visual analytics systems have been developed for specific application problems, a general understanding of how this integration can be realized is still largely missing.

Towards the goal of better understanding this interplay, our working group developed a

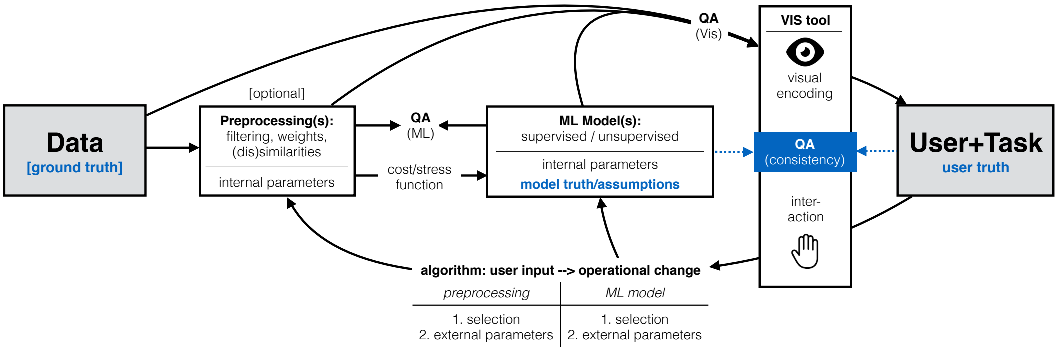


Figure 2 Our conceptual framework. The main components of the pipeline are shown in the center. Visual encoding of results at different stages of the pipeline are indicated via arrows at the top, which point trough the VIS tool to the User. User interactions are indicated through the arrows at the bottom, again via the VIS tool. Different versions of "truth" are highlighted in blue, together with a quality assurance (QA) component that helps ensuring consistency between the ML components and the user.

framework that conceptualizes how integration of machine learning methods and interactive visualizations can be implemented (see Figure 2). We identified aspects of machine learning methods, which are amenable to be controlled interactively by the user, such as the choice and parameterization of machine learning models. While some of these aspects can be automatically optimized by pre-defined cost functions, in many applications it is crucial to allow the user to control them interactively. Our framework makes the crucial interplay between automated algorithms and interactive visualizations more concrete. To show its utility, we used it to analyze several existing visual analytics systems against the framework, demonstrating that it provides useful understanding as well as guidelines when developing and evaluating them.

Based on our framework, we finally characterized a set of 11 open challenges:

1. Mapping user input to ML model adaptation – At the core of our conceptual framework lies the idea that external parameters of an ML model or preprocessor can be adapted via iterative, and direct user interactions. Some simple examples exist, such as updating a parameter of a dimension reception model based on how a user moves around points in a scatterplot. However, mapping user inputs to more complex actions, such as switching between different model types, remains an open challenge.
2. Discontinuous changes – Implementing such more complex interactions, may sometimes cause major, abrupt changes in the underlying ML components. A major challenge is how to communicate such abrupt changes in a perceptually understandable way to the user, in order to keep her in the loop.
3. Effective learning from small data – An algorithmic challenge that our envisioned human-in-the-loop scenario poses is learning from a small number of user interactions, likely in the single or low double digits. While these interactions will be used as stimuli for training the ML model's internal parameters, most ML methods require a larger set of input data to train the model, typically hundreds or thousands of input stimuli.
4. Interactive and Scalable Algorithms – Another technical challenge is that the user should not be disrupted by long response times occurring during adaptation of ML models. Therefore, training procedures must be efficient in terms of computation time. In this regard, new approximation approaches, and methods for including intermediate results will be needed.

5. Balance between Model and Visual Quality – A major challenge in a rich human-in-the-loop analysis process is assuring both ML model quality and visualization quality. However, the two types of quality preservation do not always align. For example, a visual embedding that preserves the input data structure well may not have good readability due to high dimensionality. Data sparsity and noise can cause clutter and poor group separation. While some techniques exist, the challenge is to provide a clear indication of both quality measures to the user and help them to find the right balance between the two, so meaningful analysis can be carried out.
6. Consistency between Model and Human – In current visual analytics systems, checking consistency between model and user is often done manually. The user must evaluate the model and provide feedback to the system. When a conflict between the two arises, the outcome can be biased. Such problems can be alleviated by developing automatic methods that check consistency quality, highlight inconsistency, and recommend appropriate actions. Note that, though consistency between human and machine is desirable, it does not guarantee correctness per se.
7. User Guidance – Current general purpose systems, such as R, offer multiple choices of preprocessors and ML models that can be applied to analyze data. Application users who are not ML experts, however, often find it difficult to know which choices are most suitable for the data and task at hand, and to find good parameter settings for the selected ML components. Assistance and guidance from the system is therefore of utmost importance.
8. Better Perceptual Quality Measures – While numerous quality measures have been designed for algorithmic purposes, we find few measures that have a truly perceptual motivation. Current visualization measures do not cover any complex approaches from perceptual psychology to accurately capture mechanism of human visual perception. Such models are especially difficult if they want to include human-computer interaction and data dependency. Having an accurate model of human perception would not only be helpful for guiding users through the space of visualization design choices, but also for ensuring consistency between the user and ML models as discussed above.
9. Uncertainty Description, Quantification, and Propagation – Another challenge is that we need to describe and compute uncertainty introduced by the various pieces within the pipeline of using visualization and machine learning together. Describing and quantifying uncertainty in the interplay between user, task, and ML model is a non-trivial endeavor.
10. Visualization of Uncertainty – Once we have a quantification of uncertainty, what shall we do with it? One research question deals with the visualization of such uncertainty. There is much previous work on visualization techniques to display data uncertainty of spatial data, such as volume or flow visualization. We find much less work on uncertainty visualization of abstract data, such as high-dimensional data visualization, common in ML applications.
11. Uncertainty Reduction – A related challenge is how we can reduce the amount of uncertainty. One possibility is to steer the visual analytics process toward a "sweet spot" where the process becomes less sensitive to the influence of uncertainty. Here, sensitivity analysis or similar approaches might be adapted. Another approach to reduce the uncertainty from user input (such as inaccuracies introduced by annotation uncertainty) could be automatic checks for consistency with the machine-learning model. This idea is tightly linked to having an appropriate quality assessment for consistency between model and user.

4.5 Emerging tasks at the crossing of machine learning and information visualisation

Barbara Hammer (Bielefeld University, DE), Stephen Ingram (UBC, Vancouver, CA), Samuel Kaski (Aalto University, Helsinki, FI), Eli Parviainen (Aalto University, Helsinki, FI), Jaakko Peltonen (Aalto University / University of Tampere, FI), Jing Yang (University of North Carolina, Charlotte, US), and Leishi Zhang (Middlesex University, London, UK)

4.5.1 Introduction

An ever increasing number of domains is accompanied by digital fingerprints: industry 4.0 with heterogeneous sensor streams monitoring and controlling industrial processes; smart sensor signals of everyday life which become ubiquitous in the context of smart phones, wearable devices, and digitalisation of the automotive sector; highly sensitive medical diagnostics based on a variety of different biotechnologies leading to individualised -omics sources; the financial market which is characterised by a multitude of digitally stored indicators; social life which is tightly mirrored in social media and social networks; or even politics which, increasingly, makes use of digital information and the underlying ways of decision making [1]. This digital data revolution places new challenges towards computer scientists: they are not only developing new technologies for efficient data measurement, pervasive data storage, privacy preservation, etc, but they also face the challenge to enable humans to cope with the information buried in these data and take according action. This has been identified as one of the major questions when it comes to big data, and the term ‘big data analytics’ has been coined as a key capacity of modern society [2].

Machine learning (ML) and information visualisation (InfoVis) constitute two pivotal disciplines which enable humans to unravel the information hidden in digital data. Albeit these two disciplines address similar questions and challenges, their underlying technologies and theoretical background often differ. Research directions such as the developments put under the overarching umbrella of ‘scalable visual analytics’ constitute promising attempts to bridge this gap [3], and there do exist formalisms and tools which successfully rely on aspects of both worlds [4]. The goal of this contribution is to discuss such links by zooming onto the tasks and questions which are shared by ML and InfoVis, and their respective approaches to tackle these tasks. Thereby, we do not cover the full spectrum. Rather we put spotlights on interesting aspects at two different levels of scientific granularity: differences and shared technology, respectively, as concerns central paradigms of the data processing pipeline in InfoVis and ML, on the one hand; and topics which we regard as emerging topics in the domains of ML and InfoVis, which share a common research question but which are looked at from two different points of view in the two disciplines. We discuss each of these spotlights separately in a short paragraph in the sequel.

4.5.2 Classical dimensionality reduction

Often, data are vectorial, but high dimensional, such that its direct inspection as points in the plane is impossible. Their intuitive visual access constitutes one of the classical tasks which are addressed by both, InfoVis and ML – but technologies differ [5]. InfoVis provides a number of different techniques to display such data, such as scatter plots, parallel coordinates, heat maps, glyphs, or Chernoff faces, as well as interactive exploration e.g.

based on tour methods. In this context, a major question which is investigated, is how these visualisation technologies align with human perception [6]. Conversely, ML almost solely relies on a static display of high dimensional data as a scatter plot, but it explores a variety of different approaches to learn suitable two-dimensional coordinates from the given data which preserve as much structure of the original data as possible [7]. The focus is on the different mathematical ways to formalise the concept of structure preservation, and its efficient computational modelling.

These foci constitute two different views on the problem: InfoVis concentrates on human perception and puts the user into the centre, while ML focusses on (often nonlinear) aspects in the given data and their mathematical formalisation. These different views of the same problem open the way towards new paradigms, which combine rich visual display technologies and interaction methods as offered by InfoVis with highly flexible data driven structure preservation as provided by ML technology. Such enriched data displays have a great potential for emerging areas such as biomedical data analysis where, often, heterogeneous information or additional structures have to be taken into account [8, 9, 10, 11]

4.5.3 Modelling

Both, ML and InfoVis essentially model observed data in such a way that the information buried in the data can easily be accessed by humans. Thereby, a crucial part is to identify general paradigms and workflows which allow researchers to access the given data in a principled and scientifically valid way.

For InfoVis, general workflows such as the InfoVis mantra ‘overview first, zoom and filter, then details on demand’ and clear relations of the technology to be used for display and the type of data to be displayed are well established [12, 13]. These modelling paradigms offer guidelines for the ‘scientific language’ which can be used for data visualisation and the realisation of the dynamics of such display. These principles are usually not tailored to the exact values of the data to be displayed.

For ML, the key aim is to model the given, observed data, and one overarching paradigm underlying modelling in ML is the language of probability theory and statistics: often, learning is phrased as probabilistic modelling of the given data points which are regarded as samples of an underlying data distribution; modelling refers to the inference of the latter, i.e. estimating generative probabilistic models from a finite number of given observations [14]. Thereby, computational learning theory provides a mathematical justification that this principle is valid. Such modelling is data centred, in the sense that different models result from different measurements, and the influence of the observed data on the final model can be quantified by the deviation of the resulting distribution and the prior.

In principle, probabilistic modelling is universal, being capable of modelling every possible underlying regularity – in practice, assumptions have to be made to avoid overfitting, and regularisation which is based on prior knowledge or universal priors (such as sparsity) has to be used. A good choice of priors remains a challenge in particular for sparse measurements and heterogeneous data sources. Here human intuition could help to regularise accordingly, opening up an interesting support line from the InfoVis field.

Interestingly, a Bayesian view on InfoVis, which treats data and also user interactions as observations, opens the ground towards an automation of display selection and adaptation of the views according to the data. Recently, some promising research along this line has been proposed, see e.g. [15].

4.5.4 Quantitative evaluation

Both, InfoVis and ML face the challenge to quantitatively and qualitatively evaluate their techniques. The used methods differ fundamentally, a fact which closely mirrors the user centred versus data centred view of the two disciplines.

For InfoVis, the evaluation of a system usually takes place in the form of user studies or user feedback, such as expert evaluation, lab studies, or field studies [16]. These enable a formal evaluation of important aspects of InfoVis systems such as their functionality, effectiveness, efficiency, or usability. Such evaluations are often time consuming, and they require a clear study design. Notably, these techniques do not make explicit assumptions about human perception, since humans are directly evaluating the models using their cognitive capabilities. Interesting attempts try to match human perception and formal mathematical measurements, which could result in a speed up of the design process due to the availability of computable measures mirroring human perception [17].

For ML, evaluation is almost solely data centred, and evaluation measures have its roots in statistics. Since the majority of ML technologies can be found in the field of so-called supervised learning, classical evaluation measures for ML technology refer to cost measures such as the classification error or regression error as evaluated in a cross-validation. It has been a long debate how to evaluate unsupervised methods such as clustering or dimensionality reduction for data visualisation, and widely accepted quantitative measures for the latter just emerged recently [18, 19]. One main problem in this context consists in the fact that data visualisation and unsupervised data analysis is a mathematically ill-posed problem, and it depends on the setting at hand, which aspects of the data are of interest for the user. It is often not clear how to formalise these fuzzy goals in terms of mathematical cost functions and model priors. In this respect, ML can benefit from the insights and evaluation technology which is common in InfoVis, since it enables to take the user expectation into account without the necessity to express the latter within mathematical terms.

Conversely, by focussing on an underlying data distribution and the generalisation ability of a model to new data, ML can rely on strong techniques offered by statistics. A general technology which allows to evaluate the generalisation ability and robustness of a model, for example, is provided by sampling methods such as bootstrap statistics or cross-validation [20]. Hence it is easily possible to automatically evaluate a given algorithm or model as concerns its statistical robustness – a prerequisite which is independent from the overarching goal of modelling.

4.5.5 Big and streaming data

Albeit ML and InfoVis constitute two key technologies when it comes to big data, both techniques also face a number of new challenges in this context [2]. Both disciplines have to cope with the increasing computational and memory demands when it comes to big data. Hadoop's map-reduce, as an example, constitutes a widely used technology in both domains [21].

Besides these grounds, both domains develop new data structures and algorithms to speedup computational costs for core methods such as spatio-temporal data representation or dimensionality reduction. Interesting recent proposals, for example, rely on an intricate hierarchical representation of data and a suitable summary of the information content at each hierarchical level: within InfoVis, so-called nanocubes enable to deal with tens of billions of data points efficiently [22]. In ML, a similar concept which has its roots in statistical physics has recently been proposed to speed up dimensionality reduction techniques from quadratic to only log-linear complexity [23, 24].

Often, data are not only big but arrive continuously over time. In such cases, the challenge is to face the specific data characteristics caused by its dynamic arrival. In InfoVis, streaming data visualisation deals with the problem to take user expectation and perception of temporal changes into account. As an example, dynamic graph drawing tries to optimally balance dynamic changes and constant characteristics within a visual display of dynamic graphs [25]. Besides human perception, enriched mathematical concepts such as parameterised lines can open the way to novel, efficient dynamic displays [26].

For ML, one of the core problems of streaming data analysis consists in the fact that a crucial assumption underlying classical ML is violated: data are usually no longer independently and identically distributed, rather trends occur. Thus, ML methods have to cope with the challenge of data trend, emerging and vanishing concepts, and intricate data dependencies over time, with quite a few novel approaches and theoretical models popping up to deal with these problems [27].

4.5.6 Few data

In the context of heterogeneous data and user interaction, a phenomenon, which lies on the opposite side, takes place: methods face the challenge to learn from few data only.

One example instantiation of this challenge is the detection of rare events within large data sets, popular applications being e.g. network intrusion detection, rare event detection, customer preference learning, crime detection, or change point detection [28, 29]. Here, specific ML and InfoVis techniques have to be used which are capable of dealing with highly imbalanced data sets and putting its focus on the few observed anomalities in the data, since the majority of observations belong to the class of 'normal' events in such settings.

Another application area deals with very few labeled events only, such as instantaneous learning from few examples. This becomes possible provided auxiliary information is taken into account, such as strong priors in Bayesian modelling of visual categories [30], or the wisdom of the crowd which manifests itself in social media [31].

One domain where learning from few examples would be very useful is the automated annotation of given data. Typically, interactive systems are offered by InfoVis technology which enable experts to annotate such events; still, this is usually too time consuming for the full data. Here automation as offered by ML would help. Currently, most automated annotation systems are specialised to the respective domain, covering e.g. genomic data annotation, texts, images, or specific events in time series data [32, 33]. An interplay of ML and InfoVis techniques could help to generalise these approaches towards a domain independent technique.

4.5.7 Causality

Interactive data analysis is concerned with insights into the given information such as characteristic patterns, summaries, or typical cases. Often, the causality of observations constitutes a key question humans are interested in: which measurements and observations are relevant for a certain effect and how do they relate to each other? What is the cause of a particularly interesting / annoying / relevant observation, and how can this effect be changed? While correlations of events can easily be determined based on classical statistics, the notion of causality – which event is the cause of which other event – requires a more in depth analysis. Typically, it does not suffice to analyze available observations only, rather it demands for a mediated probability or expert insight.

Interestingly, in recent years, the automated inference of causality from measured data has become more an more relevant in different areas of ML, caused by increasing data sets e.g. in neurobiology (such as action potentials of neurons, based on which neural connectivity should be predicted) [34, 35]. There do exist possibilities to infer causality in some settings, provided suitable prior assumptions are integrated into the models. One example is offered by independent component analysis, which is capable of unraveling mixed sources based on the notion of statistical independence only, and which can also be used for causality detection for linear relations. Naturally, human interaction can also help to clarify unclear cases which can occur due to highly nonlinear effects or sparse sampling; here an interactive analysis where ML and InfoVis provide different insights can be beneficial.

Having identified causal relationships, it remains a challenge to present these insights in such a way that the user can use this information for decision making in complex settings. A challenge is given by the fact that data are high dimensional, and causality is usually not only spotting relationships between simple measurements, rather it relates to significant macro-properties of the system, such as traffic jams and road network design in interactive traffic analysis. InfoVis provides a few technologies how to display such information efficiently and effectively in different contexts [36, 37].

4.5.8 Computational creativity

Automatic storytelling has been dubbed as one emerging area in InfoVis which goes beyond the mere display of data; rather it enables to build a whole story and line of argumentation around given data, supporting the arguments by suitable visualisations where appropriate [38]. Besides novel InfoVis tools, this task faces the challenge to infer a reasonable storyline automatically or with the help of the user from the given data; hence there is the need for fundamental arguing principles and inference mechanisms, typically techniques from ML and AI. Further, stories are often built around interesting exceptional events, hence rare meaningful events have to be detected automatically, as already discussed in section 4.5.6.

In ML, this question also touches on what is referred to as ‘computational creativity’: where are the relevant novel uncommon insights buried in the data? This imprecise notion can be partially matched with mathematical measures such as the entropy, which measures the amount of surprise in a data set, and successful technical systems which make use of these principles e.g. for efficient reinforcement learning have been proposed [39].

4.5.9 Collaborative work

Web and social media, among other aspects, enable an ever increasing availability of collaborative sources for data analysis: they provide basic data sources and background information based on which data analysis can be enriched, popular examples being e.g. collaborative filtering [40]; automated annotation and the wisdom of the crowd enables to rely on label information which, due to the sheer size of the participants, can be statistically very reliable; further, the web provides an environment where humans can increasingly work together and collaborate, making according platforms mandatory, examples are MOOCs or shared bioinformatics data bases.

These developments provide new possibilities but also new challenges for InfoVis and ML, such as the following: how to visualise and analyse data which comes from different sources, how to align the usually slightly different data representations and persistently store the involved information? One crucial aspects is, for example, a common data space or language shared by the collaborators, a question which is tackled under the umbrella of transfer learning in ML [41], and addressed in first systems in the InfoVis field [42].

4.5.10 Discussion

We have discussed some tasks and questions shared by InfoVis and ML, pointing out the different view of the two disciplines due to their user centred versus data centred view. This difference often results in different technologies, which can be combined to open up revenues for new, even more powerful technologies. With the advent of big data and distributed sensors, data sets and analysis tasks become ever more complex: data sources are heterogeneous, data are distributed, and massive volumes have to be addressed. At the same time the tasks, which can be tackled, are no longer restricted to simple correlations, but complex questions which relate to planning and decision making are investigated. This calls for a combination of the two fields, such that it becomes possible to address these challenged with integrated methods which can automate inference wherever possible, but which can use interactive analysis wherever expert feedback is mandatory.

References

1 Tom Khalil. Big data is a big deal. White House, Sep 2012.

2 Committee on the Analysis of Massive Data, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Their Applications, Division on Engineering and Physical Sciences, and National Research Council. *Frontiers in Massive Data Analysis.* National Academic Press, 2013.

3 Daniel A. Keim. Solving problems with visual analytics: The role of visualization and analytics in exploring big data. In *Datenbanksysteme für Business, Technologie und Web (BTW), 15. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), 11.-15.3.2013 in Magdeburg, Germany. Proceedings*, pages 17–18, 2013.

4 Tuukka Ruotsalo, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. Interactive intent modeling: information discovery beyond search. *Commun. ACM*, 58(1):86–92, 2015.

5 Stephen Ingram, Tamara Munzner, Veronika Irvine, Melanie Tory, Steven Bergner, and Torsten Möller. Dimstiller: Workflows for dimensional analysis and reduction. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2010, Salt Lake City, Utah, USA, 24-29 October 2010, part of VisWeek 2010*, pages 3–10, 2010.

6 Michael Friendly. Milestones in the history of thematic cartography, statistical graphics, and data visualization. In *13th International Conference on Database and Expert Systems Applications (DEXA 2002), Aix en Provence*, pages 59–66. Press, 1995.

7 Andrej Gisbrecht and Barbara Hammer. Data visualization by nonlinear dimensionality reduction. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 5(2):51–73, 2015.

8 Matthew Brehmer, Stephen Ingram, Jonathan Stray, and Tamara Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE Trans. Vis. Comput. Graph.*, 20(12):2271–2280, 2014.

9 Jing Yang, Yujie Liu, Xin Zhang, Xiaoru Yuan, Ye Zhao, Scott Barlowe, and Shixia Liu. PIWI: visually exploring graphs based on their community structure. *IEEE Trans. Vis. Comput. Graph.*, 19(6):1034–1047, 2013.

10 Jaakko Peltonen and Ziyuan Lin. Information retrieval approach to meta-visualization. *Machine Learning*, 99(2):189–229, 2015.

11 Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI 2015, Atlanta, GA, USA, March 29 to April 01, 2015*, pages 38–49, 2015.

12 Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *VL*, pages 336–343, 1996.

13 Melanie Tory and Torsten Möller. Rethinking visualization: A high-level taxonomy. In *10th IEEE Symposium on Information Visualization (InfoVis 2004), 10-12 October 2004, Austin, TX, USA*, pages 151–158, 2004.

14 Christopher M. Bishop. A new framework for machine learning. In *Computational Intelligence: Research Frontiers, IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong, China, June 1-6, 2008, Plenary/Invited Lectures*, pages 1–24, 2008.

15 Leanna House, Scotland Leman, and Chao Han. Bayesian visual analytics: Bava. *Statistical Analysis and Data Mining*, 8(1):1–13, 2015.

16 Catherine Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI'04, pages 109–116, New York, NY, USA, 2004. ACM.

17 D. J. Lehmann, S. Hundt, and H. Theisel. A study on quality metrics vs. human perception: Can visual measures help us to filter visualizations of interest? *it – Information Technology*, 57, 2015 2015.

18 John Aldo Lee and Michel Verleysen. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters*, 31(14):2248–2257, 2010.

19 John Aldo Lee and Michel Verleysen. Two key properties of dimensionality reduction methods. In *2014 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2014, Orlando, FL, USA, December 9-12, 2014*, pages 163–170, 2014.

20 Bradley Efron. *The Jackknife, the bootstrap and other resampling plans.* CBMS-NSF Reg. Conf. Ser. Appl. Math. SIAM, Philadelphia, PA, 1982. Lectures given at Bowling Green State Univ., June 1980.

21 Byron Ellis. *Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data.* Wiley Publishing, 1st edition, 2014.

22 Lauro Lins, James T. Klosowski, and Carlos Scheidegger. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2456–2465, 2013.

23 Zhirong Yang, Jaakko Peltonen, and Samuel Kaski. Scalable optimization of neighbor embedding for visualization. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 127–135, 2013.

24 Laurens van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

25 J. Ellson, E.R. Gansner, E. Koutsofios, S.C. North, and G. Woodhull. Graphviz and dynagraph – static and dynamic graph drawing tools. In M. Junger and P. Mutzel, editors, *Graph Drawing Software*, Mathematics and Visualization, pages 127–148. Springer-Verlag, Berlin/Heidelberg, 2004.

26 O.D. Lampe and H. Hauser. Interactive visualization of streaming data with kernel density estimation. In *Pacific Visualization Symposium (PacificVis), 2011 IEEE*, pages 171–178, March 2011.

27 Robi Polikar and Cesare Alippi. Guest editorial learning in nonstationary and evolving environments. *IEEE Trans. Neural Netw. Learning Syst.*, 25(1):9–11, 2014.

28 Joseph F. Murray, Gordon F. Hughes, and Dale Schuurmans. Machine learning methods for predicting failures in hard drives: A multiple-instance application. *Journal of Machine Learning research*, 6:816, 2005.

29 Ping Chen, Jing Yang, and Linyuan Li. Synthetic detection of change point and outliers in bilinear time series models. *Int. J. Systems Science*, 46(2):284–293, 2015.

30 Lei Le, Emilio Ferrara, and Alessandro Flammini. On predictability of rare events leveraging social media: a machine learning perspective. *CoRR*, abs/1502.05886, 2015.

31 Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Computer*

Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on, pages 178–178, June 2004.

32 Ivo Pedruzzi, Catherine Rivoire, Andrea H. Auchincloss, Elisabeth Coudert, Guillaume Keller, Edouard De Castro, Delphine Baratin, Béatrice A. Cuche, Lydie Bougueleret, Sylvain Poux, Nicole Redaschi, Ioannis Xenarios, and Alan Bridge. Hamap in 2013, new developments in the protein family classification and annotation system. *Nucleic Acids Research*, 41(Database-Issue):584–589, 2013.

33 Dengsheng Zhang, Md. Monirul Islam, and Guojun Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346–362, 2012.

34 Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv.org preprint*, arXiv:1412.3773 [cs.LG], December 2014. Submitted to Journal of Machine Learning Research.

35 Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen, and Takashi Washio. Parcelingam: A causal ordering method robust against latent confounders. *Neural Computation*, 26(1):57–83, 2014.

36 Hao Zhang, Maoyuan Sun, Danfeng (Daphne) Yao, and Chris North. Visualizing traffic causality for analyzing network anomalies. In *Proceedings of the 2015 ACM International Workshop on International Workshop on Security and Privacy Analytics*, IWSPA'15, pages 37–42, New York, NY, USA, 2015. ACM.

37 Nivedita R. Kadaba, Student Member, Pourang P. Irani, and Jason Leboe. Visualizing causal semantics using animations. In *IEEE Transactions on Visualization and Computer Graphics*, 2007.

38 Robert Kosara and Jock D. Mackinlay. Storytelling: The next step for visualization. *IEEE Computer*, 46(5):44–50, 2013.

39 Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE T. Autonomous Mental Development*, 2(3):230–247, 2010.

40 Thomas Hofmann and Justin Basilico. Collaborative machine learning. In *From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments, Essays Dedicated to Erich J. Neuhold on the Occasion of His 65th Birthday*, pages 173–182, 2005.

41 Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October 2010.

42 Kristi Morton, Magdalena Balazinska, Dan Grossman, Robert Kosara, and Jock D. Mackinlay. Public data and visualizations: How are many eyes and tableau public used for collaborative analytics? *SIGMOD Record*, 43(2):17–22, 2014.

4.6 Reproducibility and interpretability

Helwig Hauser (University of Bergen, NO), Bongshin Lee (Microsoft Research – Redmond, US), Torsten Möller (Universität Wien, AT), Tamara Munzner (University of British Columbia – Vancouver, CA), Fernando Paulovich (University of Sao Paulo, BR), Frank-Michael Schleif (University of Birmingham, GB), and Michel Verleysen (Université Catholique de Louvain, BE)

Evaluating visualizations and visualization tools is a broad area and is at the heart of visualization research. Especially considering that a visualization requires a human to be understood and used, the focus has always been on how to evaluate the user experience. However, visualization research uses more and more sophisticated algorithms including some developed in the field of machine learning. Most of these algorithms have a stochastic nature, which makes that their result (or output) may depend on various settings, such as the small variations in the data, some random initialization or random step in an optimization procedure, etc. Therefore human evaluation of visualizations include various elements related on one side to the human nature of evaluations, and on the other side to the stochastic nature of the methods. The discussion in the group during the Dagstuhl seminar has concentrated on 1) how to distinguish these two aspects, and 2) what are really the different effects that have to be measured, in terms of robustness, generalizability, stability, etc.

Evaluation of visual data analysis tools can thus be viewed under a holistic perspective. Let us consider the the process of (visual) data analysis as a special type of algorithm. It takes inputs just like any other algorithm in form of data and/or parameters. Its output is some type of number or other complex entity (as is common for any algorithm). Sometimes this output will be some kind of decision made by the user, and hence it could be seen as a classification (into 0 or 1 or any other class of possible decisions). The only difference would be that while a traditional algorithm would simply be a structured sequence of computer code, the new holistic way of algorithms could include components that are determined by the so-called user-in-the-loop. In order to better distinguish this holistic view from the traditional view, we call these hal-gorithms.

This is akin to the Turing Test. The purpose of the Turing Test is simply to find out whether the algorithm one interacts is purely a machine or has components that can only be performed by a “real” human.

Under this holistic view of an algorithm it makes sense to ask on how to evaluate the quality of this hal-gorithms. With other words, we are considering the question on how different algorithmic performance test would extend to a scenario where the human is an integral part of the hal-gorithms. Again the evaluation of the quality necessitates to distinguish between performances (or differences of performances) that result from the algorithm itself, or from the user-in-the-loop supplementary layer.

The discussion during the Dagstuhl seminar has also covered terminology. Words such as robustness, stability, generalizability and sensitivity are sometimes used without having in mind a clear definition of their respective meaning and differences. Some can cover various situations too. The following is a first attempt to clarify both the terminology and its use in the holistic context.

4.6.1 Robustness of algorithms

The term *robustness* with respect to algorithms refers to the ability of an algorithm to gracefully handle any type of input. For instance the robustness of an algorithm with regards to outliers is of great concern. The concept of robustness is not far from the concept of stability (described below),

4.6.2 Robustness of hal-gorithms

Transferring the concept of robustness to hal-gorithms can have different meanings. For example if the target visual data analysis tool was created for a specific user group ('experts'), will it handled users that are not part of this group gracefully?

4.6.3 Generalizability of algorithms

The concept of *generalizability* of an algorithm is a contribution of the machine learning community. The idea is that train the algorithm (i.e. estimate optimal parameter settings) on a small subset of the known data. The performance of the algorithm is the evaluated on a hold-out set, which allows estimating how the algorithm would *generalize* to a greater set of possible (unknown) data. Estimating how an algorithm generalizes gives some indication on how to choose between several algorithms or settings.

4.6.4 Generalizability of hal-gorithms

The concept of generalizability is not new to the visualization community. The "User Performance" and "User Experience" evaluation methods speak exactly to aspects of understanding visual encoding principles by a larger set of users. However, there is a difficulty of properly testing relatively complex (visual analysis) tools. Often times there are too many confounding factors to consider. On the other hand, many tools are created for specific applications and particular experts. Having access to a larger number of these specific users is often not possible. Further, it is often not feasible to create multiple tools for different subsets of these users (the 'training' user set). Hence, during the design of a visual analysis tool (often referred to as a Design Study) the algorithm / tool is iterated upon and refined with a set of particular users one is working with. Hence, the generalizability of these tools is not tested.

4.6.5 Stability analysis of algorithms

Stability analysis is a term that often refers to the numerical stability of algorithms or discretization schemes. It is tied to the analysis of errors in the numerical computation. Hence, it is tied to the propagation of errors over several iterations. If the errors increase, the algorithm is numerically unstable. If the errors decrease, the algorithm is stable and often an analysis of the speed of convergence is followed. Even without 'errors' stability issues may be encountered due to the stochastic nature of data. On the other hand if 'errors' also refer to possible small variations in the data, the stability concept is not far from the robustness concept, and from the sensitivity one. Stability can also be related to the objective function: does the results of an algorithm change significantly if the objective function (the criterion that is optimized by the algorithm) is slightly modified?

4.6.6 Stability analysis of hal-gorithms

In cases where the generalizability of hal-gorithms can not be tested, perhaps a restricted view can be taken and a stability analysis can be performed. I.e. perhaps it can be well defined under what conditions and circumstances our hal-gorithms can be guaranteed to perform well. Further, one can ask whether several users working together come to an answer faster or to a better answer.

4.6.7 Sensitivity analysis of algorithms

Last but not least, *sensitivity analysis* is a branch of statistics that considers the change of outputs with respect to the inputs. Here, one distinguishes between global sensitivity analysis and local sensitivity analysis. Global sensitivity analysis is considering the possible change in outputs over all possible input variables by constraining just one input. On the other hand, local sensitivity analysis constraints all inputs to a specific value and analysis the change of output with respect to a small change in input of one of the inputs.

4.6.8 Sensitivity analysis of hal-gorithms

Sensitivity analysis is perhaps the most interesting and neglected aspect of hal-gorithms. How does the result change if the particular user using the visual analysis system changes?

Participants

- Daniel Archambault
Swansea University, GB
- Francois Blayo
Ipseite SA – Lausanne, CH
- Kerstin Bunte
UC Louvain-la-Neuve, BE
- Miguel Á. Carreira-Perpiñán
Univ. of California – Merced, US
- Ignacio Díaz Blanco
University of Oviedo, ES
- David S. Ebert
Purdue University – West Lafayette, US
- Alex Endert
Georgia Inst. of Technology, US
- Thomas Ertl
Universität Stuttgart, DE
- Barbara Hammer
Universität Bielefeld, DE
- Helwig Hauser
University of Bergen, NO
- Stephen Ingram
University of British Columbia – Vancouver, CA
- Samuel Kaski
Aalto University, FI
- Daniel A. Keim
Universität Konstanz, DE
- Bongshin Lee
Microsoft Res. – Redmond, US
- John A. Lee
UC Louvain-la-Neuve, BE
- Torsten Möller
Universität Wien, AT
- Bassam Mokbel
Universität Bielefeld, DE
- Tamara Munzner
University of British Columbia – Vancouver, CA
- Ian Nabney
Aston Univ. – Birmingham, GB
- Stephen North
Infovisible – Oldwick, US
- Eli Parviainen
Aalto University, FI
- Fernando Paulovich
University of Sao Paulo, BR
- Jaakko Peltonen
Aalto University / University of Tampere, FI
- William Ribarsky
University of North Carolina – Charlotte, US
- Fabrice Rossi
University of Paris I, FR
- Frank-Michael Schleif
University of Birmingham, GB
- Michael Sedlmair
Universität Wien, AT
- Cagatay Turkay
City University – London, GB
- Jarke J. van Wijk
TU Eindhoven, NL
- Michel Verleysen
University of Louvain, BE
- Thomas Villmann
Hochschule Mittweida, DE
- Daniel Weiskopf
Universität Stuttgart, DE
- William Wong
Middlesex University, GB
- Jing Yang
University of North Carolina – Charlotte, US
- Leishi Zhang
Middlesex University, GB
- Blaz Zupan
University of Ljubljana, SI

Report from Dagstuhl Seminar 15102

Secure Routing for Future Communication Networks

Edited by

Amir Herzberg[1], Matthias Hollick[2], and Adrian Perrig[3]

1 **Bar-Ilan University – Ramat Gan, IL,** amir.herzberg@gmail.com

2 **TU Darmstadt, DE,** mhollick@seemoo.tu-darmstadt.de

3 **ETH Zürich, CH,** adrian.perrig@inf.ethz.ch

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15102 "Secure Routing for Future Communication Networks". Routing is a fundamental mechanism in communication networks, and its security is critical to ensure availability and prevent attacks; however, developing and deploying secure routing mechanism is still a challenge. Significant research effort is required to advance routing security in key areas: intra-domain routing, inter-domain routing, routing in new Internet architectures, and routing in mobile and wireless networks. The seminar covered these general aspects along with the following important guiding questions. How to systematise the topic area of routing security? What are evolutionary or revolutionary options towards more secure routing systems? How to secure inter-domain routing? How to secure intra-domain routing and routing in mobile/wireless settings? How to achieve data plane/forwarding security?

Seminar March 1–4, 2015 – http://www.dagstuhl.de/15102

1998 ACM Subject Classification C.2.0 General, C.2.1 Network Architecture and Design, C.2.2 Network Protocols

Keywords and phrases Security, Secure routing, Communication networks, Future internet, Mobile and wireless networks

Digital Object Identifier 10.4230/DagRep.5.3.28

Edited in cooperation with Michael Noisternig

1 Executive Summary

Amir Herzberg
Matthias Hollick
Adrian Perrig

Routing is a fundamental mechanism in communication networks, and its security is critical to ensure availability and to prevent attacks; however, developing and deploying secure routing mechanisms is still a challenge. Routing is the process by which information is passed via the communication network, from source to destination, via a series of intermediary nodes/routers. Routing attacks include route-hijacking, i.e., diverting traffic to an adversary-controlled router, and denial-of-service attacks exploiting the routing mechanism, i.e., preventing communication (in parts or the entire network), e.g., by malicious dropping of packets by a router.

Routing, and even more secure routing, are complex problems with many variants. In particular, the Internet is a federation of many domains (usually referred to as autonomous systems (ASes)), each managed by a separate organization; there are separate standard

Secure Routing for Future Communication Networks, *Dagstuhl Reports*, Vol. 5, Issue 3, pp. 28–40
Editors: Amir Herzberg, Matthias Hollick, and Adrian Perrig
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

protocols for routing inside an AS (intra-domain routing) and for routing from a source in one AS to a destination in a different AS (inter-domain routing). Significant efforts are dedicated to securing intra-domain routing protocols and inter-domain routing protocols; in addition, significant efforts are also dedicated to the design of completely new Internet architectures that include secure routing mechanisms.

Another categorization of routing mechanisms and challenges involves mobility. Many routing protocols, including standard Internet routing, are designed for a mostly static topology, where connections between routers are relatively stable. However, communication is increasingly performed among mobile devices. There are many efforts and challenges in the design of (secure) routing mechanisms for highly mobile networks, e.g., between tiny wireless sensors, swarms of tiny robots, or simply mobile users (e.g., upon catastrophic failure to regular infrastructure).

There is also a need to re-evaluate and possibly re-design routing mechanisms and security measures, to address changes in the way the Internet is used, and in the presence of new security challenges. In particular, is there a need to adapt routing to facilitate, and/or take advantage of, cloud services, and to support security for them? Is there a need to adapt routing to the increased threat of Denial-of-Service attacks, or to facilitate widespread provision of Quality-of-Service? Should routing be modified to take into account energy considerations, or to take advantage of and facilitate Software Defined Networking (SDN)? If modifications are made for these goals, how does this affect routing systems' attack surface? Finally, is there a need to modify routing and its security mechanisms, as a result of the recent revelations regarding the scope of abuse of routing by powerful nation-state adversaries?

In summary, to advance routing security in the aforementioned topic areas, a number of significant research problems need to be addressed, and identifying these problems was the goal of this seminar. The first objective was to facilitate brainstorming and exchange of ideas among experts working in different areas and types of secure networking, leading to an improved understanding of the different aspects of secure routing. The second objective was to identify the most important research challenges and to devise a roadmap towards addressing urgent issues. Through the seminar, we aimed at opening up new avenues of research in the area of routing security. For the given focus areas of the seminar, we contributed to the following key research challenges:

- Routing Security by Design for a Future Internet: the challenge was to overcome the limitations and confined models imposed by today's Internet. Both clean slate as well as evolutionary approaches towards a secure-by-design future Internet were discussed.
- Inter-domain Routing Security and Intra-domain Routing Security: challenges addressed in inter-domain routing were the reconciliation of potentially conflicting security interests across multiple domains and resilience against recently published attacks. Intra-domain routing is underrepresented in research; here, the seminar aimed at identifying the key research challenges towards a research roadmap.
- Routing Security in Mobile/Wireless Networks, and in Delay- and Disruption-tolerant Networks: the main goal within the seminar was to identify possible ways to provide routing security in light of the severely limited resources and special characteristics of mobile and wireless systems.
- Quality of Service (QoS) and Denial of Service (DoS) aspects of Routing Security: the challenge was to jointly consider security considerations and QoS aspects, both in theory and practice.

To address these challenges, the seminar was organized in six working groups. They are presented in Section 4 of this report. The schedule of the seminar and its working groups is presented in Table 1 below.

Table 1 Schedule for Dagstuhl Seminar 15102.

	Sunday	Monday	Tuesday	Wednesday
Breakfast from 7:30–8:45				
9:00–10:30		+ Welcome by Matthias + Intro of participants + Intro talk by Steven	Wrap-up & update Plenary discussion: A, B, F	Wrap-up & update Plenary discussion: C, D, E
Coffee				
10:50–12:10		+ Talk by Adrian Perrig + Talk by Randy + Agenda, plan, goals by Amir Picture of group	Parallel sessions C – Inter-domain D – Mobile/Wireless E – QoS/DoS, Forwarding security	Wrap-up & reserved
Lunch (12:15)				or lunch boxes for early departure
13:30–15:30		Parallel sessions A – Taxonomy B – (R)Evolution	Social event: 1h hiking/biking around Dagstuhl castle from 14:30 Parallel sessions contd.: C – Inter-domain D – Mobile/Wireless E – QoS/DoS, Forwarding security	Departure
Coffee	Arrival, registration from 15:00 to 19:00			
15:50–17:50		Parallel sessions contd.: A – Taxonomy F – Intra-domain	Parallel sessions contd.: C – Inter-domain D – Mobile/Wireless E – QoS/DoS, Forwarding security	
Dinner (18:00)				
	Cafe, wine, cheese, discussions in cafeteria, wine cellar, music room, etc.			

WG (A) Towards a taxonomy on secure routing
WG (B) Revolution and/or evolution?
WG (C) Securing inter-domain routing
WG (D) Routing security in mobile/wireless networks and delay-/disruption-tolerant networks
WG (E) Forwarding/data-plane security
WG (F) Intra-domain routing security

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

3.1 Routing is as Insecure as the Rest of the Flippin' Internet, but it's Scarier

Randy Bush – Internet Initiative Japan Inc., JP

The goal of the opening talk is to raise lessons from other (Internet) evolutions and revolutions to make us aware of the pitfalls such as 'second-system syndrome incompatibility'. Routing protocols need to protect their assets (traffic content and meta data) from threats such as traffic content inspection, modification, injection, and analysis. Both routing and DNS attacks are today's most severe security issues on the Internet, but routing threats are on the rise. Attackers are primarily spammers, but also governments and financial institutions are involved. Routing attacks target external infrastructure (e.g., IRR, Whois, RPKI) and both well-implemented and poorly designed protocols. This is a disaster happening every day but a cure is difficult to deploy. Security solutions need to provide a real benefit to some involved entities, and deployment must be simple and backwards compatible. Lessons may be learnt from bad examples such as IPv6 and better ones such as RPKI.

3.2 Routing Security Challenges

Steven Bellovin – Columbia University, US

Routing security is hard because failures occur when someone lies. The protocol, however, is executed correctly, and the liar may be distant. Attackers today are definitely spammers, but governments may be involved. Attackers can lie about prefixes for paths. It can happen internally or externally. There are currently no good incentives for deployment in the interdomain case.

3.3 SCION: A Secure Next-Generation Internet Architecture

Adrian Perrig – ETH Zürich, CH

The Internet has been successful beyond even the most optimistic expectations. This success has created a dependency on communication. Unfortunately, the current Internet suffers from numerous vulernabilities and shortcomings that limits its availability. To address these issues, we study the design of a next-generation Internet architecture that is secure, available, and offers privacy by design; that provides incentives for a transition to the new architecture; and that considers economic and policy issues at the design stage.

4 Working Groups

Core topics of the seminar have been organized in six working groups. Each working group lasted for one or two sessions of 120 minutes each and was running with one or two other working groups in parallel (one working group was merged with another one due to a low number of participants). Working group sessions were followed by a wrap-up session the next day, in which the outcome was presented to all seminar participants. The following subsections provide a summary of the discussions within each working group.

4.1 Towards a Taxonomy on Secure Routing

This working group has explored how secure routing protocols could be categorized towards a better understanding of existing solutions and the areas that need improvement. Particular focus has been put on identifying goals of secure routing protocols as this is an area that currently is poorly specified.

Principal classifiers for secure routing protocols that have been identified include

- attacks/vulnerabilities/risks on the routing protocol,
- security solutions,
- the routing abstraction/model/formalization/environment used, and
- goals at the control and data plane.

It is important that goals are clearly specified as part of a system specification. In terms of routing abstractions/models/formalizations, one may consider

- information-centric/content-centric networking (e.g., in Future Internet),
- software-defined networking (makes secure routing even more complex, but at the same time there is a chance to do things exactly as they should be),
- traditional Internet routing (intra- and inter-), and
- mobile/wireless routing.

These properties form what was called the environment of the routing protocol.

Goals

The participants have determined that a taxonomy of current routing protocols based on goals is difficult because today there are no formalisms for (routing) protocols. People talk about safety, aliveness, lightness, etc., but these issues are not defined, and models do not specify what is expected from networks. For example, a goal such as availability seems clear, but under which circumstances is usually not specified. Routing protocols typically provide a best-effort service under non-adversarial circumstances but specifications are not very clear what can be expected under which (threat) scenarios. This issue has been identified by the working group as an open problem on the network specification/modelling side.

Thus, what is needed is a model that clearly specifies what to be expected. To this end, goals need to be separately specified for the data, control, and management plane, and they may be different for different network entities (such as source vs. intermediate node vs. destination) defined by the model.

The following goals or requirements for a secure routing protocol have been identified, distinguishing the data plane from the control plane and the management plane:

- Data plane:
  - It should provide a certain service, which may be best-effort, reliable communication, or reliable communication with performance guarantees.
  - It may provide (end-to-end) path enforcement.
  - There are a number of common security/protection goals, which include confidentiality, integrity, and authenticity (CIA) as well as privacy.
  - Other goals that have been identified included net neutrality and censorship resilience. What is meant by the latter needs to be better understood.
- Control plane:
  - We may distinguish between two aspects:
    - centralized (e.g., SDN) vs. decentralized approaches, and
    - whether the routing protocol implements local state, global state, or is stateless.

    Examples include distance-vector routing, which is decentralized with local routing state, and link-state routing, which is decentralized but acquires a global state.
  - Common security goals include confidentiality, integrity, and authenticity (which may be different between the control and the data plane).
  - Freshness and trustworthiness/completeness for a global routing state. Trustworthiness in this context is not about integrity but about making sure that we can trust the global state that we aggregate.
- Management plane:
  - It must be reliable and timely.
  - In general, goals are expected to be similar to those of the data plane.

The control plane may cover one or more administrative domains, which needs to be taken into account. At the same time, there are routing protocols without a control plane (e.g., flooding). The management plane, which is used mostly for monitoring the data and control planes (but also for configuration and reboot), may implement message routing via dedicated wires or some 'obfuscated' mechanism and may be multi- or single-hop. Attacks on the management plane are similar to those on the data plane though the constraints are tighter.

A worthwhile reference that was mentioned is the paper [1], which provides a taxonomy of common concepts and definitions for security goals in communication. The paper also motivates looking at the boundary between the routing system and the environment. For example, the attack surface depends on system boundaries, e.g., malicious hardware of router, or attacks at interfaces such as power-attack/connection to 3G/4G. Another important issue at the boundary is how to bootstrap the crypto (i.e., how key management is provided).

In terms of threats, it may be interesting to look at a study published by the European Union Agency for Network and Information Security (ENISA) [2], which summarizes good practices that aim at securing an Internet infrastructure asset from Important Specific Threats. The study includes a mind map, which graphically categorizes Internet infrastructure assets into eight families: protocols, software, hardware, information, human resources, facilities, interconnection, and services; the latter which is split into four subfamilies: applications, routing, addressing, and security.

A side question was whether/how vendors would be incentivized to work with the community to implement these goals. While this has been identified as a separate issue, a clear specification of goals would at least allow checking whether the protocol actually provides these goals.

Solutions

Participants in the working group agreed that it is not possible to provide a concise and exhaustive list of solutions that implement the goals that have been identified. Therefore, only a few examples have been discussed. It was assumed that any mechanisms for the data plane must be able to rely on the control plane 'doing its job'. Examples discussed included:

- Confidentiality, integrity, and authenticity can be provided by classical cryptography, both at the control and data plane. Attacks are the 'usual suspects' (side-channel attacks, etc.).
- Reliability with performance guarantees needs to be implemented by
  - Reliability measures such as forward error correction, network coding, and (negative) acknowledgments.
  - Performance measures, which include prioritization, admission control, and resource reservation.
- Trustworthiness/completeness needs to be based on cryptography, reputation mechanisms, heuristics, and invariants.

Future Internet

Goals for the Future Internet have only been briefly discussed. Comments included the need to consider non-interference for parallel stacks and to be aware of downgrade attacks when backwards compatibility is provided, and that routing over hybrid networks may be a challenge.

Outcome

It was decided that it would be a good idea to write a survey paper on this topic. Currently, there are surveys on routing taxonomy, but they are solely on wireless routing. Some older survey that focuses on BGP exists but it does not cover newer developments such as software-defined networking. Reference [3] is also not recent and it does not seem very systematic.

In terms of the content of the paper, it was suggested that it needs to cover (1) the system environment (akin to a trusted computing base) and its boundaries, (2) goals in terms of functionality and security on data, control, and management plane, which then allows us to describe (3) mechanisms to fulfil the goals, and subsequently (4) possible attacks on the mechanisms. It was also suggested that we should consider the industry side, which sees several security issues as rather academic from the viewpoint that dectection is already good enough while protection at the cost of overhead is not worthy. From this perspective, it was agreed that the paper needs to include a clear and realistic adversary model.

4.2 Revolution and/or Evolution?

This working group was concerned with the problem of incrementally deployable improvements in secure communication and routing. It has raised two important issues towards addressing the problem:

- The identification of long-term ideal situations (the "vision"). This bears the question whether the vision is realistic.
- The identification of intermediate steps. The questions here are: what are incentives; and what is the cost?

There are a number of threats in incremental deployments. These include: (1) Not having a vision. The vision may be too vague, too tight, or too ambitious. (2) Not all intermediate steps constitute an improvement ([4]), and partial deployment might even lead to vulnerable conditions. (3) Partial deployment may be hard to achieve, e.g., ingress filtering only prevent others, so there may be no real incentives. (4) Corporate interests may result in conflicting goals.

Ultimately, the goal is to implement the vision, which includes identifying and implementing all intermediate steps.

Studying current deployments, the following systems have been identified: LISP, PKI, spoofing prevention, secure E2E communication (SSH, IPsec, SSL), anonymous communication (Crowds, TOR), digital currencies (cybercash, café, Mondex, Millicent, Bitcoin), secure email, and public-key crypto.

In terms of research activities, the Internet architecture board (IAB) recently held a workshop on Internet technology adoption and transition and published its findings in RFC 7305 [5].

The working group concluded with identifying a challenge for future work, which is to find other incentives for incremental deployments than just the operator's economic ones.

4.3 Securing Inter-Domain Routing

Secure route discovery for inter-domain routing faces a number of challenges, including:

- Privacy of control-plane data (e.g., of relationships between operators)
- Correctness of the control plane (e.g., hijacking of routes must be prevented, i.e., routers must know accurate, trustworthy routes to (all) destinations; and ownership must be asserted)
- Deployment (costs and incentives need to be considered; this may be even more challenging for partial deployment)
- Convergence (stable state vs. optimal state, efficiency)
- Policy (source / destination / transit)
- Trust relationships
- Validation of identifiers

These challenges are not the least due to pressing security threats. Examples include: (1) hijacking of resources (identifiers, links, IPs), (2) topology attacks (path redirection, link cutting), (3) availability attacks (protocol, DDoS, complexity attacks), and (4) privacy attacks ("learning" attack).

To address the challenges, a number of goals have been identified:

- Functionality: this is about reaching equilibrium for convergence.
  - A specific question concerns how to characterize such equilibria (correctness), e.g., correct path establishment to ensure correct packet delivery (if a valid path exists, it should be found), loop-freeness, and prevention of invalid name announcements.
  - Convergence must even be achievable under attack.
- Trust: creation (who do you trust for what), representation, distribution, management, agility & flexibility.
- Policies for source and destination, and for transit.
- Security in case of partial deployment.
- Privacy
  - of the topology, and
  - of the policy.

The working group has then discussed the current state-of-art in secure inter-domain routing in terms of deployment and research. The following deployed systems have been identified: RPKI, Origin Authentication (coming up), and BGPsec (close to being finished)

For BGPsec, a number of problems have been highlighted: RPKI and certificate managements, hierarchies, resource intensive (crypto, memory), DoS (fake withdrawal attacks), partial deployment, convergence, and deployment incentives.

In terms of research, several activities within the context of Future Internet research efforts can be found: (1) SCION (which provides incentives for national/local deployment), (2) FIA, (3) MobilityFirst, (4) ILNP, (5) LISP, and (6) HIP.

The working group concluded with an identification of the following research challenges:

- interdomain multipath routing (beneficial for security)
- diverse paths
- non-hierarchical trust (toleration of malicious nodes)
- impact of security on the system (measurements, verification)
- guaranteed policy-compliant path computation (if path exists, it should be found, understood, and useable)
- centralization of route computation
- finding policy-compliant path
- building network model out of specifications and the formal statements about network properties
- deployment incentives
- routing transparency
- control plane congruence with data plane
- proof of ownership in hierarchy

4.4 Routing Security in Mobile/Wireless Networks and Delay-/Disruption-Tolerant Networks

Due to a limited number of participants, it was decided to join a larger discussion within the working group on a taxonomy of secure routing protocols.

4.5 Forwarding/Data Plane Security

This working group has addressed the issue of providing a forwarding service under QoS requirements and DoS attacks, which has real-world application in the intra-routing domain. Aspects that need to be considered for this service include

- latency and latency variance – this needs to consider limited throughput but does not assume any adversary,
- (local) fast recovery,
- global monitoring, and
- path enforcement (under non-adversarial scenarios).

Fast recovery and global monitoring provide a reliability service but not end-to-end reliability. Current systems are decentralized but are moving towards central control (i.e., SDNs).

From a practical point of view, IPsec tunnels may be used to provide forwarding security. However, they are not used for various reasons. For example, mobile base stations can maintain just a few security associations, so this is a point of tension. Privacy is not considered a strict requirement since a cellular operator can disable encryption because of load (e.g., consider New Year's text message storm). However, privacy in residential applications might be a desirable marketing option.

Concerning DoS attacks and traffic analysis, current system provide no protection against traffic analysis. Existing techniques to mitigate traffic-based DoS attacks include filtering, in particular the redirection of traffic through optimized filters. However, some defenses have undesirable effects on intra-domain routing such as when path security is needed.

A threat model for forwarding security must consider both outsiders and insiders. Attacker capabilities include:

- Outsiders:
  - Traffic injection (possibly coordinated)
  - Eavesdropping (wireless links, only)
- Insiders:
  - Packet delay, dropping, misdirection, injection, corruption (payload and header), replays, etc.
  - Attacks may be directly carried out on the data plane, whereas all attacks on the control plane impact the data plane.
  - The ENISA report on BGP [6] provides a good overview of threats.

In addition to attacks on the communication links, an adversary may attack the operating system and the hardware. These system attacks 'on the box' impact the data plane as an adversary may delay data, drop data, etc. (same as before). Adversaries may also collude (e.g., by setting up a wormhole between two communication devices, which amounts to a control-plane attack affecting the data plane).

Solutions for these threats must provide clear incentives (e.g., financial incentives). This is possible when a security solution is an 'enabler' (e.g., for QoS).

It has been mentioned that while forwarding security is foremost considered for traditional, static networks, mobile wireless networks must be addressed, too. This includes cellular networks, wireless multi-hop networks, and specific settings such as medical networks.

4.6 Intra-Domain Routing Security

Secure intra-domain routing protocols are currently underrepresented in both the academic literature and in deployed systems. The reasons for this unfortunate fact are not obvious; attempts to identify why neither researchers nor industry have sufficiently addressed this important problem resemble a guessing game. In existing intra-domain protocols, however, attackers can easily create damage by launching various attacks.

The problem statement in this domain is thus how to secure intra-domain routing and make it robust against malicious entities, be it routers, end-hosts, or administrators. Such entities may launch a number of attacks including:

- Redirection attacks
  - Path shortening (LSA alteration, bogus links)
  - Prefix announcements
  - Link cutting or congestion
- Availability attacks (delay LSA or drop)

A practical problem that has been identified is that incidents are often not reported (in comparison to the inter-domain case). A first step in addressing intra-domain routing attacks is the deployment of better monitoring techniques.

Acknowledgments. We thank Raphael Reischuk for acting as a scribe during the Dagstuhl seminar and Michael Noisternig for acting as a scribe during the seminar and the collector for this report.

References

1 Avizienis et al., *Basic Concepts and Taxonomy of Dependable and Secure Computing*. IEEE Trans. on Dependable and Secure Computing, vol. 1(1), pp. 11–33. 2004.

2 *Threat Landscape and Good Practice Guide for Internet Infrastructure*. European Union Agency for Network and Information Security (ENISA), Jan. 2015. https://www.enisa.europa.eu/activities/risk-management/evolving-threat-environment/enisa-thematic-landscapes/threat-landscape-of-the-internet-infrastructure

3 Chakrabarti et al., *Internet Infrastructure Security: A Taxonomy*. IEEE Network, vol. 16(6), pp. 13–21. 2002.

4 Lychev et al., *BGP security in partial deployment: is the juice worth the squeeze?*, ACM SIGCOMM Computer Communication Review, vol. 43(4), pp. 171–182. Oct. 2013.

5 Lear (ed.), *Report from the IAB Workshop on Internet Technology Adoption and Transition (ITAT)*, IETF RFC 7305. July 2014.

6 *Secure routing: State-of-the-art deployment and impact on network resilience*. European Network and Information Security Agency (ENISA), July 2010. https://www.enisa.europa.eu/publications/archive/state-of-the-art-deployment-and-impact-on-network-resilience

Participants

- Steven Bellovin
Columbia Univ. – New York, US
- Saleem Bhatti
University of St. Andrews, GB
- Randy Bush
Internet Initiative Japan Inc. – Tokyo, JP
- Joel M. Halpern
Leesburg, US
- Amir Herzberg
Bar-Ilan University - - Ramat Gan, IL
- Matthias Hollick
TU Darmstadt, DE
- Ivan Martinovic
University of Oxford, GB
- Rossella Mattioli
ENISA – Athens, GR
- Cristina Nita-Rotaru
Purdue University – West Lafayette, US
- Michael Noisternig
TU Darmstadt, DE
- Panagiotis Papadimitratos
KTH Royal Institute of Technology, SE
- Adrian Perrig
ETH Zürich, CH
- Raphael Reischuk
ETH Zürich, CH
- Alvaro Retana
CISCO Systems – Research Triangle Park, US
- Michael Schapira
Hebrew Univ. – Jerusalem, IL
- Thomas C. Schmidt
HAW – Hamburg, DE
- Jean-Pierre Seifert
TU Berlin, DE
- Haya Shulman
TU Darmstadt, DE
- Mahesh Tripunitara
University of Waterloo, CA
- Gene Tsudik
Univ. of California – Irvine, US
- Laurent Vanbever
ETH Zürich, CH
- Matthias Wählisch
FU Berlin, DE

Report from Dagstuhl Seminar 15111

Computational Geometry

Edited by

Otfried Cheong[1], Jeff Erickson[2], and Monique Teillaud[3]

1 **KAIST – Daejeon, KR,** otfried@kaist.edu
2 **University of Illinois – Urbana, US,** jeffe@cs.uiuc.edu
3 **INRIA Nancy – Grand Est, FR,** Monique.Teillaud@inria.fr

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15111 "Computational Geometry". The seminar was held from 8th to 13th March 2015 and 41 senior and young researchers from various countries and continents attended it. Recent developments in the field were presented and new challenges in computational geometry were identified.

This report collects abstracts of the talks and a list of open problems.

Seminar March 8–13, 2015 – http://www.dagstuhl.de/15111
1998 ACM Subject Classification F.2 Analysis of Algorithms and Problem Complexity, G.2 Discrete Mathematics, G.4 Mathematical Software
Keywords and phrases Algorithms, geometry, theory, approximation, implementation, combinatorics, topology
Digital Object Identifier 10.4230/DagRep.5.3.41
Edited in cooperation with Maria Saumell

1 Executive Summary

Otfried Cheong
Jeff Erickson
Monique Teillaud

Computational Geometry

Computational geometry is concerned with the design, analysis, and implementation of algorithms for geometric and topological problems, which arise naturally in a wide range of areas, including computer graphics, robotics, geographic information systems, molecular biology, sensor networks, machine learning, data mining, scientific computing, theoretical computer science, and pure mathematics. Computational geometry is a vibrant and mature field of research, with several dedicated international conferences and journals, significant real-world impact, and strong intellectual connections with other computing and mathematics disciplines.

Seminar Topics

The emphasis of the seminar was on presenting recent developments in computational geometry, as well as identifying new challenges, opportunities, and connections to other

Computational Geometry, *Dagstuhl Reports*, Vol. 5, Issue 3, pp. 41–62
Editors: Otfried Cheong, Jeff Erickson, and Monique Teillaud
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

fields of computing. In addition to the usual broad coverage of emerging results in the field, the seminar included invited survey talks on two broad and overlapping focus areas that cover a wide range of both theoretical and practical issues in geometric computing. Both focus areas have seen exciting recent progress and offer numerous opportunities for further cross-disciplinary impact.

Computational topology and topological data analysis. Over the last decade, computational topology has grown from an important subfield of computational geometry into a mature research area in its own right. Results in this field combine classical mathematical techniques from combinatorial, geometric, and algebraic topology with algorithmic tools from computational geometry and optimization. Key developments in this area include algorithms for modeling and reconstructing surfaces from point-cloud data, algorithms for shape matching and classification, topological graph algorithms, new generalizations of persistent homology, practical techniques for experimental low-dimensional topology, and new fundamental results on the computability and complexity of embedding problems. These results have found a wide range of practical applications in computer graphics, computer vision, robotics, sensor networks, molecular biology, data analysis, and experimental mathematics.

Geometric data analysis. Geometric data sets are being generated at an unprecedented scale from many different sources, including digital video cameras, satellites, sensor networks, and physical simulations. The need to manage, analyze, and visualize dynamic, large-scale, high-dimensional, noisy data has raised significant theoretical and practical challenges not addressed by classical geometric algorithms. Key developments in this area include new computational models for massive, dynamic, and distributed geometric data; new techniques for effective dimensionality reduction; approximation algorithms based on coresets and other sampling techniques; algorithms for noisy and uncertain geometric data; and geometric algorithms for information spaces. Results in this area draw on mathematical tools from statistics, linear algebra, functional analysis, metric geometry, geometric and differential topology, and optimization, and they have found practical applications in spatial databases, clustering, shape matching and analysis, machine learning, computer vision, and scientific visualization.

Participants. Dagstuhl seminars on computational geometry have been organized in a two year rhythm since a start in 1990. They have been extremely successful both in disseminating the knowledge and identifying new research thrusts. Many major results in computational geometry were first presented in Dagstuhl seminars, and interactions among the participants at these seminars have led to numerous new results in the field. These seminars have also played an important role in bringing researchers together, fostering collaboration, and exposing young talent to the seniors of the field. They have arguably been the most influential meetings in the field of computational geometry.

The organizers held a *lottery* for the second time this year; the lottery allows to create space to invite younger researchers, rejuvenating the seminar, while keeping a large group of senior and well-known scholars involved. Researchers on the initial list who were not selected by the lottery were notified by us separately per email, so that they knew that they were not forgotten, and to reassure them that—with better luck—they will have another chance in future seminars. The seminar has now a more balanced attendance in terms of seniority and gender than in the past.

This year, 41 researchers from various countries and continents attended the seminar, showing the strong interest of the community for this event. The feedback from participants was very positive.

No other meeting in our field allows young researchers to meet with, get to know, and work with well-known and senior scholars to the extent possible at the Dagstuhl Seminar.

We warmly thank the scientific, administrative and technical staff at Schloss Dagstuhl! Dagstuhl allows people to really meet and socialize, providing them with a wonderful atmosphere of a unique closed and pleasant environment, which is highly beneficial to interactions. Therefore, Schloss Dagstuhl itself is a great strength of the seminar.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 Untraditional Geometric Queries

Peyman Afshani (Aarhus University, DK)

We consider some geometric queries that do not fit in the traditional semigroup searching/reporting model. After a brief review of the classical roots of range searching and existing classical results, we will examine a few recent results (both published and unpublished).

First, we look at recent results on "Concurrent Queries" where each point is associated with nomial data fields (e.g., "color") and the query includes both a geometric region and a list of colors. The output should be the points inside the geometric region with the specified colors. These results were presented in SODA'14 and they are joint works with Bryan Wilkinson, Yufei Tao, Cheng Sheng.

Next, we look at two different geometric queries: range summary queries and range sampling queries. After reviewing their definitions, we will briefly mention some yet unpublished results obtained in a joint work with Zhewei Wei.

We will finish with a list of open problems.

3.2 Surface Patches from Unorganized Space Curves

Annamaria Amenta (University of California – Davis, US)

Recent 3D sketch tools produce networks of three-space curves that suggest the contours of shapes. The shapes may be non-manifold, closed three-dimensional, open two-dimensional, or mixed. We describe a system that automatically generates intuitively appealing piecewise-smooth surfaces from such a curve network, and an intelligent user interface for modifying the automatically chosen surface patches. Both the automatic and the semi-automatic parts of the system use a linear algebra representation of the set of surface patches to track the topology. On complicated inputs from ILoveSketch [1], our system allows the user to build the desired surface with just a few mouse-clicks.

References

1 Seok-Hyung Bae, Ravin Balakrishnan, Karan Singh. *ILoveSketch: as-natural-as-possible sketching system for creating 3d curve models*. Proc. UIST'08, pp. 151–160.

3.3 Voronoi Diagrams of Parallel Halflines in 3D

Franz Aurenhammer (TU Graz, AT)

The Voronoi diagram for n lines and/or line segments in 3D is a complicated structure. Bisectors are complex geometric objects, and the combinatorial size is still unclear. Things get somewhat easier when the line segments are confined to have only a constant number of orientations. We consider the special case of n parallel (vertical) halflines in 3D. In this case, the intersection of the 3D diagram with any horizontal plane can be shown to be a power diagram of n weighted point sites. This enables us to study the structural properties of the Voronoi diagram of parallel halflines, and to design a relatively simple and output-sensitive algorithm for constructing it.

3.4 Faster DBSCAN and HDBSCAN in Low-Dimensional Euclidean Spaces

Mark de Berg (TU Eindhoven, NL)

DBSCAN is one of the most widely used density-based clustering methods. The clustering it produces depends on two parameters, MinPoints and ε, where MinPoints is typically fixed at a small constant, and ε essentially determines the scale at which we perform the clustering.

We present a new algorithm for DBSCAN in Euclidean spaces, whose running time is much less sensitive to the value of the parameter ε than previous approaches. As a result, our algorithm computes a DBSCAN-clustering in subquadratic time in the worst case when MinPoints is a constant, irrespective of the choice of ε. The worst-case running time of our algorithm in $\mathbb{R}^d$ is $O(n \log n)$ for $d = 2$ and $O(n^{2-\frac{2}{\lceil d/2 \rceil+1}+\gamma})$ for $d \geq 3$, where $\gamma > 0$ is an arbitrarily small constant. Our experiments show that the new algorithm is not only faster in theory, but also in many practical settings.

We also present a novel algorithm for HDBSCAN, a hierarchical version of DBSCAN introduced recently. In $\mathbb{R}^2$ our algorithm computes the HDBSCAN hierarchy in $O(n \log n)$ time in the worst case when MinPoints is a constant.

Finally, we introduce δ-approximate DBSCAN* and δ-approximate HDBSCAN, and we show how to compute these approximate versions of DBSCAN and HDBSCAN in near-linear time in any fixed dimension, for any given approximation error $\delta > 0$.

3.5 Segmentation and Classification of Trajectories

Maike Buchin (Ruhr-Universität Bochum, DE)

We consider segmentation and classification of trajectories, that is splitting and grouping trajectories such that they have similar movement characteristics. Our approach is based on a movement model parameterized by a single parameter, like the Brownian bridge movement model. We define an optimal segmentation (resp. classification) to be one that minimizes an information criterion balancing the likelihood of the model and its size. We give an efficient algorithm to compute the optimal classification for a discrete set of parameter values. For continuous parameters the problem becomes NP-hard. But we also present an algorithm that solves the problem in polynomial time under mild assumptions on the input.

3.6 Shortest Paths on Polyhedral Surfaces and Terrains

Siu-Wing Cheng (HKUST – Kowloon, HK)

Joint work of Cheng, Siu-Wing; Jin, Jiongxin; Vigneron, Antoine
Main reference S.-W. Cheng, J. Jin, "Shortest paths on polyhedral surfaces and terrains," in Proc. of the 46th Annual ACM Symp. on Theory of Computing (STOC'14), pp. 373–382, ACM, 2014.
URL http://doi.acm.org/10.1145/2591796.2591821

We present an algorithm for computing shortest paths on polyhedral surfaces under convex distance functions. Let n be the total number of vertices, edges and faces of the surface. Our algorithm can be used to compute L_1 and L_∞ shortest paths on a polyhedral surface in $O(n^2 \log^4 n)$ time. Given an $\epsilon \in (0, 1)$, our algorithm can find $(1 + \epsilon)$-approximate shortest paths on a terrain with gradient constraints and under cost functions that are linear combinations of path length and total ascent. The running time is $O(\frac{1}{\sqrt{\epsilon}} n^2 \log n + n^2 \log^2 n \log^2(n/\epsilon)$. This is the first efficient PTAS for such a general setting of terrain navigation.

3.7 Walking in Random Delaunay Triangulations

Olivier Devillers (INRIA Nancy – Grand Est, FR)

Joint work of Broutin, Nicolas; Devillers, Olivier; Hemsley, Ross
Main reference N. Broutin, O. Devillers, R. Hemsley, "Efficiently navigating a random Delaunay triangulation," in Proc. of the 25th Int'l Conf. on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA'14), DMTCS-HAL Proceedings Series, pp. 49–60, 2014.
URL https://hal.inria.fr/hal-01077251

Walking in triangulation is a widely used strategy for point location in triangulation. There are several strategies to walk between neighboring vertices or neighboring faces of a triangulation, but the analysis of such strategies under random distribution hypotheses for the point set is very difficult. This is due to the fact that the probability for an edge to be part of the walk depends on the whole set of points, thus you get dependence between these probabilities that are difficult to deal with. All these kind of walks are conjectured to have length $O(\sqrt{n})$. We propose the analysis of two walking strategies.

The cone walk is a walk amongst vertices where the dependence is reduced. The visibility is the most commonly used strategy to walk amongst faces and we analyze it using percolation theory.

3.8 Toward Parameter-Free (Friendly?) Topology Inference

Tamal K. Dey (Ohio State University – Columbus, US)

Joint work of Dey, Tamal K.; Dong, Zhe; Wang, Yusu

In topological inference from point data, a simplicial complex such as Vietoris-Rips is built on top of the data to carry out the topological analysis. This requires a user-supplied global parameter, which in some cases may be impossible to determine for the purpose of correct topology inference. We show that when the underlying space is a smooth manifold of known dimension embedded in an Euclidean space, a parameter-free sparsification of the data leads to a correct homology inference. This follows from the fact that we can compute a function called lean-set feature size over the data points with which it can be made locally uniform. The construction of the Vietoris-Rips complex on such data can be done adaptively without requiring any user-supplied parameter from which homology of the hidden manifold can be inferred. Preliminary experiments suggest that the strategy achieves correct topological (homology) inference with effective sparsification in practice.

3.9 Realization Spaces of Arrangements of Convex Bodies

Michael Gene Dobbins (Postech – Pohang, KR)

Joint work of Dobbins, Michael Gene; Holmsen, Andreas; Hubard, Alfredo
Main reference M. G. Dobbins, A. Holmsen, A. Hubard, "Realization spaces of arrangements of convex bodies," arXiv:1412.0371v2 [math.MG], 2015.
URL http://arxiv.org/abs/1412.0371v2

In this talk I introduce combinatorial types of arrangements of convex bodies, extending order types of point sets to arrangements of convex bodies, and present some results on their realization spaces. Our main results witness a trade-off between the combinatorial complexity of the bodies and the topological complexity of their realization space. First, we show that every combinatorial type is realizable and its realization space is contractible under mild assumptions. Second, we prove a universality theorem that says the restriction of the realization space to arrangements polygons with a bounded number of vertices can have the homotopy type of any primary semialgebraic set. This is joint work with Andreas Holmsen and Alfredo Hubard.

3.10 Clustering Time Series under the Frechet Distance

Anne Driemel (TU Eindhoven, NL)

The Frechet distance is a popular distance measure for curves. We study the problem of clustering time series under the Frechet distance. In particular, we give $(1+\varepsilon)$-approximation algorithms for variations of the following problem with parameters k and l. Given n univariate time series P, each of complexity at most m, we find k time series, not necessarily from P, which we call cluster centers and which each have complexity at most l, such that (a) the maximum distance of an element of P to its nearest cluster center or (b) the sum of these distances is minimized. Our algorithms have running time near-linear in the input size. To the best of our knowledge, our algorithms are the first clustering algorithms for the Frechet distance which achieve an approximation factor of $(1+\varepsilon)$ or better.

3.11 Low-quality Dimension Reduction and High-dimensional Approximate Nearest Neighbor

Ioannis Z. Emiris (University of Athens, GR)

Joint work of Anagnostopoulos, Evangelos; Emiris, Ioannis Z.; Psarros, Ioannis
Main reference E. Anagnostopoulos, I. Z. Emiris, I. Psarros, "Low-quality dimension reduction and high-dimensional Approximate Nearest Neighbor," in Proc. of the 31st Int'l Symp. on Computational Geometry (SoCG'14), LIPIcs, Vol. 34, pp. 436–450, Schoss Dagstuhl – Leibniz-Zentrum für Informatik, 2015; pre-print available at arXiv:1412.1683v1 [cs.CG].
URL http://dx.doi.org/10.4230/LIPIcs.SOCG.2015.436
URL http://arxiv.org/abs/1412.1683v1

The approximate nearest neighbor problem (ϵ-ANN) in a Euclidean space is a fundamental question, which has been addressed by two main approaches: Data-dependent space partitioning techniques, typically tree-based such as kd-trees or BBD-trees, perform well when the dimension is bounded, but are affected by the curse of dimensionality. On the other hand, Locality Sensitive Hashing (LSH) has polynomial dependence in the dimension, sublinear query time with an exponent inversely proportional to $(1+\epsilon)^2$, and subquadratic space requirement.

In this paper, we generalize the celebrated Johnson-Lindenstrauss Lemma to define "low-quality" mappings to a Euclidean space of significantly lower dimension than previously considered, such that they satisfy a requirement weaker than approximately preserving all distances or even preserving the nearest neighbor. This mapping guarantees, with high probability, that an ANN lies among the k ANN's in the projected space: the latter can be efficiently retrieved by a tree-based data structure, such as BBD-trees. Our algorithm, given n points in dimension d, achieves optimal space usage in $O(dn)$, preprocessing time in $O(dn \log n)$, and query time in $O(dn^\rho \log n)$, where ρ is proportional to $1 - 1/\ln\ln n$, for fixed $\epsilon \in (0,1)$. Moreover, our method is quite simple and easy to implement. The dimension reduction is larger if one assumes that pointsets possess some structure, namely bounded expansion rate.

We implemented our method using projection matrices whose entries are i.i.d. Gaussian variables and solve the k-ANN problem in the projected space by using software library ANN. We present experimental results in up to 500 dimensions and 10^6 points, which show that the practical performance is better than that predicted by the theoretical analysis. In particular, k seems to grow like $\sqrt{n}$ rather than n^ρ. In addition, we compare our approach to E2LSH: our method requires less space but is somewhat slower than E2LSH on the examined datasets.

3.12 The Offset Filtration of Convex Objects

Michael Kerber (MPI für Informatik – Saarbrücken, DE)

Joint work of Kerber, Michael; Halperin, Dan; Shaharabani, Doron
Main reference D. Halperin, M. Kerber, D. Shaharabani, "The Offset Filtration of Convex Objects," arXiv:1407.6132v2 [cs.CG], 2015.
URL http://arxiv.org/abs/1407.6132v2

We consider offsets of a union of convex objects. We aim for a filtration, a sequence of nested simplicial complexes, that captures the topological evolution of the offsets for increasing radii. We describe methods to compute a filtration based on the Voronoi partition with respect to the given convex objects. The size of the filtration and the time complexity for computing it are proportional to the size of the Voronoi diagram and its time complexity, respectively. Our approach is inspired by alpha-complexes for point sets, but requires more involved machinery and analysis primarily since Voronoi regions of general convex objects do not form a good cover. We show by experiments that our approach results in a similarly fast and topologically more stable method for computing a filtration compared to approximating the input by a point sample.

3.13 Minimizing Co-location Potential for Moving Points

David G. Kirkpatrick (University of British Columbia – Vancouver, CA)

Joint work of Kirkpatrick, David G.; Evans, Will; Löffler, Maarten; Staals, Frank; Busto, Daniel

Imagine a collection of entities that move in d-dimensional space each with some bound on their speed. If we know the location of an individual entity at a particular time then its location lies in a region of uncertainty at all subsequent times. We consider the problem of minimizing the ply of the uncertainty regions (defined as the maximum, over all points p in the space, of the number of uncertainty regions that contain p) by means of queries to individual entities that are restricted to one query per unit of time. This notion of co-location potential is studied in two settings, one where ply is measured at some fixed time in the future, and the other where ply is measured continuously (i.e. at all times). Competitive query strategies are described in terms of a notion of intrinsic ply (the minimum ply achievable by any query strategy, even one that knows the trajectories of all entities).

Based on joint work with Will Evans, Maarten Löffler, Frank Staals, and Daniel Busto.

3.14 Fire

Rolf Klein (Universität Bonn, DE)

Joint work of Klein, Rolf; Langetepe, Elmar; Levcopoulos, Christos

Suppose that a circular fire spreads in the plane at unit speed. A fire fighter can build a barrier at speed $v > 1$. How large must v be to ensure that the fire can be contained, and how should the fire fighter proceed? We provide two results. First, we analyze the natural strategy where the fighter keeps building a barrier along the frontier of the expanding fire. We prove that this approach contains the fire if $v > v_c = 2.6144\ldots$ holds. Second, we show that any "spiralling" strategy must have speed $v > 1.618$, the golden ratio, in order to succeed.

3.15 Approximating the Colorful Caratheodory Theorem

Wolfgang Mulzer (FU Berlin, DE)

Joint work of Mulzer, Wolfgang; Stein, Yannik
Main reference W. Mulzer, Y. Stein, "Computational Aspects of the Colorful Caratheodory Theorem," arXiv:1412.3347v1 [cs.CG], 2014.
URL http://arxiv.org/abs/1412.3347v1

Given $d+1$ point sets $P_1, \dots, P_{d+1}$ in $\mathbb{R}^d$ (the color classes) such that each set P_i contains the origin in its convex hull, the colorful Caratheodory theorem states that there is a colorful choice C which also contains the origin in its convex hull. Here, a colorful choice means a set containing at most one point from each color class. So far, the computational complexity of computing such a colorful choice is unknown.

We consider a new notion of approximation: a set C' is called a c-colorful choice if it contains at most c points from each color class. We show that for all $\varepsilon > 0$, an $\varepsilon(d+1)$-colorful choice containing the origin in its convex hull can be found in polynomial time.

3.16 The Cosheaf-Less Reeb Graph Interleaving Distance

Elizabeth Munch (University of Albany, US)

Joint work of de Silva, Vin; Munch, Elizabeth; Patel, Amit
Main reference V. de Silva, E. Munch, A. Patel, "Categorified Reeb Graphs," arXiv:1501.04147v1 [cs.CG], 2015.
URL http://arxiv.org/abs/1501.04147v1

The interleaving distance was recently defined in order to give a method for comparison of Reeb graphs. The definition draws inspiration from the interleaving distance for persistence modules via category theory and cosheaves. Here, we present this distance using the equivalent yet concrete definition which looks for function preserving maps on graphs and checks for commutativaty of a particular diagram. The distance definition also yields as a substep a new definition for the smoothed Reeb graph. This later construction can be performed in polynomial time, while the general computation of the distance is graph isomorphism hard. This is joint work with Vin de Silva and Amit Patel.

3.17 On a Line-symmetric Puzzle

Yota Otachi (JAIST – Ishikawa, JP)

Joint work of Korman, Matias; Otachi, Yota; Roeloffzen, Marcel; Uehara, Ryuhei; van Renssen, André

Given k simple polygons, the goal of the line-symmetric puzzle is to find a polygon that can be exactly covered by the k polygons without overlap. We study the computational complexity of this puzzle and show a hardness result.

3.18 Geometric Data Analysis: Matrix Sketching to Kernels

Jeff M. Phillips (University of Utah – Salt Lake City, US)

I overview some recent developments in geometric data analysis. The initial focus will be in describing how geometric analysis has been essential and central to many core problems in data mining and machine learning. Then I overview recent developments in the area of matrix sketching which has broad applications within these core data mining and machine learning problems. I highlight the geometric connections, recent developments, and broad future directions. Finally, I talk about the uses of kernels and kernel density estimates for geometric data analysis. These enforce certain analyses to be robust, and in some cases have computational advantages. In this area I identify a number of open computational geometry problems which while easy to state may have important implications in data analysis.

3.19 Richter-Thomassen Conjecture about Pairwise Intersecting Curves (and Beyond)

Natan Rubin (Ben Gurion University – Beer Sheva, IL)

A long standing conjecture of Richter and Thomassen states that the total number of intersection points between any n simple closed (i.e., Jordan) curves in the plane which are in general position and any pair of them intersect, is at least $(2 - o(1))n$.

Very recently, we established an even stronger form of the above conjecture, which states that the overall number of proper intersection points must exceed, in asymptotic terms, the number of the touching pairs of curves.

If time permits, we discuss this result in connection with other fundamental questions concerning string graphs and arrangements of curves in the plane.

This is joint work in progress with Janos Pach and Gabor Tardos.

3.20 Controlling Modular Robotic Systems: Some Ideas from Computational Geometry

Vera Sacristan (UPC – Barcelona, ES)

A self-reconfiguring modular robot consists of a large number of independent units that can rearrange themselves into a structure best suited for a given environment or task. For example, it may reconfigure itself into a thin, linear shape to facilitate passage through a narrow tunnel, transform into an emergency structure such as a bridge, or surround and manipulate objects in outer space. Since modular robots comprise groups of identical units, they can also repair themselves by replacing damaged units with functional ones. Such robots are especially well-suited for working in unknown and remote environments.

In this talk I will introduce various types of units for modular robots that have been designed and prototyped by the robotics community, present the current challenges in the field, discuss how computational geometry can help in solving some of them, and present some current results and strategies, as well as open problems.

3.21 A Dynamic Programming Algorithm to Find Subsets of Points in Convex Position Optimizing some Parameter

Maria Saumell (University of West Bohemia – Pilsen, CZ)

Given a set S of n points in the plane, we may consider the problem of finding a subset of S of maximum cardinality such that they are the vertices of a convex polygon and their convex hull is empty of other points of S. This problem can be solved in cubic time by a dynamic programming algorithm [1]. We show that this algorithm can be adapted to solve a variety of other optimization problems related to convex polygons, in particular, the problem of computing largest monochromatic islands in a bicolored point set [2], or the problem of finding cliques of maximum size in the visiblity graph of a simple polygon [3, 4].

References

1 David Avis, David Rappaport. *Computing the largest empty convex subset of a set of points.* Proc. SoCG'85, pp. 161–167.

2 Crevel Bautista-Santiago, José Miguel Díaz-Báñez, Dolores Lara, Pablo Pérez-Lantero, Jorge Urrutia, Inmaculada Ventura. *Computing optimal islands.* Oper. Res. Lett. 39(4):246–251 (2011).

3 Sergio Cabello, Maria Saumell. *A randomized algorithm for finding a maximum clique in the visibility graph of a simple polygon.* Discrete Math. Theor. Comput. Sci. 17(1):1–12 (2015).

4 Sergio Cabello, Josef Cibulka, Jan Kynčl, Maria Saumell, Pavel Valtr. *Peeling potatoes near-optimally in near-linear time.* Proc. SoCG'14, pp. 224–231.

3.22 A Middle Curve Based on Discrete Fréchet Distance

Ludmila Scharf (FU Berlin, DE)

Joint work of Ahn, Hee-Kap; Alt, Helmut; Buchin, Maike; Scharf, Ludmila; Wenk, Carola
Main reference H.-K. Ahn, H. Alt, M. Buchin, L. Scharf, C. Wenk, "A Middle Curve Based on Discrete Fréchet Distance," to appear in Proc. of the 2015 European Workshop on Computational Geometry (EuroCG'15).

Given a set of polygonal curves we seek to find a "middle curve" that represents the set of curves. We ask that the middle curve consists of points of the input curves and that it minimizes the discrete Fréchet distance to the input curves. We develop algorithms for three different variants of this problem.

3.23 On Perturbations of the Expansion Cone

André Schulz (Universität Münster, DE)

Joint work of Schulz, André; Igamberdiev, Alexander

An expansive motion is an assignment of infinitesimal velocities to points in the plane such that all pairwise distances are (infinitesimal) nondecreasing. The space of the infinitesimal velocities forms a polyhedral cone. After a perturbation we obtain a polyhedron, whose corners represent geometric graphs induced by the tight inequalities. One set of perturbation parameters gives the polytope of pointed pseudo-triangulations. We reprove this result and show how a different set of parameters can be used to define a polyhedron whose corners represent a different class of planar Laman graphs. These graphs have no nonempty convex polygon (a necessary but not a sufficient condition). As a consequence we obtain a new description of the associahedron.

3.24 Topological Data Analysis

Donald Sheehy (University of Connecticut – Storrs, US)

URL http://donsheehy.net

I will present a top-down survey of some topics in topological data analysis (TDA). Consider the following model of data analysis.

$$U \longrightarrow (X \to R) \longrightarrow S$$

U is the universe, a population, or some "underlying" thing to be studied. The "data" comes in the form of real-valued functions on some (possibly unknown) space X. S is for signatures or summaries. A major goal of TDA is to define and compute signatures that are "topologically invariant" in the sense that

$$Sig(f(X)) = Sig(f(h(X)))$$

whenever h is a homeomorphism. I will show how many of the known results and many open research directions in TDA can be understood by systematically adding noise, error, discretization, or new hypotheses into this model.

3.25 Restricted Constrained Delaunay Triangulations

Jonathan Shewchuk (University of California – Berkeley, US)

Joint work of Shewchuk, Jonathan; Lévy, Bruno; Khoury, Marc; van Kreveld, Marc

The constrained Delaunay triangulation is a geometric structure that adapts the Delaunay triangulation to enforce the presence of specified edges. The restricted Delaunay triangulation is a geometric structure drawn on a smooth surface embedded in three-dimensional space, having properties similar to those of the Delaunay triangulation in the plane. We combine these two structures to address a question of Bruno Levy: can we define mathematically well-behaved constrained Delaunay triangulations on smooth surfaces?

We define the restricted constrained Delaunay triangulation to be the dual of a restricted extended Voronoi diagram, which is a generalization of the extended Voronoi diagram introduced by Raimund Seidel as a dual of the constrained Delaunay triangulation. The topological space on which we define the restricted extended Voronoi diagram is a 2-manifold created by cutting slits in the input surface (one slit for each specified edge constraint) and gluing two extrusions onto each slit. We define a metric on this 2-manifold that is similar to the three-dimensional Euclidean metric but is modified so that vertices on one "side" of an edge constraint cannot influence the portion of the Voronoi diagram on the other "side". The Voronoi diagram on the 2-manifold under this metric dualizes to a triangulation of the original surface if certain sampling conditions are met.

3.26 Beyond the Euler Characteristic: Approximating the Genus of General Graphs

Anastasios Sidiropoulos (Ohio State University – Columbus, US)

Main reference K.-i. Kawarabayashi, A. Sidiropoulos, "Beyond the Euler Characteristic: Approximating the Genus of General Graphs," in Proc. of the 47th Annual ACM Symp. on Theory of Computing (STOC'15), pp. 675–682, ACM, 2015; pre-print available at arXiv:1412.1792v1 [cs.DS].
URL http://dx.doi.org/10.1145/2746539.2746583
URL http://arxiv.org/abs/1412.1792v1

Computing the Euler genus of a graph is a fundamental problem in graph theory and topology. It has been shown to be NP-hard by [Thomassen 1989] and a linear-time fixed-parameter algorithm has been obtained by [Mohar 1999]. Despite extensive study, the approximability of the Euler genus remains wide open. While the existence of an $O(1)$-approximation is not ruled out, the currently best-known upper bound is a trivial $O(n/g)$-approximation that follows from bounds on the Euler characteristic.

In this paper, we give the first non-trivial approximation algorithm for this problem. Specifically, we present a polynomial-time algorithm which given a graph G of Euler genus g outputs an embedding of G into a surface of Euler genus $g^{O(1)}$. Combined with the above $O(n/g)$-approximation, our result also implies a $O(n^{1-\alpha})$-approximation, for some universal constant $\alpha > 0$.

Our approximation algorithm also has implications for the design of algorithms on graphs of small genus. Several of these algorithms require that an embedding of the graph into a surface of small genus is given as part of the input. Our result implies that many of these algorithms can be implemented even when the embedding of the input graph is unknown.

3.27 The Cosheaf Reeb-graph Interleaving Distance

Vin de Silva (Pomona College – Claremont, US)

Topological data analysis is typically carried out in a persistent framework [1]: a data set is converted to a filtered family of topological spaces, and the homological invariants of this system (rather than of any individual space in the family) are provably stable [2,3]. The family is typically parametrized by a real variable, which represents the scale at which the discrete data set is blurred to make it into a space.

Taking a more general view of persistence [4] as the study of functors on small sites and certain 'interleaving' relationships between them, we see that merge trees and Reeb graphs are susceptible to the same treatment. A merge tree can be viewed as a set-valued functor on the real line, and a Reeb greeph can be viewed as a set-valued cosheaf on the category of real intervals. In both cases there is defined an interleaving metric [5,6] that is provably stable with respect to perturbations of the initial data.

References

1 Herbert Edelsbrunner, David Letscher, Afra Zomorodian. *Topological persistence and simplification.* Discrete & Computational Geometry, 28:511–533 (2002).
2 David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. *Stability of persistence diagrams.* Discrete & Computational Geometry, 37(1):103–120 (2007).
3 Frederic Chazal, David Cohen-Steiner, Marc Glisse, Leonidas Guibas, Steve Oudot. *Proximity of persistence modules and their diagrams.* Proceedings of the 25th annual Symposium on Computational Geometry: 237–246 (2009).
4 Peter Bubenik and Jonathan A. Scott. *Categorification of persistent homology.* Discrete & Computational Geometry, 51(3):600–627 (2014).
5 Kenes Beketayev, Damir Yeliussizov, Dmitriy Morozov, Gunther Weber, Bernd Hamann. *Measuring the distance between merge trees.* Topological Methods in Data Analysis and Visualization III: Theory, Algorithms, and Applications, Mathematics and Visualization, 151–166 (2014).
6 Vin de Silva, Elizabeth Munch, Amit Patel. *Categorified Reeb Graphs.* Unpublished manuscript, arXiv:1501.04147 (2015).

3.28 Augmenting Embedded Paths and Trees to Optimize their Diameter

Fabian Stehn (Universität Bayreuth, DE)

We consider the problem of augmenting a graph with n vertices embedded in a metric space, by inserting one additional edge in order to minimize the diameter of the resulting graph. We present algorithms for the cases when the input graph is a path (running in $O(n \log^3 n)$ time) or a tree (running in $O(n^2 \log n)$ time). For the case when the input graph is a path in $\mathbb{R}^d$, where d is a constant, we present an algorithm that computes a $(1+\varepsilon)$-approximation in $O(n + 1/\varepsilon^3)$ time.

3.29 Flip Distances in Triangulations and Rectangulation

Csaba Toth (California State University – Northridge, US)

Joint work of Ackerman, Eyal; Allen, Michelle; Barequet, Gill; Cardinal, Jean; Hoffmann, Michael; Kusters, Vincent; Löffler, Maarten; Mermelstein, Joshua; Souvaine, Diane; Toth, Csaba; Wettstein, Manuel
Main reference J. Cardinal, M. Hoffmann, V. Kusters, C. D. Tóth, M. Wettstein, "Arc Diagrams, Flip Distances, and Hamiltonian Triangulations," in Proc. of the 32nd Symp. on Theoretical Aspects of Computer Science (STACS'15), LIPIcs, Vol. 30, pp. 197–210, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2015.
URL http://dx.doi.org/10.4230/LIPIcs.STACS.2015.197

It is shown that every triangulation (maximal planar graph) on $n \geq 6$ vertices can be flipped into a Hamiltonian triangulation using a sequence of less than $n/2$ combinatorial edge flips. The previously best upper bound uses 4-connectivity as a means to establish Hamiltonicity. But in general about $3n/5$ flips are necessary to reach a 4-connected triangulation. Our result improves the upper bound on the diameter of the flip graph of combinatorial triangulations on n vertices from $5.2n - 33.6$ to $5n - 23$. We also show that for every triangulation on n vertices there is a simultaneous flip of less than $2n/3$ edges to a 4-connected triangulation. The bound on the number of edges is tight, up to an additive constant.

For n noncorectilinear points in a unit square $[0, 1]^2$, a rectangulation is a subdivision of $[0, 1]^2$ into $n + 1$ rectangles by n axis-aligned line segments, one passing through each point. It is shown that a sequence of $O(n \log n)$ elementary *flip* and *rotate* operations can transform any rectangulation to any other rectangulation on the same set of n points. This bound is the best possible for some point sets, while $\Theta(n)$ operations are sufficient and necessary for others.

3.30 Road Map Construction and Comparison

Carola Wenk (Tulane University, US)

Map construction is a new type of geometric reconstruction problem in which the task is to extract the underlying geometric graph structure described by a set of movement-constrained trajectories, or in other words reconstruct a geometric domain that has been sampled with continuous curves that are subject to noise. Due to the ubiquitous availability of geo-referenced trajectory data, the map construction task has widespread applications ranging from a variety of location-based services on street maps to the analysis of tracking data for hiking trail map generation or for studying social behavior in animals.

Several map construction algorithms have recently been proposed in the literature, however it remains a challenge to measure the quality of the reconstructed maps. We present an incremental map construction algorithm based on the Frechet distance. And we present different distance measures for comparing two road maps which amounts to comparing two uncertain embedded geometric graphs. One approach is based on comparing the set of paths in the graphs, and the other uses persistent homology of the offset filtration to compare the local topology of the graphs. We also introduce local signatures based on these distance measures, which allow us to identify regions where the maps differ the most.

3.31 Completely Randomized RRT-Connect: A Case Study on 3D Rigid Body Motion Planning

Nicola Wolpert (University of Applied Sciences – Stuttgart, DE)

Joint work of Schneider, Daniel; Schoemer, Elmar; Wolpert, Nicola
Main reference D. Schneider, E. Schömer, N. Wolpert, "Completely Randomized RRT-Connect: A Case Study on 3D Rigid Body Motion Planning," in Proc. of the 2015 IEEE Int'l Conf. on Robotics and Automation (ICRA'15), pp. 2944–2950, IEEE, 2015.
URL http://dx.doi.org/10.1109/ICRA.2015.7139602

Nowadays sampling-based motion planners use the power of randomization to compute multidimensional motions at high performance. Nevertheless the performance is based on problem-dependent parameters like the weighting of translation versus rotation and the planning range of the algorithm. Former work uses constant user-adjusted values for these parameters which are defined a priori. Our new approach extends the power of randomization by varying the parameters randomly during runtime. This avoids a preprocessing step to adjust parameters and moreover improves the performance in comparison to existing methods in the majority of the benchmarks. Our method is simple to understand and implement. In order to compare our approach we present a comprehensive experimental analysis about the parameters and the resulting performance. The algorithms and data structures were implemented in our own library RASAND, but we also compare the results of our work with OMPL and the commercial software KineoTM Kite Lab.

4 Open Problems

On Monday evening (19:15–20:30), March 9, 2015, we held an open problem discussion. The session scribe was Joe Mitchell and the session chair was Jeff Erickson. The problems span a range of topics, including fundamental algorithms, discrete geometry, combinatorics, and optimization.

▶ Problem 1 (Don Sheehy). *A metric problem:* Given n points P in $\mathbb{R}^d$. For a curve γ, define $len(\gamma) = \int_\gamma N(x)dx$, where $N(x)$ is the Euclidean distance from the nearest point of P to the point x. Let $d_N(p,q) = \inf_\gamma len(\gamma)$, where the infimum is over all paths starting at p and ending at q. Define $W_{ab} = (1/4)||a-b||^2$, for $a, b \in P$. Define $d_S(p,q) = \inf_{p=v_0,\ldots,v_k=q} \sum_i W_{v_{i-1}v_i}$, where the points v_i are all points of P. Conjecture: $d_N = d_S$.

Note that it is true for 2 points; this is the source of the "1/4" in the definition. From this it follows that the piecewise linear path γ that determines d_S has $len(\gamma) = d_S(p,q)$. So, for all $p, q \in P$, $d_S(p,q) \geq d_N(p,q)$. Moreover, it's easy to check that any edge traversed in the piecewise linear path determining d_S must be Gabriel, i.e. it must have a diametral ball empty of other points of P. This follows from the Pythagorean theorem, as any point inside the diametral ball would create a shortcut and, thus, a shorter path.

The problem is motivated by density-based distances. The metric d_N is a natural density-based distance arising from the nearest neighbor density estimator. We originally believed d_S would be a good approximation, but never found an example where they differ.

▶ Problem 2 (Jeff Erickson). *A question in elementary topology:* Any generic closed curve in the plane can be continuously deformed into a simple closed curve through a series of elementary local transformations resembling Reidemeister moves:

- Remove an empty loop: $\propto \Rightarrow ($
- Remove an empty bigon: $\between \Rightarrow)($
- Flip an empty triangle: $\forall \Rightarrow \rotatebox[origin=c]{180}{$\forall$}$

How many moves are required in the worst case, as a function of the number of self-intersection points? A proof of Steinitz's theorem[1] by Grünbaum[2] yields an $O(n^2)$ upper bound. A more recent algorithm of Feo and Provan[3] yields an upper bound of $O(nD)$ moves, where D is the diameter of the graph. On the other hand, the $\sqrt{n} \times (\sqrt{n}+1)$ "torus knot" curve provably requires at least $\binom{\sqrt{n}}{3} = \Omega(n^{3/2})$ moves. I conjecture that the lower bound is tight.

▶ Problem 3 (Tamal Dey). *Deciding triviality of cycles:* Let K be a finite simplicial complex linearly emedded in $\mathbb{R}^3$. Let C be any given 1-cycle in K. We are interested in detecting if C is *trivial* in the first homology group, that is, if there is a set of triangles in K whose boundaries when summed over $\mathbb{Z}_2$ give C. This problem can be solved in $O(M(n))$ time by first reducing the boundary matrix of K (triangle-edge matrix) to Echelon form and then reducing a column corresponding to C to see if it becomes empty column or not. Here $M(n)$ is the matrix multiplication time whose current best bound is $O(n^{2.37..})$.

▶ Conjecture 1. Let K be a finite simplicial complex linearly embedded in $\mathbb{R}^3$ with a total of n simplices. Given a 1-cycle C in K, one can detect if C is trivial in the first homology group (with $\mathbb{Z}_2$ coefficient) in $O(n^2)$ time.

If K is a 2-manifold, the detection can be performed in $O(n)$ time by a simple depth-first walk in K. If K is a 3-manifold with connected boundary, the algorithm in "An efficient computation of handle and tunnel loops via Reeb graphs [D.-Fan-Wang] *ACM Trans. Graphics (SIGGRAPH 2013)*, Vol. 32(4), 2013" can be modified to accomplish the task in $O(n^2)$ time. The question remains open for general simplicial complexes. Although, the conjecture is posed here for K embedded in $\mathbb{R}^3$ and for a 1-cycle C, it can be posed for a finite simplicial complex embedded linearly in $\mathbb{R}^d$ and a given p-cycle C in it.

▶ Problem 4 (Nina Amenta). *A problem of unique polyhedron determination:* Let P be a simplicial (triangulated) three-dimensional polyhedron, not necessarily convex. Given the combinatorial structure of P, that is, the graph of its 1-skeleton, and the dihedral angle at every edge. Assume the dihedrals are all bounded away from 0, although they could be positive (convex) or negative (concave). Does this uniquely determine the vertex positions (up to rotation, translation, scale)? (Mazzeo and Montcouquiol, 2011, Journal of Differential Geometry, proved that uniqueness holds for convex polyhedra; highly nontrivial proof.)

▶ Problem 5 (Michael Gene Dobbins). *Realizing order types by k-gons:* We say an arrangement of convex bodies is orientable when the bodies do not pair-wise cross (each pair of bodies has exactly 2 common supporting tangents) and among every three bodies, each body appears exactly once on the boundary of their convex hull. We define the order type of an orientable arrangement as the orientation of each triple of bodies: (+) if the bodies appear in counter-clockwise order around the boundary of their convex hull, and (−) if they appear in clockwise order.

For a fixed integer k, how complicated can the set of arrangements of k-gons of a fixed order type be?

[1] Every 3-connected planar graph is the 1-skeleton of a 3-polytope, and vice versa.

[2] Branko Grünbaum. *Convex Polytopes.* John Wiley & Sons, 1967.

[3] Thomas A. Feo and J. Scott Provan. Delta-wye transformations and the efficient reduction of two-terminal planar graphs. *Operations Research* 41(3):572–582, 1993.

With Andreas Holmsen and Alfredo Hubard, we were able to show that the k-gon realization space of an arrangement can have the homotopy type of any primary semialgebraic set, but the arrangement used for this construction was not orientable. Orientable arrangements are a natural class of arrangements to consider, since the orientations on triples in such an arrangement satisfy the chirotope axioms, and as such are more closely related to configurations of points. We conjecture that universality also holds for orientable arrangements. That is, we conjecture that the set of arrangements of k-gons of a fixed order type modulo projectivities can have the homotopy type of any primary semialgebraic set.

▶ Problem 6 (Joe Mitchell). *Two problems:* (a) Given n points in $\mathbb{R}^3$ in general position, is it always the case that there exists a triangulation (tetrahedralization) of S whose dual graph is Hamiltonian? (The dual graph has a node for each tetrahedron, and an edge between facet-sharing tetrahedra. We look for a Hamiltonian path.) In $\mathbb{R}^2$ it is always the case that a Hamiltonian triangulation exists. In $\mathbb{R}^3$ it suffices to consider points in convex position (after which, if a Hamiltonian triangulation is found, the interior points can be inserted, one by one, and the corresponding tetrahedra repartitioned to maintain Hamiltonicity).

(b) Given a unit-radius ball ("planet") in $\mathbb{R}^3$, find a minimum-length set X (path, cycle, or tree), outside the ball, such that X does not penetrate the interior of the ball and all of the surface of the ball is illuminated by X. The shortest known path (see SoCG video paper Timothy M. Chan, Alexander Golynski, Alejandro López-Ortiz, Claude-Guy Quimper, "The asteroid surveying problem and other puzzles". SoCG 2003:372-373) consists of a union of two segments and a connecting spiral curve; the shortest known cycle is the "baseball curve" consisting of 4 semicircles on the surface of the bounding cube; is the shortest tree any different from the shortest path?

▶ Problem 7 (Michael Kerber). *A problem of well centeredness:* A d-simplex σ in $\mathbb{R}^d$ is *well-centered* if the circumsphere of σ is inside $CH(\sigma)$. Is there a point set P of n points in $\mathbb{R}^d$ such that the Delaunay diagram of P has at least $c \cdot n^{\lceil d/2 \rceil}$ well-centered d-simplices? What if $d = 3$?

(Related to Pitteway triangulations.)

▶ Problem 8 (Joe Mitchell). *The guarding game:* In 2014 I posed the "guarding game": For a given set S of n points in the plane, player 1 (the "guarder") is to pick a subset, G, of S, of size $k = |G|$, at which he places guards; separately, without seeing what play 1 does, play 2 (the "polygonalizer") is to give a simple polygonalization, P, of S (the set S is the vertex set of P). The guarder wins if G guards P; otherwise, the polygonalizer wins. What is a reasonable value for k (as a function of n, or possibly of the number, i, of points of S interior to $CH(S)$) to make the game close to "fair"? What is the best strategy for each player?

Participants

- Peyman Afshani
Aarhus University, DK
- Annamaria Amenta
Univ. of California – Davis, US
- Franz Aurenhammer
TU Graz, AT
- Maike Buchin
Ruhr-Universität Bochum, DE
- Sergio Cabello
University of Ljubljana, SI
- Siu-Wing Cheng
HKUST – Kowloon, HK
- Otfried Cheong
KAIST – Daejeon, KR
- Jinhee Chun
Tohoku University – Sendai, JP
- Mark de Berg
TU Eindhoven, NL
- Vin de Silva
Pomona College – Claremont, US
- Olivier Devillers
INRIA Nancy – Grand Est, FR
- Tamal K. Dey
Ohio State University – Columbus, US
- Michael Gene Dobbins
Postech – Pohang, KR
- Anne Driemel
TU Eindhoven, NL
- Ioannis Z. Emiris
University of Athens, GR
- Jeff Erickson
Univ. of Illinois – Urbana, US
- Jie Gao
SUNY – Stony Brook, US
- Marc Glisse
INRIA Saclay – Île-de-France, FR
- Leonidas J. Guibas
Stanford University, US
- Michael Kerber
MPI für Informatik – Saarbrücken, DE
- David G. Kirkpatrick
University of British Columbia – Vancouver, CA
- Rolf Klein
Universität Bonn, DE
- Joseph S. B. Mitchell
SUNY – Stony Brook, US
- Wolfgang Mulzer
FU Berlin, DE
- Elizabeth Munch
University of Albany, US
- Yota Otachi
JAIST – Ishikawa, JP
- Jeff M. Phillips
University of Utah – Salt Lake City, US
- Natan Rubin
Ben Gurion University – Beer Sheva, IL
- Vera Sacristan
UPC – Barcelona, ES
- Maria Saumell
University of West Bohemia – Pilsen, CZ
- Ludmila Scharf
FU Berlin, DE
- André Schulz
Universität Münster, DE
- Raimund Seidel
Universität des Saarlandes, DE
- Donald Sheehy
University of Connecticut – Storrs, US
- Jonathan Shewchuk
University of California – Berkeley, US
- Anastasios Sidiropoulos
Ohio State University – Columbus, US
- Fabian Stehn
Universität Bayreuth, DE
- Monique Teillaud
INRIA Nancy – Grand Est, FR
- Csaba Toth
California State University – Northridge, US
- Carola Wenk
Tulane University, US
- Nicola Wolpert
University of Applied Sciences – Stuttgart, DE

Report from Dagstuhl Seminar 15112

Network Calculus

Edited by

Florin Ciucu[1], Markus Fidler[2], Jörg Liebeherr[3], and Jens Schmitt[4]

1 **University of Warwick, GB,** `florin@dcs.warwick.ac.uk`
2 **Leibniz Universität Hannover, DE,** `markus.fidler@ikt.uni-hannover.de`
3 **University of Toronto, CA,** `jorg@comm.utoronto.ca`
4 **University of Kaiserslautern, DE,** `jschmitt@cs.uni-kl.de`

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15112 "Network Calculus". At the seminar, about 30 invited researchers from academia and industry discussed the promises, approaches, and open challenges of the Network Calculus. This report gives a general overview of the presentations and outcomes of discussions of the seminar.

Seminar March 8–11, 2015 – http://www.dagstuhl.de/15112
1998 ACM Subject Classification C.2.1 Network Architecture and Design – Packet switching networks
Keywords and phrases Deterministic, Stochastic Network Calculus, Queueing Theory, Effective Bandwidth, Performance Evaluation
Digital Object Identifier 10.4230/DagRep.5.3.63
Edited in cooperation with Hao Wang, University of Kaiserslautern, DE

1 Executive Summary

Florin Ciucu
Markus Fidler
Jörg Liebeherr
Jens Schmitt

The network calculus has established as a versatile methodology for the queueing analysis of resource sharing based systems. Its prospect is that it can deal with problems that are fundamentally hard for alternative methodologies, based on the fact that it works with bounds rather than striving for exact solutions. The high modelling power of the network calculus has been transposed into several important applications for network engineering problems, traditionally in the Internet's Quality of Service proposals IntServ and DiffServ, and more recently in diverse environments such as wireless networks, sensor networks, switched Ethernets, Systems-on-Chip, as well as smart grids.

The goal of this Dagstuhl seminar was to gather the deterministic and stochastic network calculus community, to discuss recent research activities, to identify future research questions, and to strengthen cooperation. Topics of this Dagstuhl seminar were:

Wireless systems: for the analysis of wireless networks, a question of interest is how the stochastic properties of wireless channels impact delay and backlog performance. The usual statistical models for radio signals in a propagation environment do not lend themselves easily to a queueing model. Promising methods that were elaborated in the seminar are effective capacities and a recent network calculus of fading channels.

Lower bounds and tightness of bounds: based on the ability to solve some fundamentally hard queueing problems, the stochastic network calculus is regarded as a valuable alternative to the classical queueing theory. The derivation of performance bounds in the stochastic network calculus, e.g., for backlog, and delay, frequently exploits well known tail estimates, such as Chernoff bound and others. The tightness of these bounds and alternative more accurate models and techniques, such as Martingale bounds, were a topic of the seminar.

Network topology: a remarkable quality of the network calculus is that it includes a variety of systems that can be composed to arbitrary network topologies. Various analytical as well as numerical approaches have been explored to analyze different types of topologies, such as line topologies or feed-forward networks. The goal of this seminar was to identify relevant classes of topologies, their defining properties, and corresponding methods.

Parallel systems: the area of performance evaluation of parallel systems has recently become increasingly important due to the prevalence of modern parallel computational models. It is thus a great opportunity for the network calculus community to develop new models and methods which can enable a fundamental and broad understanding of the performance of parallel systems. At the seminar, recent approaches to parallel systems have been discussed.

Related methods: the network calculus has a number of rather unexplored and unexploited connections to related methods in the areas of competitive analysis, adversarial queueing theory, and robust queueing theory that may offer a significant potential for future research. At the seminar, researchers from related fields provided valuable new input to the network calculus community.

During the seminar, we discussed and (partly) answered the following questions:

What are the requirements on a wireless network calculus? Given the increasing importance of wireless communications, the seminar featured two sessions comprising seven presentations on wireless systems, where different approaches and their applications were discussed. Subsequently, a wireless roadmap discussion was centered around the following questions:

- How to model wireless channels and systems?
- What are the most relevant future systems and technologies?
- Which assumptions are needed, which can be safely made?
- What kind of results are needed, which theories can provide these?

With regard to the questions above, we highlight some of the main aspects that were elaborated on during the seminar. The methods that were presented include

- effective capacities,
- impairment models (duality with left-over service curves of scheduling),
- (min, x)-calculus for fading channels,
- capacity-delay-error boundaries,
- central limit theorem,
- Martingale bounds.

Providing different pros, a common basis of many of these methods was found to be due to the prevailing use of moment generating functions (Laplace transforms or Mellin transforms). Relevant systems that were discussed are cognitive radio, 3GPP, MIMO, spatial multiplexing, automatic repeat request, and medium access control. Some fundamental aspects of modelling wireless systems are the assumptions that are required today. Typical choices include

- service increments:
  - independent,
  - Markovian, Gilbert-Elliott channel,
- in-order delivery,
- error-free, instantaneous feedback channel,
- instantaneous retransmission of erroneous data,
- channel state information.

During the discussion, the need for transfer domains beyond Gilbert-Elliott models was raised. Also, the introduction of a notion of time into information-theoretic concepts, such as channel capacity, was discussed and finite-block length capacity results were brought up. Topics of further interest included spatial aspects of wireless networks, interference, and multi-hop networks in general. Regarding the solutions that can be obtained, a tradeoff between exactness and analytical closed forms became apparent. In particular, in system optimization analytical solutions were mentioned to be most useful to obtain derivatives of relevant performance measures. The discussion also touched upon some more general aspects such as qualitative vs. quantitative results, where many practical applications may not require exact results but can benefit from measurable rules of thumb.

What are most promising future research topics in the network calculus? This question was elaborated on in group work sessions, where the task was to identify an upcoming, relevant research topic where performance evaluation can be expected to make a key contribution. The discussion was guided by the following questions:

- What are the requirements for theory, which assumptions can be made?
- Which results would be needed from theory?
- How would a model/approach look like?
- What would be the best case outcome?
- Which body of theory could provide such results?
- What would be a good topic/method/approach for a PhD dissertation in this area?

Relevant topics in the network calculus were found to include cross-layer design, industrial communication, systems on chip, networks on chip, data center communication, and big data. A strategic orientation may also focus on new and unorthodox problems such as

- just-in-time manufacturing,
- renewable energy, smart grid,
- caching,
- financial engineering,
- road traffic,

where the intuitive concept of envelopes as used by the network calculus may be beneficial for many applications in industry. Methodological aspects that may pose relevant and interesting challenges were discussed in the areas of:

- re-entrant lines, particularly stability of such systems,
- max-min problems,
- derivative constraints, e.g., in modelling of batteries,
- network topologies, particularly non-feed forward networks.

Making network calculus happen: computational aspects, application modelling, tool support. Clearly, for network calculus to become a standard technique in performance modelling and analysis of networked and distributed systems it is crucial to arrive at computable solutions, demonstrate its strengths in diverse applications and provide software tools to support performance engineers in their daily tasks. As these different issues are interrelated on many levels two sessions with nine presentations were devoted to them. Among the different issues raised during these presentations and the corresponding discussions were the following:

- What are suitable novel application domains for network calculus? What are their requirements?
- How can network calculus computations be made more scalable? Where are fundamental limits for the network analysis? How do current software tools perform?
- What is the "killer" application for network calculus, and, in particular, for stochastic network calculus?
- How can network calculus' scope be extended to open up for new application domains?

Some (partial) answers to these important questions could be hinted at by the presentations and the subsequent discussions:

- Currently, some of the most promising application domains of (deterministic) network calculus were identified in industrial control, automotive and aerospace industries; also, interesting steps using (stochastic) network calculus in the modelling of smart energy grids were presented.
- The hardness of feedforward network analysis is by now understood, good heuristic approaches are on the way; however, cyclic dependencies and feedback systems are still open problems to some degree.
- The modelling of parallel systems using network calculus seems a promising building block to address novel attractive applications.
- Software tool support for network calculus, in particular for the stochastic version, is under construction and requires a community effort.

Looking over the fence: related methods. The research goals of network calculus and its methodologies, such as system performance evaluation, Markov chain analysis, or large deviations, intersect with those of other research communities. The objective of the session "Related Methods" was to create a forum where researchers from diverse research communities present their research approaches and discuss them with network calculus researchers. Thus, the session exposed the network calculus community to recent trends in system performance evaluation. Moreover, since speakers in this session had previously no or only limited exposure to network calculus, the session created an opportunity to disseminate the network calculus research agenda to other communities. The session was subtitled as "Looking over the fence", indicating an interest in learning new methodologies and the desire for cross- and interdisciplinary interactions. The session featured speakers from four countries (Canada, France, Israel, USA), from three disciplines (mathematics, theoretical computer science, operations research), presenting recent research on approaches on topics such as robust queueing theory, adversarial queueing theory, and competitive analysis.

This report provides an overview of the talks that were given during the seminar. Also, the seminar comprised a one minute madness session for introduction and for statements on the network calculus, a breakout session for group work on promising future research topics in the network calculus, as well as a podium discussions on wireless network calculus. The discussions, viewpoints, and results that were obtained are also summarized in the sequel.

We would like to thank all presenters, scribes, and participants for their contributions and lively discussions. Particular thanks go to the team of Schloss Dagstuhl for their excellent organization and support.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

3.1 Network Calculus for Parallel Processing

George Kesidis (The Pennsylvania State University, US)

We begin with an overview of classical Markovian results in fork-join queues and cloud-computing jargon. We then present preliminary results on the use of network calculus for parallel processing (fork join) systems such as MapReduce. We derive a probabilistic bound on delay through a single parallel processing stage. We also provide a numerical result using a publicly available dataset of a Facebook data-center that includes the total job arrival rate and workload statistics of the tasks of different types of MapReduce jobs at both the mapper and reducer stages. Finally, we discuss how to extend to tandem queues.

3.2 Wireless Network Calculus

Yuming Jiang (NTNU – Trondheim, NO)

In this talk, an overview of the difficulty, the key underlying issues and an overall picture of Wireless Network Calculus, i.e. extension/application of SNC to wireless networks, is first presented. This is followed by a brief introduction of our achieved research results in Wireless Network Calculus. The last part is devoted to the introduction of a fundamental problem in Wireless Network Calculus, which is end-to-end (e2e) QoS analysis of wireless networks where there is interference among neighbor hops. Some preliminary ideas to deal with this analysis are presented.

3.3 Energy Efficient Effective Capacity for 5G Networks

Eduard Jorswieck (TU Dresden, DE)

URL http://5glab.de/

In wireless 5G networks a paradigm change of services and applications to machine-to-machine low delay communications (tactile internet) requires to guarantee round trip times below 1–10 ms. On the other hand, the energy efficiency in 5G should be improved by a factor of 1000, too. Therefore, we propose a new performance metric which combines both conflicting objectives into the efficient effective capacity defined as the ratio of effective capacity to total consumed energy. The maximization of this metric leads to a fractional programming problem which can be solved efficiently by the Dinkelbach algorithm. The extension of the efficient effective capacity to the elements of multiuser networks, i.e., to the multiple access and broadcast channel is not available yet, because expressions for the effective capacity region are missing. In order to develop 5G networks with latency requirements/guarantees, we need to solve the following problems:

1. Derive the effective capacity region for multiple access channels.
2. Compute the effective capacity region for broadcast channels.
3. Derive the effective capacity for multihop (relaying) networks for different relaying protocols (amplify-and-forward, decode-and-forward, compress-and-forward, compute-and-forward, noisy network coding).

3.4 Effective Capacity – Through Physical and Data-Link Layers

Sami Akin (Leibniz Universität Hannover, DE)

Alongside the growth in social networks, mobile computing and pervasive communications, and the innovations in lower layer technologies, we see the need to re-visit network design strategies and develop better protocols. Can we design better higher layer strategies that inform, or are informed by, the underlying physical layer? With sufficient co-existence mechanisms, what novel cognitive radio network architectures are required? Hence, in this presentation, we discuss Effective Capacity from a physical layer perspective and investigate the effects of physical layer features on buffer performance in data-link layers by considering the cognitive radio framework as a working ground.

3.5 Performance of In-Network Processing for Visual Analysis in Wireless Sensor Networks

Hussein Al-Zubaidy (KTH Royal Institute of Technology, SE)

Joint work of Al-Zubaidy, Hussein; Gyorgy, Dan; Viktoria, Fodor
Main reference H. Al-Zubaidy, D. Gyorgy, F. Viktoria, "Performance of in-network processing for visual analysis in wireless sensor networks," in Proc. of the 14th IFIP Networking Conf. (Networking 2015), pp. 1–9, IEEE, 2015.
URL http://dx.doi.org/10.1109/IFIPNetworking.2015.7145292

Nodes in a sensor network are traditionally used for sensing and data forwarding. However, with the increase of their computational capability, they can be used for in-network data processing, leading to a potential increase of the quality of the networked applications as well as the network lifetime. Visual analysis in sensor networks is a prominent example where the processing power of the network nodes needs to be leveraged to meet the frame rate and the processing delay requirements of common visual analysis applications. The modelling of the end-to-end performance for such networks is, however, challenging, because in-network processing violates the flow conservation law, which is the basis for most queuing analysis. In this work we propose to solve this methodological challenge through appropriately scaling the arrival and the service processes, and we develop probabilistic performance bounds using stochastic network calculus. We use the developed model to determine the main performance bottlenecks of networked visual processing. Our numerical results show that an end-to-end delay of 2–3 frame length is obtained with violation probability in the order of 10^{-6}. Simulation shows that the obtained bounds overestimates the end-to-end delay by no more than 10%.

3.6 Capacity-Delay-Error Boundaries: A Composable Model of Sources and Systems

Nico Becker (Leibniz Universität Hannover, DE)

Joint work of Fidler, Markus; Lübben, Ralf; Becker, Nico
Main reference M. Fidler, R. Lubben, N. Becker, "Capacity-Delay-Error Boundaries: A Composable Model of Sources and Systems," IEEE Trans. onWireless Communications, 14(3):1280–1294, 2014.
URL http://dx.doi.org/10.1109/TWC.2014.2365782

It is presented a notion of capacity-delay-error (CDE) boundaries as a performance model of networked sources and systems. It is shown that the model has the property of additivity, which enables composing CDE boundaries obtained for sources and systems as if in isolation. Results for essential sources, channels and for the composition of sources and channels coders are presented.

3.7 Service-Martingales: Theory and Applications to the Analysis of Random Access Protocols

Felix Poloczek (TU Berlin, DE)

Joint work of Poloczek, Felix; Ciucu, Florin
Main reference F. Poloczek, F. Ciucu, "Service-martingales: theory and applications to the delay analysis of random access protocols," in Proc. of the 2015 IEEE Conf. on Computer Communications (INFOCOM'15), pp. 945–953, IEEE, 2015.
URL http://dx.doi.org/10.1109/INFOCOM.2015.7218466

We propose a martingale extension of effective capacity, a concept which has been instrumental in the teletraffic theory to model the link-layer wireless channel and to analyze QoS metrics. Together with a recently developed concept of an arrival-martingale the proposed *service-martingale* concept enables the queuing analysis of a bursty source sharing a MAC channel. In particular, we derive the first rigorous stochastic delay bounds for a Markovian source sharing either an ALOHA or CSMA/CA channel. By leveraging the powerful martingale methodology, the obtained bounds are remarkably tight.

3.8 Queuing Analysis of Wireless Systems: A Waste of Time?

James Gross (KTH Royal Institute of Technology, SE)

For some time now, there is a significant research activity with respect to queuing analysis of wireless systems based on effective capacity. These contributions follow a certain pattern: Identify what is hot in information theory and provide the corresponding queuing analysis. However, such contributions are limited by the additional insight they provide (in comparison to the original publication), while on the other hand the models are usually too theoretic to have practical value. In this talk I mainly illustrate these circumstances based on my own work, and intend to provoke discussions around the future value of queuing-related analysis of wireless systems. A few possible ways forward are finally presented, too.

3.9 SLA Calculus – Modelling Software Systems with Network Calculus

Peter Buchholz (TU Dortmund, DE)

Quantitative properties of modern software systems are often defined as part of a service level agreement (SLA) that fixes the maximal load and the maximal delay. Evaluation of the software system in order to validate the SLA is a challenging task since the system is to a large extend unknown and unpredictable. Thus, performance analysis has to be based on the SLAs without additional information about the basic system. The talk presents a new approach to analyze software architectures based on the ideas available in network-order real time calculus. In this way, bounds for departure processes are computed from available bounds for the arrival and delay processes. With the technique systems of composed services can be easily analyzed resulting in SLAs for the composed service. It is shown, how the solutions can be used to help a user and a provider to analyze and determine SLAs. Furthermore, open questions and limitations of the proposed approach are outlined.

3.10 Modelling Avionics Communicating Systems: Successes, Failures, Challenges

Marc Boyer (ONERA – Toulouse, FR)

This talk gave some perspectives on “the application modelling side, what is required from NC, what is still missing, what are success and failure stories”. The talk presented how the modelling of AFDX has been done in an accurate way, whereas the one of SpaceWire has not. Thereafter, seven challenges on modelling are listed.

3.11 Industrial Application of Network Calculus

Kai-Steffen Jens Hielscher (Universität Erlangen-Nürnberg, DE)

Joint work of Herpel, Thomas; Hielscher, Kai-Steffen; Klehmet, Ulrich; German, Reinhard
Main reference T. Herpel, K.-S. Hielscher, U. Klehmet, R. German, “Stochastic and deterministic performance evaluation of automotive CAN communication,” Computer Networks, 53(8):1171–1185, 2009.
URL http://dx.doi.org/10.1016/j.comnet.2009.02.008

In this talk we present the application of deterministic Network Calculus for two real-world examples: Communication of embedded controllers in automotive networks in cooperation with Audi AG and industrial Ethernet communication for industry automation in cooperation with Siemens AG. In the automotive example, the industry partner provided the topology and information about periodic CAN and FlexRay messages. The goal was to decide on which of the different busses inside a car interoperating electronic control units (ECUs) should be placed to avoid the violation of the hard real-time bounds. To achieve this, the CAN media access method had to be modelled in Network Calculus. Since the busses are interconnected by a central gateway, the service of this gateway also has to be modeled. Besides the scheduling strategy, this involved considering the aggregation of numerous interfering flows.

Industrial automation today mainly uses variants of industrial Ethernet like Profinet RT. Since these technologies do not provide guarantees like traditional field busses, our industry partner Siemens uses Network Calculus to calculate bounds for real-time traffic. For this purpose, they are integrating a Network Calculus Engine into their existing network planning tool. The tool already contains topology information and necessary information to generate arrival curves for scheduled flows. Other flows generated by user programs can be integrated by semi-automatic static code analysis. Since the end users often integrate hardware like web cams and HMI terminals into the network that generates non-real-time traffic, traffic profiles for these applications have been defined. To ensure that the limits provided in the profiles are not exceeded, traffic shaping has to be introduced into the network for the non-real-time flows.

References

1 T. Herpel, K.-S. Hielscher, U. Klehmet, and R. German. Stochastic and deterministic performance evaluation of automotive CAN communication, in *Computer Networks*, vol. 53, no. 8, pp. 1171–1185, 2009, Performance modelling of Computer Networks: Special Issue in Memory of Dr. Gunter Bolch.

2 S. Kerschbaum, K.-S. Hielscher, U. Klehmet, and R. German, A framework for establishing performance guarantees in industrial automation networks, in *Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance*, ser. Lecture Notes in Computer Science, K. Fischbach and U. Krieger, Eds., Springer International Publishing, 2014, vol. 8376, pp. 177–191.

3.12 On the Scalability of Real Time Calculus

Kai Lampka (Uppsala University, SE)

Joint work of Lampka, Kai; Perathoner, Simon; Suppiger, Urban; Thiele, Lothar
Main reference U. Suppiger, S. Perathoner, K. Lampka, L. Thiele, “A simple approximation method for reducing the complexity of modular performance analysis,” Technical Report 329, ETH Zurich, August 2010.
Main reference http://www.tik.ee.ethz.ch/db/public/tik/?db=publications&form=report_single_publication&publication_id=3494

With Real Time Calculus and the related tool-support [8], it can be observed that the computation of the commonly used piece-wise linear pseudo-periodic functions, may require significant demands of computation and memory resources. The resulting overheads might render system analysis inefficient, if not infeasible, or often enforce simplifications, respectively overapproximations in the modelling. Simplifying overapproximations of signal frequencies or processing patterns, yield analysable models and guarantees conservativeness of results. However, it results in a non-tight bounding of performance metrics and ultimately yields potentially over-provisioned system designs. This shortcoming is the starting point for precise prefixing of bounding functions, as it only exploits overapproximations on the unneeded parts of functions. In order to achieve this the presentation presents the following innovations.

- The presentation introduces the concept of curve prefixing. This allows one to present curves precisely only on the interval $[0, c]$. For the range $(c, +\infty)$ the concept uses simplifying overapproximations which makes periodic tail descriptions of curves obsolete.

- The presentation formally establishes the framework for computing backlog and delay bound which are as tight as if one would have used the original curve representation. It thereby lifts the presenter's previous work in this direction from the level of an approximation method to the level of a precise analysis technique.
- The presentation contains an industrial, real-time constraint communication system. The system contains over 200 devices and integrates different real-time applications in a single (non-partitioned) architecture.

The concept of curve prefixing and tail overapproximations makes a clear distinction to today's implementations of Network or Real-time Calculus, e.g., as provided by the Matlab-based MPA-toolbox [5]. There, curve prolongation is the default behaviour, at each component the least common multiple of the periods of two input curves gives the period of the resulting output curve. The proposed approach therefore clearly increases the scalability of RTC-based system analysis as demonstrated by the industrial case study. But most importantly, it works on top of the existing tools and thereby avoids re-implementation of RTC.

References

1 J.-Y. L. Boudec and P. Thiran. Network Calculus: a theory of deterministic queuing systems for the Internet, volume 2050 of *LNCS*. Springer, 2001.

2 A. Bouillard and E. Thierry. An algorithmic toolbox for network calculus, in *Journal of Discrete Event Dynamic Systems (JDEDS)*, 18(1):3–49, 2008.

3 R. L. Cruz. A calculus for network delay. part i: Network elements in isolation and part ii: Network analysis, in *IEEE Transactions on Information Theory*, 37(1):114–141, January 1991.

4 R. Henia, A. Hamann, M. Jersak, R. Racu, K. Richter, and R. Ernst. System level performance analysis – the SymTA/S approach, chapter 2, pages 29–72, The Institution of Electrical Engineers, London, United Kingdom, 2006.

5 Modular performance analysis framework. http://www.mpa.ethz.ch.

6 U. Suppiger, S. Perathoner, K. Lampka, and L. Thiele. Modular performance analysis of large-scale distributed embedded systems: an industrial case study, Technical Report 330, ETH Zurich, November 2010.

7 U. Suppiger, S. Perathoner, K. Lampka, and L. Thiele. A simple approximation method for reducing the complexity of modular performance analysis, Technical Report 329, ETH Zurich, August 2010.

8 E. Wandeler, L. Thiele, M. Verhoef, and P. Lieverse. System architecture evaluation using modular performance analysis – a case study, in *International Journal on Software Tools for Technology Transfer*, 8(6):649–667, October 2006.

3.13 Network Calculus Tool Support – Expectations and Reality

Steffen Bondorf (University of Kaiserslautern, DE)

The first part of this talk will be covering the Disco Deterministic Network Calculator (DiscoDNC), an open-source network calculus tool [1].

Steffen Bondorf has been working with the network calculus tool support offered by the DISCO group for some time [2] before he eventually took over the role as its maintainer.

Since then, he has put effort into improving the tool in different aspects [3] – one of which is lowering the barrier to start working with deterministic network calculus. For that, the code has been restructured, the API reworked, functional tests have been created and their computations have been documented in detail.

This work led to several inquiries from researchers seeking to make use of network calculus results in order to evaluate their work. Unfortunately, there is a gap between the expectations that those researcher had regarding tool support and the reality at hand. Most notably, the preceding modelling step required to apply network calculus emerged as the single most problematic hurdle on the way towards deriving delay and backlog bounds. The DiscoDNC, however, strictly depends on the network calculus model, i.e., service curves and arrival curves need to be given in order to analyze a network.

In his talk, Steffen shares his experiences from being approached by academics making their first steps in the area of network calculus. He will depict common misconceptions along the lines of an example, showing that the effort to analyze a "simple" network with roughly 200 nodes can result in actually analyzing a so-called server graph connecting 1140 queues (servers) connected by nearly 7000 links. This observation motivates Steffen's work on improving the efficiency of network calculus analyses.

The second part of this talk will depict several enhancements to the computational efficiency of network calculus analyses. These improvements can be divided into two groups:

- Technical solutions allowing the DiscoDNC to derive bounds faster and
- Conceptual improvements in network calculus itself.

The former part covers the reuse intermediate results and the potential to parallelize the execution of a network analysis – both possible thanks to the modularity of (algebraic) deterministic network calculus.

The latter part will conclude the talk by providing some insight into an upcoming result [4] enabling to significantly reduce the analysis effort in sink trees with token-bucket arrival curves and rate-latency service curves.

References

1 The Disco Deterministic Network Calculator. http://disco.cs.uni-kl.de/index.php/projects/disco-dnc

2 S. Bondorf and J. Schmitt. Statistical Response Time Bounds in Randomly Deployed Wireless Sensor Networks, in *Proceedings of the 35th IEEE Conference on Local Computer Networks (LCN 2010).*

3 S. Bondorf and J. Schmitt. The DiscoDNC v2 – A Comprehensive Tool for Deterministic Network Calculus, in *Proceedings of the 8th International Conference on Performance Evaluation Methodologies and Tools (ValueTools 2014).*

4 S. Bondorf and J. Schmitt. Boosting Sensor Network Calculus by Thoroughly Bounding Cross-Traffic, in *Proceedings of the 34th IEEE International Conference on Computer Communications (INFOCOM 2015).*

3.14 Exact Delays in Networks

Anne Bouillard (ENS/INRIA, FR)

In this talk, we present a method based on linear programming to compute exact worst-case delay bounds under network calculus assumptions. We assume that the network is feed-forward; that the arrival/service curves are piecewise linear concave/convex and that the service policy is FIFO. The proposed method encodes every NC constraint into linear constraints, possibly with boolean variables. Then the solution of the LP is the exact worst-case delay. This algorithm is compared against existing method; derived into two simpler LPs that respectively compute good approximations of the upper bound and lower bound of the exact worst-case delay.

3.15 Optimal Joint Path Computation and Rate Allocation for Real-time Traffic

Giovanni Stea (University of Pisa, IT)

Computing network paths under worst-case delay constraints has been the subject of abundant literature in the past two decades. Assuming Weighted Fair Queueing scheduling at the nodes, this translates to computing paths and reserving rates at each link. The problem is NP-hard in general, even for a single path; hence polynomial-time heuristics have been proposed in the past, that either assume equal rates at each node, or compute the path heuristically and then allocate the rates optimally on the given path. In this paper we show that the above heuristics, albeit finding optimal solutions quite often, can lead to failing of paths at very low loads, and that this could be avoided by solving the problem, i.e., path computation and rate allocation, *jointly* at *optimality.* This is possible by modelling the problem as a mixed-integer second-order cone program and solving it optimally in split-second times for relatively large networks on commodity hardware; this approach can also be easily turned into a heuristic one, trading a negligible increase in blocking probability for one order of magnitude of computation time. Extensive simulations show that these methods are feasible in today's ISPs networks and they significantly outperform the existing schemes in terms of blocking probability.

3.16 How Can Network Calculus Help Smart Grids?

Yashar Ghiassi (University of Waterloo, CA)

This work is motivated by the challenges that arise when integrating large scale renewable energy integration. The significant fluctuations injected to the grid by renewable energy sources must be captured by storage systems. The role of storage in smart grids resembles the role of buffers and shapers in computer networks. We use this analogy to employ the buffer-overflow bounds from network calculus to size storage systems for given maximum loss of power and rate of power probabilities. This framework applies to a large range of applications in smart grids given that storage is an integral element in smart grids.

3.17 Computable Bounds in Fork-Join Queueing Systems

Amr Rizk (University of Massachusetts – Amherst, US)

Joint work of Rizk, Amr; Poloczek, Felix ; Ciucu, Florin
Main reference A. Rizk, F. Poloczek, F. Ciucu, "Computable Bounds in Fork-Join Queueing Systems," SIGMETRICS Perform. Eval. Rev., 43(1):335–346, ACM, 2015.
URL http://dx.doi.org/10.1145/2745844.2745859

A Fork-Join (FJ) queueing system is characterized by an upstream fork station that splits incoming jobs into N tasks to be further processed by N parallel servers, each with its own queue; the response time of one job is determined, at a downstream join station, by the maximum of the corresponding tasks's response times. FJ queueing systems help modelling multi-service systems subject to synchronization constraints. One prominent example are MapReduce clusters. In this work we provide first computable stochastic bounds on the waiting and response time distributions in FJ systems for renewal and non-renewal arrivals. Further, we consider blocking and non-blocking server behavior and prove that delays scale as $\mathcal{O}(\log N)$ in the non-blocking case, a law which is known for first moments under renewal input only. We show simulation results indicating that our bounds are tight, especially at high utilizations.

3.18 Window Flow Control in Network Calculus

Michael Beck (University of Kaiserslautern, DE)

This talk is concerned with the long-standing problem of feedback in Network Calculus (NC), in particular Stochastic Network Calculus (SNC). While there are plenty and elegant results on the deterministic side of NC, corresponding theorems are missing in SNC. This in turn limits the areas where SNC could be applied. In this talk – presenting preliminary work – the feedback-inequality in its original form and its connection to a Window Flow Controller is given. An overview follows, presenting the generalizations on the feedback-inequality. This is concluded with the solution to the feedback-inequality for the continuous-time and bivariate case. While this is a necessary step, it is not sufficient for a full analysis of a WFC. It provides, however, some insights, especially on a paradox behavior concerning dynamic window sizes, which perform worse compared to their static window counterparts. At last it is shown that under (very) strict assumptions an analysis of the stochastic WFC is possible.

3.19 Scaling Laws in the Network Calculus Bounds vs. Exact results

Almut Burchard (University of Toronto, CA)

In this talk, I described how exact results can be recovered from performance bounds in the stochastic network calculus in certain important limits. For example, it is well-understood that the output bound agrees, in the long-time limit, with the arrival rate; similarly, the exact fail decay of the backlog can be recovered from the delay bound. A more delicate question is the growth of end-to-end delays with the path length. For heavy-tailed and self-similar processes, such delays grow with a power-law, but the exact power is not known. I illustrated the importance of simple scaling laws for the evaluation of simulations.

3.20 Routing and Scheduling for Bursty Adversarial Traffic – Adversarial Queuing Theory

Adi Rosén (CNRS / University Paris-Diderot, FR)

In this talk we mainly consider the setting of Adversarial Queuing Theory.

The main part of the talk gives a simple, deterministic, local-control routing and scheduling protocol that applies to any network topology. This protocol guarantees that, for any input traffic for which stability is possible, stability is indeed achieved, and moreover the buffers at the nodes are polynomially-bounded as well as each packet has polynomially-bounded delivery time. This part of the talk is based on the paper [1].

This main part of the talk is complemented by a short (partial) survey of results in Adversarial Queueing Theory, as well as results on the achievable throughput in networks with fixed, bounded buffers. The latter results, compared to the results in Adversarial Queuing Theory, suggest that the question of stability and the question of throughput under bounded buffers are different questions with answers that do not relate to each other.

References

1 W. Aiello, E. Kushilevitz, R. Ostrovsky, and A. Rosén. Adaptive packet routing for bursty adversarial traffic, in *JCSS*, vol. 60, no. 3, pp. 482–509, 2000.

3.21 Managing Queues with Bounded Buffers: Micro-decisions from a Competitive Lens

Gabriel Scalosub (Ben Gurion University – Beer Sheva, IL)

Network Calculus has traditionally assumed flow conservation, and that no traffic is lost while traversing the network. In consequence, it has primarily focused on understanding the performance of systmes in terms of delay, and provisioning of capacity. Nevertheless, packet loss is a feature of common networks, most predominantly the Internet, where packets are dropped due to buffer overflows, congestion control mechanisms, and and ever growing demand for more bandwidth which is not always available. In this talk we present models and algorithms for dealing with such packet loss, focusing primarily on buffer-management mechanisms. These results are cast within a competitive framework, which subscribes to other related models, such as AQT. We present both classic results in this framework, as well as some more recent results, and emphasize the characteristics of performing buffer-management, which are sometimes counter-intuitive, and sometimes lead to surprising results. Among other things, this talk may also serve as a teaser for Network Calculus to try and incorporate packet loss (and working with bounded buffers) into its framework.

3.22 Robust Queueing Theory

Nataly Youssef (MIT – Cambridge, US)

Joint work of Bandi, Chaithanya; Bertsimas, Dimitris; Youssef, Nataly
Main reference C. Bandi, D. Bertsimas, N. Youssef, "Robust Queueing Theory," Operations Research, 63(3):676–700, 2015.
URL http://dx.doi.org/10.1287/opre.2015.1367

We propose an alternative approach for studying queues based on robust optimization. We model the uncertainty in the arrivals and services via polyhedral uncertainty sets which are inspired from the limit laws of probability. Using the generalized central limit theorem, this framework allows to model heavy-tailed behavior characterized by bursts of rapidly occurring arrivals and long service times. We take a worst-case approach and obtain closed form upper bounds on the transient and steady-state system time in multi-server queues and feedforward networks. These expressions provide qualitative insights which mirror the conclusions obtained in the probabilistic setting for light-tailed arrivals and services and generalize them to the case of heavy-tailed behavior. We also develop a calculus for analyzing a steady-state network of queues based on the following key principle: (a) the departure from a queue, (b) the superposition, and (c) the thinning of arrival processes have the same uncertainty set representation as the original arrival processes. The proposed approach (a) yields results with error percentages in single digits relative to simulation, and (b) is to a large extent insensitive to the number of servers per queue, network size, degree of feedback, traffic intensity, and somewhat sensitive to the degree of diversity of external arrival distributions in the network.

4 Working Groups

4.1 Working group A

Steffen Bondorf

The working group started with discussing recent developments combining network calculus (NC) with optimization. In deterministic network calculus (DNC), work in this area started as early as 2008 when the basic problem of existing algebraic tandem analyses was identified.

It was suggested to derive an optimization problem from the network calculus "constraints" to solve it. Since then, the optimization-based analysis has been advanced to ultimate tightness, i.e., the best results possible with the given NC constraints, and has been extended to encompass the entire network instead of a tandem of servers only. In the stochastic network calculus (SNC) branch, work recently suggested to make use of its modelling capabilities in robust optimization. Thus creating a robust queueing theory.

These developments were caused by deficiencies in network calculus that are not easy to overcome. The discussion identified the following major issues:

- Lack of decision variables: Being restricted to the analysis of systems, NC relies on the provision of an exact model. It lacks the capability to directly support system engineering by finding assignments for open parameters such that a given requirement is fulfilled. Complementary methodologies as add-ons can help NC to increase its applicability in this area.
- Bounds instead of exact results: Network calculus itself is concerned with bounding a performance indicator instead of providing an exact result. Quality of bounds is a problem of both branches, DNC and SNC.
- Computational effort: Moreover, the computational effort involved in deriving bounds can be very high. For example, in the DNC analysis of tandems of FIFO multiplexing servers can already be very involved, yet, it is not ultimately tight. Although NC only derives bounds, oftentimes an additional tradeoff is required to derive results at all – especially in the analysis of reasonably sized networks.

The working group then turned to the aspect thought to be the common cause of the above problems: The model used for network calculus. From the beginning, i.e., Cruz' first papers, its simplicity was considered as defining NC's beauty. NC does not take many assumptions into account whereas classic performance tend to have too many to keep track of all of them properly. However, in order to overcome the problems identified in the discussion, the model may be considered simplistic. It is, e.g., even simpler than visual models for simulation as used by OMNeT++ or others. Summarizing, this kind of beauty defines the limitations of network calculus as well. The constrained expressiveness hampers the ambition to derive better bounds while simultaneously not allowing for fast and easy derivations either.

Given that the NC model's expressiveness currently seems to restrict leaps forward, the question about creating a potentially better model appeared. The group members asked themselves if it was possible to take such a disruptive step that incorporates the knowledge and experience the community generated over the past years. I.e., can we redesign the calculus such that its main deficiencies will be gone for good?

4.2 Working group B

Yashar Ghiassi

We started the meeting by discussing the application domains of network calculus. We classified the applications to two groups: emerging applications and traditional ones. Examples of emerging applications are vehicular transportation, energy systems (battery and EVs), financial engineering, and inventory control and manufacturing systems. Examples of traditional applications are communication networks and computation networks (e.g., cloud, embedded).

In the second half of our meeting we tried to discuss possible interesting problems (in each of the applications listed above) for which network calculus is helpful. The first interesting set of problems are facility location and dynamic topology; e.g., how do we optimally size and locate storage systems in smart grids? As another important and open problem we discussed feedback networks problems and the possibility that network calculus extends to that area of research. Routing algorithms was the third possible research direction that we discussed. Finally, we discussed a series of control related problems: state-dependent scheduling, traffic lights/signals, avoiding underflow (finance), and ramp limitations (electricity).

4.3 Working group C

Amr Rizk

The main focus of the discussion within this working group was on identifying requirements for advancing the Network Calculus (NC), as well as, main technical problems that are not solved (yet!) in the NC framework. We identified two pillars that would help the advancement of Network Calculus in the sense of increasing the user, as well as, the researcher community, i.e., (i) bringing NC to standardization and (ii) teaching NC at a graduate level. One success story of a related performance evaluation research topic that made a key contribution through the transition to (IETF) standardization is fair scheduling. Hence, it is of utmost importance to visualize the impact of the NC framework with implementations and case studies of actual deployment. A particular strength of NC lies in providing fundamental characterizations (limits) on basic networking elements that can be compiled into complex communication scenarios. We believe that a collection of such results with appropriate deployment examples would be very instructive for potential adoption. However, there are still basic elements and protocols that do not lend themselves to the (stochastic) Network Calculus, such as feedback and lossy systems. The conclusion of the discussion was that there are still many open challenges/problems to be solved within the Network Calculus framework.

5 Seminar Programme

Monday	
09:00-09:30	Welcome and general introduction
09:30-10:30	One minute madness: introduction of participants
11:00-12:00	Seed talk: George Kesidis
14:00-15:30	Wireless network calculus Yuming Jiang Eduard Jorswieck Sami Akin Hussein Al-Zubaidy
16:00-17:00	Wireless network calculus Nico Becker Felix Poloczek James Gross
17:00-17:45	Wireless network calculus: roadmap discussion
evening	Network calculus pub quiz

Tuesday	
09:00-10:00	Seed talk: Peter Buchholz
10:00-12:00	Group work: future network calculus topics
14:00-15:30	Making network calculus happen: computational aspects application modelling and tool support (CAT) Marc Boyer Kai-Steffen Hielscher Kai Lampka Steffen Bondorf
16:00-17:45	Making network calculus happen: computational aspects application modelling and tool support (CAT) Anne Bouillard Giovanni Stea Yashar Ghiassi-Farrokhfal Amr Rizk Michael Beck

Wednesday	
09:00-10:30	Looking over the fence: related methods Almut Burchard Adi Rosen Gabriel Scalosub Nataly Youssef
11:00-12:00	Feedback from group work
12:00-12:15	Seminar resume and farewell

Participants

- Sami Akin
Leibniz Univ. Hannover, DE
- Hussein Al-Zubaidy
KTH Royal Institute of Technology, SE
- Michael Beck
University of Kaiserslautern, DE
- Nico Becker
Leibniz Univ. Hannover, DE
- Daniel Berger
University of Kaiserslautern, DE
- Steffen Bondorf
University of Kaiserslautern, DE
- Anne Bouillard
ENS – Paris, FR
- Marc Boyer
ONERA – Toulouse Research Center, FR
- Peter Buchholz
TU Dortmund, DE
- Almut Burchard
University of Toronto, CA
- Florin Ciucu
University of Warwick, GB
- Markus Fidler
Leibniz Univ. Hannover, DE
- Reinhard German
Univ. Erlangen-Nürnberg, DE
- Fabien Geyer
TU München, DE
- Yashar Ghiassi-Farrokhfal
University of Waterloo, CA
- James Gross
KTH Royal Institute of Technology, SE
- Kai-Steffen Jens Hielscher
Univ. Erlangen-Nürnberg, DE
- Yuming Jiang
NTNU – Trondheim, NO
- Eduard Jorswieck
TU Dresden, DE
- George Kesidis
Pennsylvania State University – University Park, US
- Kai Lampka
Uppsala University, SE
- Jörg Liebeherr
University of Toronto, CA
- Krishna S. Pandit
TU Darmstadt, DE
- Felix Poloczek
TU Berlin, DE
- Amr Rizk
University of Massachusetts – Amherst, US
- Adi Rosén
CNRS / Univ. Paris-Diderot, FR
- Gabriel Scalosub
Ben Gurion University – Beer Sheva, IL
- Jens Schmitt
University of Kaiserslautern, DE
- Giovanni Stea
University of Pisa, IT
- Hao Wang
University of Kaiserslautern, DE
- Nataly Youssef
MIT – Cambridge, US

Report from Dagstuhl Seminar 15121

Mixed Criticality on Multicore/Manycore Platforms

Edited by

Sanjoy K. Baruah[1], Liliana Cucu-Grosjean[2], Robert I. Davis[3,2], and Claire Maiza[4]

1 **University of North Carolina at Chapel Hill, US,** baruah@cs.unc.edu
2 **INRIA Roquencourt – Le Chesnay, FR,** liliana.cucu@inria.fr
3 **University of York, GB,** rob.davis@york.ac.uk
4 **VERIMAG – Gières, FR,** claire.maiza@imag.fr

Abstract

This report provides an overview of the discussions, the program and the outcomes of the first Dagstuhl Seminar on Mixed Criticality on Multicore/Manycore Platforms. The seminar brought together researchers working on challenges related to executing mixed criticality real-time applications on multicore and manycore architectures with the main purpose of promoting a closer interaction between the sub-communities involved in real-time scheduling, real-time operating systems / runtime environments, and timing analysis as well as interaction with specialists in hardware architectures.

Seminar March 16–20, 2015 – http://www.dagstuhl.de/15121
1998 ACM Subject Classification C.3 Special-purpose and Application-based Systems – Real-time and embedded systems
Keywords and phrases Mixed-Criticality, Real-time systems, Multicore/Manycore Platforms, fixed priority; probabilistic scheduling, varying-speed processors, model combination
Digital Object Identifier 10.4230/DagRep.5.3.84
Edited in cooperation with Adriana Gogonel

1 Executive Summary

Liliana Cucu-Grosjean
Robert I. Davis
Claire Maiza
Sanjoy K. Baruah

Real-time systems are characterised not only by the need for functional correctness, but also the need for timing correctness. Today, real-time embedded systems are found in many diverse application areas including; automotive electronics, avionics, and space systems. In these areas, technological progress is resulting in rapid increases in both software complexity and processing demands. To address the demand for increased processor performance, silicon vendors no longer concentrate on increasing processor clock speeds, as this approach has led to problems with high power consumption and excessive heat dissipation. Instead, technological development has shifted to multicore processors, with multiple CPUs integrated onto a single chip. The broad technology trend is towards much larger numbers of cores, referred to as manycore, requiring network-on-chip rather than bus interconnects.

Mixed Criticality on Multicore/Manycore Platforms, *Dagstuhl Reports*, Vol. 5, Issue 3, pp. 84–142
Editors: Sanjoy K. Baruah, Liliana Cucu-Grosjean, Robert I. Davis, and Claire Maiza
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Requirements on Size Weight and Power consumption, as well as unremitting cost pressures, are pushing developments in avionics and automotive electronics towards the adoption of powerful embedded multicore processors, with a longer term vision of migrating to manycore. With the adoption of such technology comes the opportunity to combine different applications on the same platform, potentially dramatically reducing assembly and production costs, while also improving reliability through a reduction in harnessing. Different applications may have different criticality levels (e.g. safety-critical, mission-critical, non-critical) designating the level of assurance needed against failure. For example, in automotive electronics, cruise control is a low criticality application, whereas electric steering assistance is of high criticality. In an aerospace context, flight control and surveillance applications in Unmanned Aerial Vehicles are of high and low criticality respectively. The very low acceptable failure rates (e.g. 10^{-9} failures per hour) for high criticality applications imply the need for significantly more rigorous and costly development and verification processes than required by low criticality applications.

Combining high and low criticality applications on the same hardware platform raises issues of time separation and composition; it must be possible to prevent the timing behaviour of high criticality applications from being disturbed by low criticality ones, otherwise both need to be engineered to the same rigorous and expensive standards. Simple methods of achieving this separation, such as time partitioning or allocation to different cores can however be wasteful of processing resources. They may require more expensive hardware than necessary, increasing production costs, which is something industry is strongly motivated to avoid. Time composability is needed so that the timing behaviour of applications, determined in isolation, remains valid when they are composed during system integration. Without time composability integration of complex applications would become infeasible expensive. The transformation of real-time embedded systems into mixed criticality multicore and manycore systems is recognised as a strategically important research area in Europe and the USA.

The seminar focused on the two key conflicting requirements of Mixed Criticality Systems: separation between criticality levels for assurance and sharing for resource efficiency, along with the related requirement of time composability. The key research questions addressed were:

- How to provide effective guarantees of real-time performance to applications of different criticality levels via intelligent sharing of resources while respecting the requirements for asymmetric separation / isolation between criticality levels?
- How to provide asymmetric time separation between applications with different levels of criticality so that the impact of lower criticality applications on those of higher criticality can be tightly bounded independent of the behaviour or misbehaviour of the former, without significantly compromising guaranteed real-time performance?
- How to provide time composability for applications of different criticality levels, so that the timing behaviour of applications determined in isolation remains valid when they are composed during system integration?

The sessions of the seminar were structured around a set of themes. Particular attention was given to the interfaces between themes, as these are the areas that can benefit most from improved understanding and collaboration. The discussion groups were organized around the following themes that correspond to research challenges in mixed criticality systems (MCS):

- Platforms and Experimental Evaluation (see Section 5.1);
- Worst-Case Execution Time (see Section 5.2);
- Criticality (see Section 5.3);
- Probabilistic (see Section 5.4).

Organization of the Seminar

The seminar took place from 15th to 20th March 2015. The first day started with a keynote talk by Prof. Alan Burns (University of York), one of the most influential researchers in the Real-Time Systems field over the last 25 years. Alan reviewed advances in MCS research and underlined current open problems. An overview of his talk is provided in Section 3. The first day ended with presentations and feedback on real implementations (see Section 4) as well as identifying the main themes for group discussion.

The following three days started with presentations, which were followed by discussions either within the identified groups or in an open format.

The second day started with discussions about the motivation for mixed-criticality systems presented by three different participants (see Sections 4.4, 4.5 and 4.6). Different notations are used by different sub-communities and several presentations underlined these differences (see Sections 4.7, 4.8 and 4.9). An outline of the main ideas for probabilistic analysis of real-time systems provided the topics for the discussion group on probabilistic MCS (see Sections 4.10 and 4.11).

The morning of the third day commenced with discussions on the relation between time and MCS (see Section 4.11), which continued into the afternoon's hiking activity.

Starting from the fourth day a slot dedicated to anonymous mixed criticality supporters was added to the program allowing researchers new to the topic to identify open problems in MCS from the perspective of their different domains.

As detailed later in this report, the seminar enabled the real-time community to make important progress in articulating and reaching a common understanding on the key open problems in mixed criticality systems, as well as attracting new researchers to these open problems (see Section 6). The seminar also provided an ideal venue for commencing new collaborations, a number of which are progressing towards new research publications, see Section 7.

The seminar has helped define a research agenda for the coming years that could be supported by follow-up events, given the strong interest expressed by the participants of this seminar.

As organizers, we would like to thank Prof. Reinhard Wilhelm for encouraging us to submit the seminar proposal, Dagstuhl's Scientific Directorate for allowing us to run a seminar on mixed criticality systems, and to the staff at Schloss Dagstuhl for their superb support during the seminar itself. Finally, we would like to thank all of the participants for their strong interaction, presentations, group discussions, and work on open problems, sometimes into the early hours of the morning. We were very pleased to hear about the progress of new found collaborations, and to receive such positive feedback about the seminar itself. Thank you to everyone who participated for a most enjoyable and fruitful seminar.

2 Table of Contents

Executive Summary

Keynote

Overview of Talks

Working Groups

Open Problems

New collaborations

3 Keynote

3.1 Mixed Criticality – A Personal View

Alan Burns (University of York, GB)

In this talk I want to address four topics:

- The notion of mixed criticality
- A overview of the literature on mixed criticality
- An augmented system model for mixed criticality
- Open Issues in mixed criticality research

The third of these topics is addressed in a separate abstract in Section 3.2. Notes on the other topics are provided below. As this is an extended abstract, derived from a talk, I will not include citations of the many works that have been published on Mixed Criticality. For accurate accrediting of the work alluded to below I refer readers to the Review from York (updated every 6 months and funded by the MCC project) available from: http://www-users.cs.york.ac.uk/~burns/.

It is important to be clear on the notion of 'criticality' as it is used in the, now extensive, literature on mixed critically. To me the notion is primarily based on the consequences and, to some extent, the likelihood of failure. A classification is therefore obtained by some form of hazard (or risk) analysis following a process usually defined in a Standard. All potential hazards much be mitigated during the design and implementation of both the hardware and software architectures. Software components, perhaps implemented within a run-time thread or task, will be assigned a criticality level (although different names are used for this classification in different Standards and application domains). If the late running of a task can contribute to a potential hazard then there must be evidence to support the view that such a deadline miss is sufficient unlikely. Such evidence will come from WCET analysis of the code and scheduling theory. It may also rely on run-time checks and enforcement.

The level of hazard, and the assignment of an assurance, integrity or criticality level will dictates the level of hardware redundancy and the procedures required in the design, verification and implementation of the code. There is considerable cost implications in (justifiable) begin able to reduce the classification of the software within a system.

To me, 'mixed criticality' is a means of dealing with the inherent uncertainty in a complex system. It is a means of providing efficient resource usage in the context of this uncertainty. It is also the means of protecting the more critical work when faults occur; including where assumptions are violated (rely conditions are false).

A mixed critically system is therefore not a mixture of hard and soft deadlines, nor is it a mixture of critical and non-critical components. Moreover it is not only concerned with delivering isolation and non-interference. And it is certainly not about dropping tasks to make a system schedulable. All of these ideas are, I believe, misconceptions about the nature of a mixed criticality system.

So if a mixed criticality approach is a means of dealing with uncertainly, where does this uncertainty come from? The primary source of uncertainly, as recognised in Vestal's initial paper, comes from WCET estimation. We know that WCET cannot be known with certainty. All estimates have a probability of being wrong (too low). But all estimates are attempting to be safe (pessimistic). In particular $C(LO)$ is a valid engineered estimate with the belief that $C(LO) > \text{WCET}$[1]. Beliefs can be misplaced of course, which is why systems must be built to be resilient in the face of faults. But there must be a high level of confidence, perhaps expressed as a probability, that the assertion $C(LO) > \text{WCET}$ is true.

Other forms of uncertainty can come from the environment of the system. An event driven system must make assumptions about the intensity of the events it must deal with (in a timely fashion). Again this cannot be known with certainty. So the 'Load' parameters (however they may appear in the scheduling analysis) need to be estimated (safely). In particular, the minimum arrival interval (T) for a sporadic task (i.e. the assumed minimum interval between two events from the same source) as assumed at even the lowest criticality level of the system much be safe, that is $T(LO) < T(real)$.

Critical systems need to demonstrate survivability. Faults will occur and some level must be tolerated. One source of faults is that relate to the assumptions upon which the verification of the timing behaviour of the system was based eg. WCET, arrival rates, etc. A common notion in the fault tolerance literature is the idea of a fault model. Fault models provide a means of assessing/delivering survivability. For example:

- full functional behaviour with a certain level of faults;
- Graceful Degradation for more severe faults.

Graceful Degradation is a controlled reduction in functionality, aiming to preserve safety. So within the mixed criticality domain: if any task executes for more than $C(LO)$ and all HI-criticality tasks execute for no more than $C(HI)$ then it can be demonstrated that all HI-criticality tasks meet their deadlines.

As a strategy for Graceful Degradation a number of schemes in Mixed Criticality literature have been proposed:

- Drop all lower critical work
- Drop some, using notions of importance etc.
- Extend periods and deadlines (elastic task model)
- Reduce functionality within low and high criticality tasks

This strategy should perhaps be extended to issues concerning the $C(HI)$ bound also being wrong!

If tasks are dropped/aborted then this cannot be arbitrary – the approach must be related back to the software architecture and task dependencies. So if task A is closely coupled to task B then either drop both or neither. Recovery must also relate to the needs of the software (e.g. dealing with missing/stale state).

What I want to emphasis with the above discussion is that the dropping of functions can never be seen as part of the normal behaviour of the system. That would not be acceptable to any system's developer. Rather it is a means of protected the most critically functions during a system overload, which itself is due to a fault, with may occur due to the inherent uncertainly in the system's behaviour and environment.

Another issue that arises in the mixed criticality literature is the use of a 'criticality mode' to capture the behaviour of the system when all functions are being deliver, and other modes that relates to reduced functionality. We have tended to call these modes (in a dual-criticality system) LO-criticality mode when all is well, and HI-criticality mode when only HI-criticality functions are guaranteed. This terminology is I feel misleading, the normal behaviour of the system when all functions are timely should be called 'normal'.

[1] Many papers on mixed criticality assume two critically levels, LO and HI, and two estimates of WCET related to these two levels: $C(LO)$ and $C(HI)$, with $C(LO) \leq C(HI)$.

After a fault, and degraded functionality it should be possible for the system to return to full functionality (i.e. normal mode). After all, a 747 can fly with 3 engines, but its nice to get the 4th one back. Fault recovery is therefore also an issue for mixed criticality behaviour. Indeed, as I have tried to explain, much of the work on mixed criticality systems need to be more cognisant of the available literature on fault tolerance.

Since Vestal's paper there has been at least 180 articles published (one every 2 weeks!).

Some top level observations follow. For uniprocessors:

- For FPS, AMC seems to be the 'standard' approach.
- For EDF, schemes that have a virtual deadline for the *HI*-criticality tasks seem to be standard.
- Server based schemes have been revisited.
- Not too much work on the scheduling schemes actually used in safety-critical systems, e.g. cyclic executives and non-preemptive (or cooperative) FPS.

For multiprocessor systems there are a number of schemes (extensions from uni-criticality systems). Similarly for resource sharing protocols. Work on communications is however less well represented. As indicated above, there is lots of work on graceful degradation (although few papers use that term).

Almost all papers stick to just two criticality levels. But remember *LO*-criticality does not mean no-criticality! Some papers pay lip service to multiple levels, but not many. It is still not clear what is the model we require for, say, 4 or 5 levels? To me it does not seem to make sense to have five estimates of WCET.

Notwithstanding the obvious synagy with fault tolerance there is actually little work on linking mixed criticality to fault tolerance in general. There is also little work on probabilistic assessment of uncertainty. There is some implementation work, but arguable not enough. Similarly, there is some comparative evaluations, but again not enough. There is however good coverage of formal issues such as speed-up factors.

I will finish by recording the open issues that I have identified from reading the extensive mixed criticality literature.

1. As well as looking at mixing criticality levels within a single scheduling scheme (e.g. different priorities within FPS) we need to look at integrating different schemes (e.g. Cyclic Executives for safety-critical, FPS for mission critical on the same processor).
2. More work is needed to integrate the run-time behaviour (monitoring and control) with the assumptions made during static verification.
3. We need to be more holistic in terms of ALL system resources (especially communications media).
4. There are a number of formal aspects of scheduling still to be investigated.
5. We need to be sure that techniques scale to at least 5 levels of criticality.
6. There are still a number of open issues with regard to graceful degradation and fault recovery.
7. There is little work as yet on security as an aspect of criticality.
8. We need protocols for information sharing between criticality levels.
9. We need better WCET analysis to reduce the (safe) C(HI) and C(LO) values (or at least improve our confidence in the numbers used).
10. We should look to have an impact on the Standards relevant to the application domains we hope to influence.
11. Better models for system overheads and task dependencies are needed.
12. How many criticality levels to support (and how many estimates of the sources of uncertainty to accommodate)?
13. We do not as yet have the structures (models, methods, protocols, analysis etc) that allow tradeoffs between sharing and separation to be evaluated.

To conclude, the Dagstuhl seminar is both timely and necessary in moving our research forward.

3.2 Keynote addenda: An Augmented Model for Mixed Criticality

Alan Burns (University of York, GB)

Inevitably not all papers on mixed criticality have used the same system or task model. But following on from the initial paper of Vestal [1] the most common form of the mixed criticality model is one that has a small number of criticality levels and that for each level tasks have an assigned estimation of worst-case execution time (WCET), C. Early publications on mixed criticality have often further restricted the model to have just two criticality levels, HI and LO, and therefore only two computations parameters $C(HI)$ and $C(LO)$. As we move back to using four or five criticality levels then the question arises – do we really have this number of ways of estimating WCET?

A criticality level determines many aspects of how a software function, embedded in a run-time task, is to be produced and verified. But it does not follow that distinct means of estimating or measuring execution time are available at each criticality level. In this short note we argue that two estimates are sufficient for a suitably expressive model to be defined.

Assume that the application domain of the defined system has four levels of criticality: $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$ (with A being the highest level) and one non-critical level E. Code in E, if it exists, will have no or soft deadlines and not be crucial for any function of the system. Nevertheless, it may include house-keeping functions that are useful and should be executed if possible.

Level D is the lowest criticality level. We term this the *normal* mode of the system in that during normal, fault-free, execution all code from all four criticality levels are guaranteed to meet defined timing constraints. To validate normal behaviour it is necessary for all critical code to have an estimate of its WCET that is appropriate for level D criticality. We call these estimates $C(normal)$ – in existing literature this would be called $C(\mathsf{D})$.

As code of a particular criticality level has to be produced and verified to the standard dictated by the assigned level, there must be an estimate of WCET that is linked to that criticality level. So level A has a $C(\mathsf{A})$ estimate, level B a $C(\mathsf{B})$ estimate etc. In general, we can say that all critical code has an estimate commensurate with its own criticality, we term this $C(self)$.

To summarise, all task have two estimates of WCET: $C(self)$ and $C(normal)$, with $C(self) \geq C(normal)$. For tasks of the lowest level of criticality (level D in our framework), these two estimates are the same.

The run-time behaviour of our system, following the basic idea of Vestal, is as follows:

- If all critical tasks execute for no more than their $C(normal)$ values then all critical deadlines are met.
- All tasks are prevented, by run-time monitoring, to execute for more than $C(self)$.
- If any task of criticality level X executes for more than $C(normal)$ then all tasks of critically level X and higher must continue to meet their deadlines, using estimates of WCET of $C(self)$ for tasks of criticality X and $C(normal)$ for tasks of criticality higher than X.

- In the above scenario, tasks of criticality levels below X are no longer guaranteed (and may be subject to forms of graceful degradation necessary to ensure the continuing correct execution of levels X and higher).

Note that this is a different behaviour from the one used by AMC [2], for example. In that protocol the second case would use estimates of $C(\mathsf{X})$ for the higher criticality tasks (not $C(normal)$). Of course the augmented model presented here can be directly expressed in the original Vestal model if all intermediate estimates of WCET between $C(normal)$ and $C(self)$ are assigned the same value as $C(normal)$.

As well as simplifying the model, the above behaviour is supported by at least some industrial practice. Code of level A is likely to be subject to coding standards that restrict the expressive power of the programming language employed. For example, recursion may be prohibited as may the arbitrary use of pointers and 'while' loops. It follows that level A code is more predictable and less likely to execute for more than $C(normal)$.

This augmented model is motivated by the fact that software development processes are unlikely to deliver more than two estimates of WCET. However, this does not mean that run-time behaviour cannot use computed estimates that facilitate more fine-grain control over graceful degradation. For example, if a task of level A executed for more than $C(\mathit{self})$ (i.e. $C(\mathsf{A})$) then the above model will allow all tasks of all lower criticality levels to be abandoned (to ensure level A work is preserved). However a Real-Time Systems engineer could quite reasonable argue that this reaction is overly conservative. It would be quite straightforward to use the scheduling analysis to compute a value of $C(\mathsf{C})$, with $C(\mathit{self}) > C(\mathsf{C}) > C(normal)$, and enforce the run-time behaviour: if a level A task executes for more than $C(normal)$ but no higher than $C(\mathsf{C})$ then only tasks of level D need to degrade. Sensitivity analysis for fixed priority scheduling has already been used [3] to solve a related problem.

This useful step, of computing intermediate WCET estimates, does not however detract from the application model being advocated here. This model restricted the number of external/given estimates of WCET to two.

Vestal [1], and much follow on work, has focused on WCET as the main source of uncertainty in the model of the system. Other forms of uncertainly exist – including load from the environment, faults in the hardware, power from (perhaps unreliable) sources etc. For all of these sources of uncertainty we argue that there should be two estimates. One for the normal all inclusive criticality of the system and one that reflects the particular criticality of the component.

References

1 S. Vestal. Preemptive scheduling of multi-criticality systems with varying degrees of execution time assurance. In *Proc. of the IEEE Real-Time Systems Symposium (RTSS)*, pages 239–243, 2007.

2 S.K. Baruah, A. Burns, and R. I. Davis. Response-time analysis for mixed criticality systems. In *Proc. IEEE RTSS*, 2011, pages 34–43.

3 T. Fleming and A. Burns. Incorporating the notion of importance into mixed criticality systems. In L. Cucu-Grosjean and R. I. Davis, editors, *Proc. 2nd Workshop on Mixed Criticality Systems (WMC), RTSS*, 2014, pages 33–38.

4 Overview of Talks

Mixed-criticality needs feedback from real implementation

4.1 Mixed Criticality in Multicore Automotive Embedded Systems

Sebastien Faucou (University of Nantes, FR)
sebastien.faucou@univ-nantes.fr

Introduction

The automotive industry pursues an effort toward the standardization of in-vehicle embedded systems technologies. If we focus on the topics of interest of this seminar, two standards stands out: ISO 26262, the functional safety standard; and AUTOSAR OS, the RTOS component of the standardized AUTOSAR architecture. Studying these standards gives some insight on the way mixed criticality is handled today in automotive embedded systems and allows to identify direction for future works.

ISO 26262: ASIL and freedom from interference

ISO 26262 defines two key concepts. The first one is the risk classification scheme composed of four *ASILs* (Automotive Safety Integrity Level) that range from A (the least critical) to D (the most critical). ASILs are attached to hazardous events and mapped to software components as a result of the hazard analysis. Mixed criticality requirements arise when the software components of a system have different ASILs.

The second key concept of ISO 26262 is *freedom from interference*. Freedom from interference is established when no error can propagate from low-criticality components to high criticality components. Freedom from interference does not implies full isolation but rather that interferences between criticality classes are bounded. It encompasses functional and extra-functional concerns, including timeliness and communication. In a system built on top of a shared platform, if freedom from interference can not be proved, then every component shall be designed with the requirements associated with the highest ASIL among all the co-hosted components.

AUTOSAR OS

AUTOSAR OS extends OSEK/VDX OS with several features, including *protection facilities* and support for multicore platforms. Among the protection facilities, timing protection monitors the run-time behaviour of the jobs, assuming a sporadic model. This allows to use for instance the schedulability tests developped for Vestal's model [1] and may be for some of its extensions [2] in order to validate the capacity of the system to survive to timing faults and preserve the timeliness of its most critical functions, *ie.* establishing freedom from interference in the time domain.

Protection facilities also include the possibility to partition the memory and the peripherals of the platforms between *OS-Applications* (set of tasks, interrupt handlers and shared resources) and to enforce this partitionning at run-time. These features contribute to freedom from interference in the communication domain

MC in multicore automotive embedded systems

The two parts of AUTOSAR OS protection facilities presented above are usefull but with the advent of multicore platforms, this is not sufficient. Indeed, these mechanisms do not adress the management of shared hardware resources in a mixed-criticality context. Examples of shared hardware resources include the memory bus, the SRAM banks, the shared cache level found in high-ends microcontroller for infotainment. These shared resoources are channels that allow low criticality tasks to interfere on the execution of highest criticality ones.

Two directions for futures works

Vestal's model and its extensions offer a solid theory for real-time scheduling of mixed criticality systems. Some works have been carried on in order to evaluate the pertinence of this theory in the context of Linux-based el-time ystems [3]. The same type of works remains to be done in the context of smaller (embedded) real-time systems, taking into account and exploiting some distinctive features such as: limited hardware resources, static software, sub-millisecond deadlines, etc.

According to the current state of the art, the second direction that should be considered a priority is the design of methods to bound interferences in multicore systems. Once again, the distinctive features of automotive embedded systems should be exploited to propose low footprint mechanisms, amenable to static analysis.

References

1 S. Vestal. Preemptive scheduling of multi-criticality systems with varying degrees of execution time assurance. In *Proc. of the IEEE Real-Time Systems Symposium (RTSS)*, pages 239–243, 2007.

2 S.K. Baruah, A. Burns, and R. I. Davis. Response-time analysis for mixed criticality systems. In *Proc. IEEE RTSS*, 2011, pages 34–43.

3 Huang-Ming Huang, Christopher D. Gill, and Chenyang Lu. Implementation and evaluation of mixed-criticality scheduling approaches for periodic tasks. In *RTAS*, pages 23–32, 2012.

4.2 Efficiently Safe: Decoding the Dichotomy in Mixed-Criticality Systems

Arvind Easwaran (Nanyang Technological University, SG)
arvinde@ntu.edu.sg

An increasing trend in embedded systems is towards open computing environments, where multiple functionalities are developed independently and integrated together on a single computing platform. This trend is evident in industry-driven initiatives such as ARINC653 Integrated Modular Avionics (IMA) in avionics and AUTOSAR in automotive. An important notion behind this trend is the safe partitioning of separate functionalities, primarily to achieve fault containment. This raises the challenge of how to balance the conflicting requirements of partitioning for safety assurance and efficient resource sharing for economical benefits. The concept of *mixed-criticality*, first introduced by Vestal [1], appears to be important in meeting these dichotomous goals.

In many safety-critical systems, the correct behavior of some functionality (e.g., flight control) is more important ("critical") to the overall safety of the system than that of another (e.g., in-flight entertainment). In order to certify such systems as being correct, they are conventionally assessed under certain assumptions on the worst-case run-time behavior. For example, the estimation of Worst-Case Execution Times (WCETs) of code for highly critical functionalities involves very conservative assumptions that are unlikely to occur in practice. Such assumptions make sure that the resources reserved for critical functionalities are always sufficient. Thus, the system can be designed to be fully safe from a certification perspective, but the resources are in fact severely under-utilized in practice.

In order to close such a gap in resource utilization, Vestal [1] proposed the mixed-criticality task model that comprises of different WCET values. These different values are determined at different levels of confidence ("criticality"), based on the following principle. A reasonable low-confidence WCET estimate, even if it is based on measurements, may be sufficient for almost all possible execution scenarios in practice. In the highly unlikely event that this estimate is violated, as long as the scheduling mechanism can ensure deadline satisfaction for highly critical applications, the resulting system design may still be considered as safe.

To ensure deadline satisfaction of critical applications, mixed-criticality studies make pessimistic assumptions when a single high-criticality task executes beyond its expected (low-confidence) WCET. They assume that the system will either immediately ignore all the low-criticality tasks (e.g., [2, 3, 4]) or degrade the service offered to them (e.g., [5, 6, 7]). They further assume that all the high-criticality tasks in the system can thereafter request for additional resources, up to their pessimistic (high-confidence) WCET estimates. Although these strategies ensure safe execution of critical applications, they have a serious drawback as pointed out in a recent article [5]. When a high-criticality task exceeds its expected WCET, the likelihood that all the other high-criticality tasks in the system will also require more resources is very low in practice. Therefore, to penalize all the low-criticality tasks in the event that some high-criticality tasks require additional resources seems unreasonable.

In practice, most mixed-criticality systems are comprised of independently developed components. For wide applicability, it is then natural that mixed-criticality strategies must consider the impact of WCET violations across component boundaries. To the extent possible, these strategies must limit this impact to within components, so that other components in the system (high- as well as low-criticality ones) can continue their execution uninterrupted. Considering the fact that different approaches may be used to compute the WCET estimates within different components, we believe this is a reasonable requirement because otherwise components can be unfairly penalized due to ill-computed WCET estimates of other components. One extreme solution that addresses this requirement is the worst-case reservation-based approach that completely isolates components but severly under-utilizes the resources. On the other hand, most of the recent mixed-criticality studies such as those mentioned above, completely ignore these component boundaries but still under-utilize resources due to unrealistic assumptions.

Challenges

Based on the above discussions, we now summarize some of the main challenges in designing an "*efficiently safe*" mixed-criticality system.

1. What is a good scheduling and execution strategy that can use component boundaries to provide partitioning between functionalities, but at the same time is resource-efficient and adequately supports low-criticality tasks? Some initial results in this direction are presented in a recent article [8].

2. Is there any motivation to provide hierarchical scheduling for component-based mixed-criticality systems? Since mixed-criticality scheduling strategies naturally isolate the critical tasks from non-critical ones, can we meet the partitioning requirements using a non-hierarchical scheduling framework?
3. The failure of existing studies to understand the implication and feasibility of abruptly stopping/modifying the active low-criticality tasks is another important shortcoming that has been highlighted [5]. Can we also address this issue by limiting the impact of WCET violations to within components?

References

1 S. Vestal. Preemptive scheduling of multi-criticality systems with varying degrees of execution time assurance. In *Proc. of the IEEE Real-Time Systems Symposium (RTSS)*, pages 239–243, 2007.

2 S. K. Baruah, A. Burns, and R. I. Davis. Response-time analysis for mixed criticality systems. In *Proc. IEEE RTSS*, 2011, pages 34–43.

3 S. Baruah, V. Bonifaci, G. D'Angelo, H. Li, and A Marchetti-Spaccamela. The Preemptive Uniprocessor Scheduling of Mixed-Criticality Implicit-Deadline Spo- radic Task Systems. In *ECRTS*, 2012.

4 A. Easwaran. Demand-based Scheduling of Mixed-Criticality Sporadic Tasks on One Processor. In *RTSS*, 2013.

5 A. Burns, S. Baruah, K. M. Phan, and I. Shin Towards a more practical model for mixed criticality systems In *WMC*, 2013.

6 H. Su and D. Zhu An elastic mixed-criticality task model and its scheduling algorithm. In *DATE*, 2013.

7 P. Huang, G. Giannopoulou, N. Stoimenov, and L. Thiele. Service adaptions for mixed-criticality systems. In *ASP-DAC* 2014. In Proceedings of the Asia and South Pacific Design Automation Conference (ASP-DAC).

8 X. Gu, A. Easwaran, K. M. Phan, and I. Shin Resource Efficient Isolation Mechanisms for Mixed-Criticality Systems (Technical Report: Nanyang Technological University). http://ntu.edu.sg/home/arvinde/preprints/ECRTS15.pdf In *ECRTS*, 2015.

4.3 Adding Cache and Memory Management to the MC^2 (Mixed Criticality on Multicore) Framework

James H.Anderson (The University of North Carolina at Chapel Hill, US)
anderson@cs.unc.edu

Keywords: cache coloring, set partitioning, way partitioning, memory banks, multicore

The multicore revolution is having limited impact in safety-critical application domains. The key reason is the "one out of m" problem: when checking real-time constraints on a platform with m cores, analysis pessimism can easily negate the processing capacity of the "additional" $m-1$ cores. Two major approaches have been investigated previously to address this problem: mixed-criticality allocation strategies that seek to provision less critical software components less pessimistically, and hardware management strategies that seek to make the underlying platform itself more predictable. While both approaches seem somewhat promising, neither by itself has proven capable of practically resolving the "one out of m" problem. In this talk, the results of an ongoing development effort will be discussed in which both approaches are being applied together. This effort is based on a new variant of the MC^2 (mixed-criticality on multi core) [1, 2, 3] framework that enables tasks to be isolated by criticality level with respect to the hardware resources they access. Experimental results will be presented that demonstrate the efficacy of the overall framework (if such results are available by the time the workshop is held).

References

1 J. Herman, C. Kenna, M. Mollison, J. Anderson, and D. Johnson. RTOS support for multicore mixed-criticality systems. In *RTAS*, 2012.

2 M. Mollison, J. Erickson, J. Anderson, S. Baruah, and J. Scoredos. Mixed criticality real-time scheduling for multicore systems. In *ICESS*, 2010.

3 B. Ward, J. Herman, C. Kenna, and J. Anderson. Making shared caches more predictable on multicore platforms. In *ECRTS*, 2013.

... and the industry has created the need for the mixed-criticality

4.4 Mixed-criticality in Railway Systems: A Case Study on Signaling Application

A. Cohen (INRIA, FR),V. Perrelle (Technological Research Institute SystemX, FR), D. Potop-Butucaru (INRIA, FR), E. Soubiran (Alstom Transport, FR), Z. Zhang (INRIA & Technological Research Institute SystemX, FR)
albert.cohen@inria.fr

Since the early 2000's almost every new metro project in the world make use of a standardized railway signalling system called Communication Based Train Control (CBTC) (IEEE 1474). Previously to CBTC, conventional signalling train control systems were relying almost exclusively on track circuits, wayside signals and operating procedures to ensure train protection and operation. In order to ensure better operational performance (e.g. effective utilization of the transit infrastructure), CBTC systems rest on three pillars: "*Automatic train control (ATC) based on high-resolution train location determination, independent of track circuits*"; "*high-capacity and bidirectional train-to-wayside data communications*"; and "*train-borne and wayside computing units that execute vital functions*". Functions are classified within three families that are: Automatic Train Protection (ATP), Automatic Train Operation (ATO) and Automatic Train Supervision (ATS). The level of criticality differs from a family to another and without loss of generality, one can state that ATP functions are mostly safety critical functions (SIL4 regarding to CENELEC 50126), whereas ATO and ATS gather functions of low criticality (from SIL0 to SIL2). As a matter of fact, CBTC systems are in essence Mixed-critical systems. Furthermore the mainstream evolution of those systems tends toward more functional integration on more powerful computing units. ATP and ATO functions that were traditionally distributed on different computing units (both on wayside and train-borne) tends now to be deployed on the same computing units and thus sharing resources.

FSF (Safe and reliable embedded system) is an IRT SystemX project positioned on two topics, the first one is about the conception of signalling applications (typically ATO/ATP

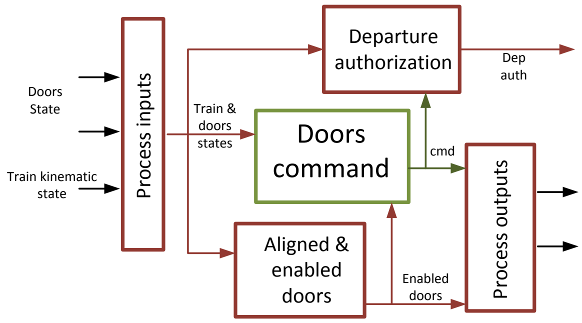


Figure 1 Simplified view of a mixed-critical path in the Passenger Exchange component.

application) that contain both critical and non-critical parts and the second one is on execution platforms that execute those applications while offering high guarantee of safety and availability. Industrial expectations around the execution platform include the use of multi-core COTS, the use of modern RTOS that offer spatial and temporal isolation, the use of safety and availability architectural patterns (e.g. voting and redundancy), and the whole being finally hidden behind a "system abstraction layer". On top of this platform, a tooled framework is prototyped and allow one to develop, verify and deploy component based applications where components may arbitrary contains both vital and non-vital code. The project has started in May 2013, the aim of this communication is to propose a first return of experience and a positioning on how MICS will be addressed in FSF.

Alstom Transport has defined an applicative case study that, while being limited to one single ATC function, is representative of the complexity in term of vital/non-vital code interweaving, operational performance and availability. The system function is called "passenger exchange". This function takes control on the train when this one is safely docked at a station; it organizes the exchange of passengers (train and station doors opening/closing) while protecting them from any untimely train movement or non-aligned doors opening and finally gives the departure authorization when all safety conditions are met. The functional specification is made of more than 300 requirements (natural language + SysML), and the functional architecture is made of about twenty sub functions.

PE is designed as a system component with a vital and a non-vital part. At this level a component is roughly a packaging unit that exposes to the exterior world a set of ports (in or out) and that is characterized by a set of behaviours that depend on the operational context. One shall notice that there are no restrictive design constraints on dataflow dependency between the vital and non-vital part. This component is then implemented as a set of software components which are this time exclusively vital or non-vital.

The vital and the non-vital parts need to communicate together. To illustrate this fact, we give an example from the case study. Let's take two constraints from the vital requirements. The first one states that the component shall not transmit a departure authorization when the doors are open or opening. It obviously implies that the component isn't sending any command to open the doors. The second requirements state that the system shall not send commands to doors which are not safe to open. (e.g. because they are not aligned) To meet these requirements, a vital subcomponent compute which doors shall be enabled. This information is given to a non-vital component which compute which commands to send and when to send them to achieve the assigned mission. Since this last component is not bound by the same safety constraints, we can't give the same confidence to its output. Hence, we need to process these output with a vital component. The door commands are matched against the enable set of doors initially computed and truncated if necessary to fulfil the requirements. Finally, another vital component reads the door state and the commands which have just been computed to decide whether to give a departure authorization or not.

This example shows that in our case study, there can be numerous communications between small components of mixed criticality.

Synchronous approaches

Synchronous languages

Data-flow synchronous languages, such as Lustre [1] or Signal [2] have been designed in the 80's for program real-time safety critical embedded systems. Since then, they have been widely used in industrial applications [3]. These languages emphasise a correct-by-construction approach, ensuring bounded memory and execution time. Moreover, they are praised for their predictable behaviour and formally defined semantics.

Recently, the problem of scheduling multi-rate, mixed-critical synchronous programs have been addressed. At first for uni-processor [6] then for multi-processors [14]. Outside the scope of mixed-criticality there were also several attempts to distribute synchronous data-flow languages [4, 5]. Recent work have been done to develop these languages to target multi-core platforms through the programming of parallelism [12]. This work introduces futures in Lustre-like languages giving the guarantee that the sequential semantics is preserved.

Automatic allocation, partitioning, and scheduling

Due to their use in the avionics industry, synchronous languages have been considered early on as an input formalism for the automatic or semi-automatic synthesis of real-time implementations. Most significant in this direction are previous results by previous work by Sorel *et al.* [7] on the AAA/SynDEx methodology and tool for distributed, but not time-triggered, real-time implementation of multi-periodic synchronous specifications, previous work by Caspi *et al.* on the use of Lustre/Scade in the real-time implementation of Simulink over multi-processor platforms based on the time-triggered partitioned bus TTA [8], and previous work by Forget *et al.* [9] on the specification and implementation of multi-periodic applications over a time-triggered platform using the Prelude language.

But none of these approaches allow us to take into account all the characteristics of our case study in order to allow automatic mapping. In particular, none of them has support for ensuring the time and space separation between application parts with different *criticalities*.

This is why we considered in this project a new tool, named LoPhT [11, 10], which allows the automatic mapping of applications onto platforms following the ARINC 653 time and space partitioning mechanisms. The LoPhT tool has the flow pictured in Fig. 2. It takes as input deterministic functional specifications provided by means of synchronous data-flow models with multiple modes and multiple relative periods. These specifications are extended to include a real-time characterization defining task periods, release dates, and deadlines. Task deadlines can be longer than the period to allow a faithful representation of complex end-to-end flow requirements. The specifications are also extended with allocation constraints and partitioning information meant to represent the criticality of the various tasks, as well as information on the preemptability of the various tasks. Starting from such specifications, the LoPhT tool performs a fully automatic allocation and off-line scheduling onto partitioned time-triggered architectures. Allocation of time slots/windows to partitions can be fully or partially provided, or synthesized by LoPhT. The mapping algorithms of LoPhT take into

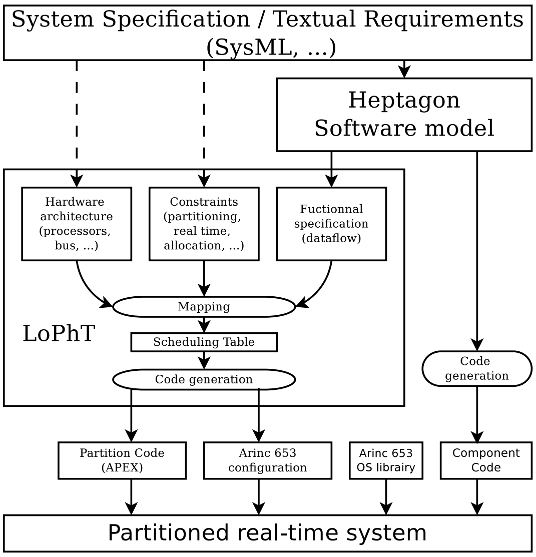

System Specification / Textual Requirements (SysML, ...)
Heptagon Software model
Hardware architecture (processors, bus, ...)
Constraints (partitioning, real time, allocation, ...)
Fuctionnal specification (dataflow)
Mapping
LoPhT
Scheduling Table
Code generation
Code generation
Partition Code (APEX)
Arinc 653 configuration
Arinc 653 OS librairy
Component Code
Partitioned real-time system

Figure 2 The design flot.

account the communication costs. The off-line mapping algorithms of LoPhT use advanced mapping techniques such as software pipelining and pre-computed preemption to improve schedulability and minimize the number of context switches.

Case study

The case study PE has been implemented and a first demonstrator has been produced. The challenge for this first demonstrator was to propose a framework for on the one hand the design and implementation of components and on the other hand the design of signalling application its partitioning and scheduling.

Choice of software modelling language. We chose to use the language HEPTAGON, very similar to LUSTRE and featuring novel constructions and novel optimisations. Two criteria have influenced the choice of the language. First, the functional specification defined at system level and allocated to software components have been written in a reactive and mostly equational way. It was thus very natural to implement it in a synchronous data-flow language. Second, the normative referential (CENELEC 50128) recommends the use of formal methods for the development of critical software while making no restrictive assumption on the language used for the non critical part. Synchronous languages are a good trade-off since they enable the use of formal methods (for instance model checking or abstract interpretation) while providing a sufficient power of expression to implement non-critical components. Finally, having a single language to develop both critical and non-critical components allows not only the early simulation of functional behaviour without integration effort but also the rationalisation of competence in the software development team.

Scheduling and partitioning with LoPhT. Entering in the flow of the LoPhT tool, detailed above, requires the definition of its input specification. For the functional specification part, direct translation is possible from Scade/Heptagon to the input formalism of LoPhT (which is also a data-flow synchronous language).

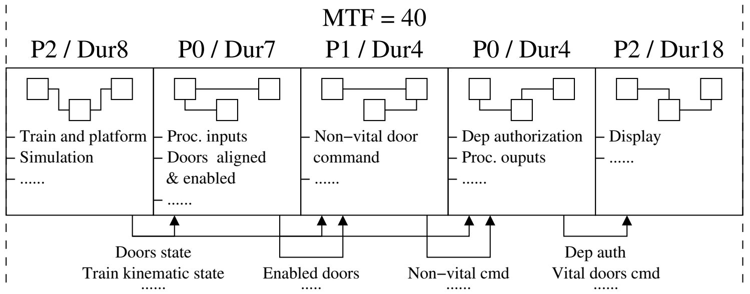


Figure 3 The partitional scheduling result of LoPhT.

Technical realisation. We developed the Passenger exchange components following a five step process:

1. In a SysML environment, we produced a component design that realize the Passenger exchange function. System requirements are traced and refined to define atomic components that correspond to software components and that are either safety-critical, mission-critical or non-critical.
2. We matched every atomic component to a Heptagon node realizing the functional behaviour.
3. Depending on the SIL of the component verification activities have been led but are out of the scope of this communication.
4. Thanks to a dataflow model we have produced a small signalling application that gathers several components including Passenger exchange, Train/Station interfaces and a simulation of other system functions (train driving...). From Heptagon point of view the application is trivially a node assembly. At this stage, a first executable code is produced to simulate the application behaviour, however no insurance is given on spatial isolation.
5. In LoPhT, the application functions are decomposed into three partitions, which are "P0: critical", "P1: non-critical" and "P2: environment". Meanwhile the function durations are given (we suppose that each function takes one time scale unit). The scheduling result is presented in the Figure 3. Five windows are created. The first one has eight functions of the "environment" partition. In this window, the states of the doors and the train kinematic are analyzed and finally sent to the corresponding windows. The second window containing seven functions of the "critical" partition, which decide the critical control commands. The third window is composed of four functions of the "non-critical" partition. In this window, the non-critial control commands are generated. The fourth window does the remaing critical works and the last window gives the feedbacks to the "environment" component, such as a display screen. The code generated by LoPht is simulated and tested on the POK OS [13].

To interface the software model with LoPhT, we needed to add a bit of glue code. Each Heptagon node representing an atomic function has been wrapped with a static memory. This wrapper exposes the two `reset` and `step` functions needed by the synchronous paradigm. We extended the C backend of Heptagon to be able to generate these wrappers automatically.

References

1 P. Caspi, D. Pilaud, N. Halbwachs and J.A. Plaice. LUSTRE: A Declarative Language for Real-time Programming. *Proceedings of the 14th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages, ACM*, 178–188, 1987.

2 A. Benveniste, P. Le Guernic and C. Jacquemot. Synchronous Programming with Events and Relations: The SIGNAL Language and Its Semantics. *Sci. Comput. Program., Elsevier North-Holland, Inc.*, 16, 103–149 1991.

3 A. Benveniste, P. Caspi, S. A. Edwards, N. Halbwachs, P. L. Guernic, L. Robert and D. Simone. The synchronous languages 12 years later. *Proceedings of The IEEE*, 64–83, 2003.

4 P. Aubry and P. Le Guernic. On the desynchronization of synchronous applications. *11th International Conference on Systems Engineering, ICSE*, 96, 1996.

5 P. Caspi, A. Curic, A. Maignan, C. Sofronis, S. Tripakis and P. Niebert. From Simulink to SCADE/Lustre to TTA: A Layered Approach for Distributed Embedded Applications. *Proceedings of the 2003 ACM SIGPLAN Conference on Language, Compiler, and Tool for Embedded Systems, ACM*, 2003.

6 S. Baruah. Semantics-preserving Implementation of Multirate Mixed-criticality Synchronous Programs. *Proceedings of the 20th International Conference on Real-Time and Network Systems, ACM*, 11–19, 2012.

7 M. Marouf, L. George and Y. Sorel. Schedulability analysis for a combination of non-preemptive strict periodic tasks and preemptive sporadic tasks. *Proceedings ETFA*, 2012.

8 P. Caspi, A. Curic, A. Magnan, C. Sofronis, S. Tripakis and P. Niebert. From Simulink to SCADE/Lustre to TTA: a Layered Approach for Distributed Embedded Applications. *Proceedings LCTES*, 2003.

9 C. Pagetti, J. Forget, F. Boniol, M. Cordovilla and D. Lesens. Multi-task Implementation of Multi-periodic Synchronous Programs. *Discrete Event Dynamic Systems*, 21, 307–338, 2011.

10 T. Carle and D. Potop-Butucaru. Predicate-aware, Makespan-preserving Software Pipelining of Scheduling Tables. *ACM Trans. Archit. Code Optim*, 11, 12:1–12:26, 2014.

11 T. Carle, D. Potop-Butucaru and Y. Sorel and D. Lesens. From dataflow specification to multiprocessor partitioned time-triggered real-time implementation. *INRIA*, 2012.

12 A. Cohen, L. Gérard and M. Pouzet. Programming Parallelism with Futures in Lustre. *Proceedings of the Tenth ACM International Conference on Embedded Software, ACM*, 197–206, 2012.

13 J. Delange, L. Pautet, and P. Feiler. Validating Safety and Security Requirements for Partitioned Architectures. *Proceedings of the 14th Ada-Europe International Conference on Reliable Software Technologies, Springer-Verlag*, 30–43, 2009.

14 E. Yip, M. Kuo, D. Broman, and P. S. Roop. Relaxing the Synchronous Approach for Mixed-Criticality Systems. In *Proceedings of the 20th IEEE Real-Time and Embedded Technology and Application Symposium (RTAS)*, pages 89–100. IEEE, 2014.

4.5 Confidence in Mixed-Criticality Multi-Core

Zoë Stephenson and Mark Pearce (Rapita Systems Ltd., UK)
{zstephenson,mpearce}@rapitasystems.com

Keywords: WCET, assurance, assurance deficit, argument, partitioning

For aerospace applications, CAST-32 [1] indicates that applicants need to show both that applications running on a multi-core processor have the desired behaviour, and that the characteristics of the computing platform are understood and controlled.

We believe that our process for measurement-based worst-case execution time (WCET) estimation can be extended to account for mixed-criticality systems, many-core systems and measurement-based testing for characteristics other than timing.

Hypothesis-driven Analysis

As an example, our method for exploring execution time is to perform standard functional testing and *measure* end-to-end execution times, *hypothesise* that the longest measured time is the worst that can occur, *search* for evidence to contradict this, and *repeat* the exercise until further challenge to the hypothesis is no longer practicable. We use RapiTime for on-target measurements and WCET path prediction as part of this process, but the method is not restricted to a specific tool. Any lingering aspects that cannot be addressed through this approach need to be accounted for with additional margins and protection mechanisms (CAST-32 calls these "safety nets").

Meeting CAST-32

CAST-32 contains many recommendations asking the applicant to demonstrate understanding of the target. It is explicit in MCP_Software_2 that the verification environment should be representative of the final intended hardware environment. By including iterative testing on the target as part of the analysis process, we ensure that this is the case. We can then use the guidance of CAST-32 to drive the search for counter-evidence challenging the execution time hypothesis, which provides insight into the behaviour of the processor.

Beyond Execution Time

CAST-32 is concerned with other effects that can occur because of the multi-core platform – delays in access to resources (data, devices, locks...), denial of access, out-of-order accesses or incorrect accesses. These are compatible with the hypothesis-driven approach. For example, a hypothesis that accesses are always in order may be established, and then testing improved to try to cause out-of-order behaviour. In the eventuality that the erroneous behaviour can be triggered, the testing process itself tells the analyst what to recommend to avoid triggering the behaviour.

Extension to Mixed Criticality

Mixing criticalities implies some protection between those different criticalities. This may be seen as another type of on-target constraint that the analyst can measure and challenge. It is likely that this will be needed as multi- and many-core systems become more prevalent. We expect that additional support may be needed from suppliers to be able to provide testing to show that protection works for a specific application in a specific set of configurations in a specific test environment.

Beyond Two Cores

The strategy in CAST-32 is to test what will run on each core individually, and then test them together. When viewed from the perspective of hypothesis-and-challenge testing, we suggest that it may be useful to test individual cores first in combination with a range of "test" behaviours on other cores, to try to undermine the application on the core under test with expected and unexpected use of shared resources.

Challenges

The recommendations of CAST-32 ask the applicant to explain what has been done to understand and control the target behaviour. This is a complex task, with evidence both about the behaviour of the software on target and about the process of exploring the unknown behaviour of that target. We advise using a structured argument to present this explanation as an assurance case.

Regardless of the approach taken, a significant challenge is to measure the behaviour on a multi-core system without causing further interference. To this end, we advise engagement to refine existing debug and trace capabilities so that interference is minimal and bounded, and ideally entirely non-existent.

References

1 CAST. *Multi-core Processors*. Position Paper CAST-32, May 2014

4.6 Challenges in Mixed Criticality Systems Design – Integration Issues

Rolf Ernst (TU Braunschweig, DE)
ernst@ida.ing.tu-bs.de

Current industrial developments lead to a growing number of tasks with different safety criticalities sharing the same components of an embedded system. At the same time, high performance is becoming more important. Prominent examples are the automotive and avionics domains. In the talk, we explain the complex side effects of switched Ethernet for automotive applications which make mixed critical designs hard. A main challenge arises from the many dependencies between the numerous layers of an architecture that are typically not overseen by a single person in the design process. We propose applying dependency analysis to identify possible hidden effects between function executions and between components. We conclude that mixed criticality are as complex as the underlying architectures and mechanisms. Solutions to individual problems, such as scheduling, are not sufficient, because safety (like security) is dominated by the weakest link. Challenges often arise from integration mechanisms that shall improve efficiency. Research should, therefore, address effective and efficient mechanisms for bounding interference on all levels, not only of time. Another important topic are mechanism which work under errors.

Hard and soft, low and high, how mixed-criticality makes the difference between important and urgent?

4.7 Real-time Performance Evaluation and VT Control mechanisms for the timing correct use of shared main memory

Kai Lampka (Uppsala University, SE)
kai.lampka@it.uu.se

Joint work of Lampka, Kai; Georgia Giannopolou; Nikolay Stoimenov (ETH Zurich); Jonas Flodin; Yi Wang
Main reference K. Lampka, G. Giannopoulou, R. Pellizzoni, Z. Wu, N. Stoimenov, "A formal approach to the WCRT analysis of multicore systems with memory contention under phase-structured task sets," Real-Time Systems, 50(5):736–773, 2014.
URL http://dx.doi.org/10.1007/s11241-014-9211-y

This presentation considers sets of real-time tasks executing in parallel on different cores and sharing parts of the memory hierarchy. For quantifying and handling contention at the DRAM controller this presentation presents the following recent innovations.

Worst-case response time analysis based on Timed Automata

The proposed method exploits the so-called superblock model of the work of Schranzhofer et al. [2], respectively the PRedictable Execution Model (PREM) of Pellizzoni et al. [3]. This limits the time non-determinism inherent to the occurrence of cache misses, respectively memory (data) fetches. To achieve scalability we suggest to replace some of the Timed Automata models with an abstract representation based on access request arrival curves, rather than using an individual component TA model for each core and its real-time workload.

Memory access bandwidth control

The proposed adaptive budgeting technique controls the access frequencies of applications to the main memory. To bound the interference of co-running soft real-time tasks, past works have proposed periodic server- based memory access reservation mechanisms [7, 5, 4, 6]. As the computed budgets are commonly extremely pessimistic, they reflect the worst-case rather than the normal resource use, it can be assumed that tasks under memory access budgeting experience a severe degradation of their average response time. For the hard real-time tasks, commonly implementing system control functions, this degradation is irrelevant, what matters is the guarantee that all deadlines are met. However, for user-centric soft-real time applications performance degradation should be reduced. The presented approach addresses this obstacle by dynamically changing sizes of budgets or simply ignoring them once a hard real-time tasks has terminated before its set worst case response time and there is no job release of some other hard real-time task.

References

1 K. Lampka, G. Giannopoulou, R. Pellizzoni, Z. Wu, and N. Stoimenov. A formal approach to the WCRT analysis of multicore systems with memory contention under phase-structured task sets. *Real-Time Systems*, 50(5-6):736–773, 2014.

2 A. Schranzhofer, R. Pellizzoni, J.-J. Chen, L. Thiele, and M. Caccamo. Timing analysis for resource access interference on adaptive resource arbiters. In *Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 213–222, 2011.

3 G. Yao, R. Pellizzoni, S. Bak, E. Betti, and M. Caccamo. Memory-centric scheduling for multicore hard real-time systems. *Real-Time Systems Journal*, 48(6):681–715, Nov 2012.

4 W. Jing. Performance isolation for mixed criticality real-time system on multicore with xen hypervisor. Master's thesis, Uppsala University, Department of Information Technology, 2013.

5 H. Yun, G. Yao, R. Pellizzoni, M. Caccamo, and L. Sha. Memory access control in multiprocessor for real-time systems with mixed criticality. In *Real-Time Systems (ECRTS), 2012 24th Euromicro Conference on*, pages 299–308, 2012.

6 H. Yun, G. Yao, R. Pellizzoni, M. Caccamo, and L. Sha. Memguard: Memory bandwidth reservation system for efficient performance isolation in multi-core platforms. In *Real-Time and Embedded Technology and Applications Symposium (RTAS), 2013 IEEE 19th*, pages 55–64, 2013.

7 M. Behnam, R. Inam, T. Nolte, and M. Sjödin. Multi-core composability in the face of memory-bus contention. *SIGBED Rev.*, 10(3):35–42, Oct. 2013.

8 J. Flodin, K. Lampka, and W. Yi. Combining Performance Monitoring and Resource Budgeting on Multi-core for Real-Time Guarantees. In *MCC 2013 – Sixth Swedish Workshop on Multicore Computing*, 2013.

4.8 System-level, Inter-Criticality, Multi-Core Resource Sharing with Scalable Predictability

Gabriel Parmer (The George Washington University, Washington, DC, US)
gparmer@gwu.edu

Background

Multi-core systems have proven to be a double-sided sword for embedded and real-time systems. They provide increases in computational power that promise to not only consolidate previously distributed systems together, but also to increase the computational capability, thus intelligence and functionality, of embedded systems. However, these parallel systems present a significant challenge due to the interference between tasks caused by increased resource sharing between cores. For example, different cores often share hardware resources such as last-level caches (LLC) and memory buses. Past research has addressed each of these in turn by, for example, partitioning memory [1] or cache [2]. An inescapable challenge not addressed by these techniques is the *interference caused by the sharing relationships of data-structures within software due to cache coherency.* This problem is complementary to previous approaches, and it is particularly important: a store to a cache-line can (on our 40-core, 4 socket, cache-coherent hardware) take three cycles, or *more than 27μs*, depending on coherency behavior.

Note that this is relevant to all shared structure access, and is orthogonal to the mechanisms for mutual exclusion and their resource sharing protocols. When considering such implementations, even predictable (FIFO) spin-locks that are known to be scalable in the average case (MCS locks), have worst-case latencies of 50μs which increases to 65μs if even a single cache line is modified within the critical section.

The impact of the overhead for loads and stores that access data-structures on shared cache lines not only impacts a task's response time, but also increases the interference between competing tasks. One task's data-structure access pattern in the kernel can increase the latency of another. This cross-talk makes temporal isolation difficult across criticalities. A high criticality task, tested in isolation on a system could suffer memory access latency spikes when a low criticality task is added to the system that contends a shared kernel data-structure.

Scalable Predictability

Just as the real-time community designs techniques to provide isolation given access to shared hardware resources such as cache and memory buses, this talk will discuss the major challenges in designing software to enable controlled access to data-structures that are shared between tasks of different criticalities across cores. Specifically, our goal is to provide access to data-structures shared between cores that not only scales with increasing core counts in terms of average-case performance (i.e. the per-operation overhead doesn't increase), but also in terms of the *worst-case latencies.* We call this *scalable predictability*, and it is a strong form of scalability that focuses on the *worst-case overheads from cache-line coherency traffic* – in addition to the average behaviors that are often the focus of scalable systems – and on avoiding coherency traffic all-together. Scalable predictability means that the latency bounds provided on a single core and in isolation of all lower criticalities, don't increase with a rising number of cores.

Techniques for Scalable Predictability

This talk will be discussing recent work that will appear in RTAS, and more recent research into further techniques for scalable predictability. Methods and mechanisms to handle concurrent data-structure access are derived or borrowed from techniques in the High-Performance Computing and scalable software construction realms.

Existing techniques. Here we'll briefly survey existing concurrency control mechanisms for shared data-structures, and assess them for their scalability properties. We assume that the data-structures to be protected are both accessed on multiple cores, and by tasks of multiple criticalities (though the code that defines the access methods for the data-structure is of high assurance [3]). In other words, they are typical kernel data-structures.

- *Predictable locks.* Locks are often backed by at least one common cache-line. The cache-coherency traffic due to this cache-line "bouncing" between caches results in the significant overheads discussed previously.
- *Read-write locks.* Many data-structures are read mostly, thus enabling parallel access for readers to the data structures increases parallelism. However, these locks suffer from the same deficiency with respect to scalable parallelism as normal predictable locks: large worst-case latencies due to bouncing the lock's cache line (or in some cases, multiple cache-lines).
- *Read-Copy-Update (RCU).* RCU enables very low overhead reads to data-structures, often without *any* writes to shared structures. However, modifications to such structures involves ensuring that no readers are still accessing the modified portions before returning. This heavily penalizes writers that require coherency traffic for consensus. Thus RCU is mainly used in read-mostly workloads.
- *Reference counting.* Object liveness is often interrelated with mutual exclusion as references to an object can only be removed with proper coordination between the object, and the data-structure that is referencing it. Reference counting is the pervasive technique

often used to track this, but it suffers from average-case scalability overheads, let alone issues with scalable predictability.

Techniques for scalable predictability. This talk will include an overview of a few techniques we propose as serving as a foundation for a community investigation of worst-case scalability. These include:

- *Lock-less lookup structures with fine-grained consistency control.* Operating systems often must map between an opaque identifier (*e.g.* a file descriptor, mailbox id, process identifier), and the data-structure that backs it. The very data-structures that typically provide this map often require protection with locks. Thus, regardless of which criticality is accessing the namespace, and on which core, data-structure modification, and lock cache-line bouncing significantly impact high-criticality latencies. Lock-less data-structures that rely on liveness based on the very timing properties provided by the system (see quiescence below) provide a strong foundation for worst-case scalability.
- *Explicit mapping of namespaces to cache-line accesses.* Though the objects tracked in the kernel can be located without any cache-line modifications using the previous technique. However, once located, any modifications to the object must not impact the cache-lines of other objects. The goal is to enable the users of the API to tailor their access to the namespace to explicitly avoid accesses and modifications to cache lines for any other object. When data-structures require modification, the goal here is to enable the modifications to be at the finest possible granularity, so that scalable predictability is only compromised when cores and criticalities access the exact same object.
- *Quiescence-based liveness.* By avoiding any shared cache-line modification on kernel object lookup, the liveness question must be answered. Can a kernel object be deallocated, or are there parallel accesses to it on another core? Quiescence-based memory reclamation answers this question by ascertaining a point in the future when all references accessed before the object is freed, cannot persist. Real-time and predictable systems offer a significant benefit here: we can base our quiescence period on the latency bounds provided by the system itself.

Case Study: The SPeCK Kernel

This talk will discuss a case study we've conducted for scalable predictability in the SPeCK kernel [4] which is our new kernel for the Composite component-based OS. Through a combination of using the techniques listed above, it is able to provide scalable predictability guarantees for many of its most important operations. Notably, the operations used by computation in hard real-time components are worst-case scalable, and even operations that have traditionally never had scalable solutions, such as TLB coherence on page unmap, are usable in real-time computation. Additionally, SPeCK provides the features required by Composite systems: all resource management policies and most system abstractions are defined in user-level components including scheduling, memory mapping management, and I/O.

References

1 H. Kim, D. deNiz, B. Andersson, M. Klein, O. Mutlu, and R. (Raj) Rajkumar, "Bounding memory interference delay in COTS-based multi-core systems," in *RTAS*, 2014.

2 H. Kim, A. Kandhalu, and R. Rajkumar, "A coordinated approach for practical OS-level cache management in multi-core real-time systems," in *ECRTS*, 2013.

3 E. Armbrust, J. Song, G. Bloom, and G. Parmer, "On spatial isolation for mixed criticality, embedded systems," in *2nd International Workshop on Mixed Criticality Systems (WMC)*, 2014.

4 Q. Wang, Y. Ren, M. Scaperoth, and G. Parmer, "Speck: A kernel for scalable predictability," in *Proceedings of the 21st IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2015.

4.9 Mixed Criticality Support on Networks-on-Chip

Leandro Soares Indrusiak (University of York, UK)
lsi@cs.york.ac.uk

Overview

Networks-on-Chip (NoCs) are a widely used on-chip interconnect architecture for large multi and many-core processors. They provide packet-switching infrastructure for multiple types of system-wide communications, such as message passing between tasks running on different cores, data transfers between external memories and local scratchpads, or paging and coherency mechanisms for multi-level caches. In the work reported here, we focus on the first two types of communication, which deal with coarse-grain communications (i.e. messages and data blocks rather than cache lines). Thus, we consider that mixed-criticality application tasks executing over such a processor exchange data packets of different criticality levels through the NoC infrastructure. This leads to a situation where the transmission of a packet has potential impact over the latency of all the others. Therefore, the design of the NoC infrastructure must logically separate packets of different criticality, so that their distinct requirements can be satisfied even though they share the same interconnect.

Mixed-Criticality Networks-on-Chip

WPMC [1] is a protocol applied to NoCs with virtual channels (VCs) that are arbitrated at the flit-level using a priority-preemptive mechanism. This means that each output port will send out, in every cycle, a data word (flit) from the input VC with the highest priority. WPMC aims to provide hard real-time guarantees to all criticality levels (i.e. all packets will arrive by their deadlines even in the worst-case scenario) and supports sporadic as well as periodic traffic patterns. It follows Vestal's assumption [2] that application components of high criticality will be given more generous upper bounds for their timing behaviour, e.g. due to more strict analysis or to larger safety margins; and that components of low criticality, which are likely to be analysed with less strict techniques or which are given smaller safety margins, will have tighter upper bounds for their timing behaviour. In line with that approach, WPMC assumes that high-criticality traffic is likely to have potentially larger packets, or having packets injected more often into the NoC, as this would be a safer upper bound on the load it may impose to the NoC.

A key idea of WPMC, which was also used in the AMC scheduling algorithm [3], is that traffic of high criticality could also be analysed with the same techniques and safety margins used to profile low criticality traffic, and thus be given tighter upper bounds to its timing behaviour. The tight upper bounds can be used to dimension the NoC in such a way that all packets will always meet their deadline, as long as they don't exceed their low criticality upper bounds (i.e. maximum packet size, minimum packet inter-arrival interval). WPMC uses runtime monitoring to check whether all high- criticality traffic stays within their low-criticality upper bounds. The moment one of them exceeds that bound, the system is said to change into a high-criticality mode. To guarantee the timely delivery of all high-criticality packets under that mode, the NoC is allowed to drop all low-criticality traffic (as a way to achieve graceful degradation). Thus, to ensure the system is dimensioned to cope with the high-criticality mode, it must be able to support only the high-criticality traffic, but considering their more generous upper bounds. In [1], we defined the NoC mechanisms to perform the runtime monitoring, signalise mode change, and to change the NoC arbitration policies to drop low-criticality traffic. We also provided schedulability analysis to evaluate whether a given NoC is properly dimensioned to cope with the traffic produced by a given (set of) application(s) under the default low-criticality mode, as well as during and after a change to the high-criticality mode is detected.

A number of extensions and improvements to WPMC are currently being researched and developed, including:

- An alternative mode-change propagation strategy that floods the network and forces the whole NoC to change its criticality level.
- An improved credit-based flow control that allows low criticality packets to be transferred without impact on high criticality packets, even after a mode change.
- The set of conditions that must be satisfied before a mode change from high to low-criticality mode, which has not been supported by WPMC.
- Task allocation heuristics that can be used to improve schedulability of a given mixed-criticality application mapped onto a specific NoC.

This contribution will provide an overview of WPMC, and will present the progress on each one of the extensions and improvements mentioned above.

References

1 A. Burns, J. Harbin and L.S. Indrusiak. A Wormhole NoC Protocol for Mixed Criticality Systems In *Proc. IEEE RTSS*, pages 184–195, 2014.

2 S. Vestal. Preemptive scheduling of multi-criticality systems with varying degrees of execution time assurance. In *Proc. of the IEEE Real-Time Systems Symposium (RTSS)*, pages 239–243, 2007.

3 S.K. Baruah, A. Burns, and R. I. Davis. Response-time analysis for mixed criticality systems. In *Proc. IEEE RTSS*, 2011, pages 34–43.

How do we map criticalities to certification levels – a probabilistic attempt

4.10 Mapping cricalities to certification levels – a probabilistic attempt

Liliana Cucu-Grosjean and Adriana Gogonel (INRIA, FR)
{liliana.cucu,adriana.gogonel}@inria.fr

Some context

The main feature of time critical embedded systems concerns the respect of temporal constraints. The correctness of each computation within these systems depends on both the

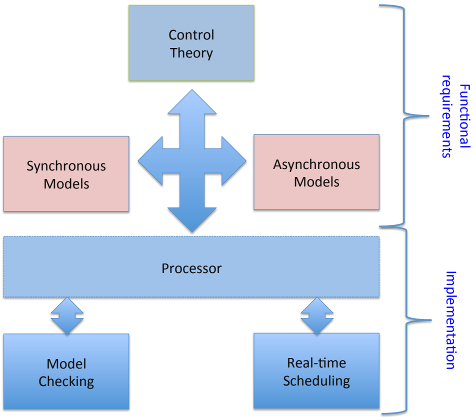


Figure 4 Different phases of the design of a time critical embedded system.

logical results of the computation and the time at which these results are produced.

The design of a time critical embedded system may have basically three main phases: (i) the description of the physical process that should be controlled, (ii) the description of the functional requirements that should be fulfilled and (iii) the description of the implementation of the time critical embedded system. During the first phase the characteristics of the physical process are described using control theory. Then a model is proposed using synchronous or asynchronous modelling and this model is verified using model checking. At the end of the second phase the designer has a model of the system that is correct with respect to the expected functional requirements. During the last phase (of implementation), the processors are taken into account and the time feasibility of the system is checked using methods like formal verification or real-time schedulability analysis. The relations between different phases of conception is provided in Figure 4.

The pessimism of all existing solutions comes mainly from the implementation phase where an absolute value is considered for the worst case execution time of a program. The arrival of modern and more complex processors (e.g., use of caches, multi- and many-core processors) increases the timing variability of programs, i.e., the absolute worst case execution time is becoming significantly larger. For instance, larger execution times require an increased number of processors or more powerful processors.

Our open problem

An intuitive solution to overcome this pessimism is the introduction by Steve Vestal [1] of the notion of mixed criticality for time critical embedded systems. This solution defines several possible values for the worst case execution time of a program on a processor and it has propagated from the original work on scheduling theory [2] to synchronous languages [3], predictable processors [4], model checking [5], etc.

Nevertheless today the mixed criticality solutions are heterogeneous and they are proposed for different phases of design without a common framework. In conclusion we identify as vital **the need for a modular framework unifying heterogeneous solutions of the design problem of mixed criticality systems without re-writing the entire theory of time critical embedded systems**.

Our intuition is that **probabilistic description of some parameters or properties of existing models is a possible solution** to the problem of designing time critical embedded systems.

Nevertheless **the introduction of probabilities is not trivial** as not every probabilistic approach may be used to study time critical embedded systems. Indeed the introduction of probabilistic descriptions in all phases of the design of time critical embedded systems should be done such that the two following properties are ensured:

1. worst case values are rare events;
2. probabilistic worst case reasoning is applicable.

References

1 S. Vestal. Preemptive scheduling of multi-criticality systems with varying degrees of execution time assurance. In *Proc. of the IEEE Real-Time Systems Symposium (RTSS)*, pages 239–243, 2007.

2 A. Burns and R. I. Davis. Mixed Criticality Systems – A Review. *Department of Computer Science, University of York, Report. Fourth edition, July 31*, 2014.

3 E. Yip, M. Kuo, D. Broman, and P. S. Roop. Relaxing the Synchronous Approach for Mixed-Criticality Systems. In *Proceedings of the 20th IEEE Real-Time and Embedded Technology and Application Symposium (RTAS)*, pages 89–100. IEEE, 2014.

4 M. Zimmer, D. Broman, C. Shaver, and E. A. Lee. FlexPRET: A Processor Platform for Mixed-Criticality Systems. In *Proceedings of the 20th IEEE Real-Time and Embedded Technology and Application Symposium (RTAS)*, pages 101–110. IEEE, 2014.

5 A.J. Boudjadar and A. David and J. Kim and K. G. Larsen and M. Mikucionis and U. Nyman and A. Skou. Degree of Schedulability of Mixed-Criticality Real-Time Systems with Probabilistic Sporadic Tasks. In *the book Theoretical Aspects of Software Engineering Conference*, 2014

4.11 Response Time Analysis for Fixed-Priority Tasks with Multiple Probabilistic Parameters

Dorin Maxim (The Polytechnic Institute of Porto, PT)
dorin@isep.ipp.pt

Joint work of Maxim, Dorin; Cucu-Grosjean, Liliana;
Main reference D. Maxim, L. Cucu-Grosjean, "Response Time Analysis for Fixed-Priority Tasks with Multiple Probabilistic Parameters," in Proc. of the IEEE 34th Real-Time Systems Symposium (RTSS'13), pp. 224–235, IEEE, 2013.
URL http://dx.doi.org/10.1109/RTSS.2013.30

Introduction

We consider a system of n synchronous tasks $\{\tau_1, \tau_2, \ldots, \tau_n\}$ to be scheduled on one processor according to a preemptive fixed-priority task-level scheduling policy. Without loss of generality, we consider that τ_i has a higher priority than τ_j for $i < j$. By synchronous tasks we understand that all tasks are released simultaneously the first time at $t = 0$.

Each task τ_i generates an infinite number of successive jobs $\tau_{i,j}$, with $j = 1, \ldots, \infty$. All jobs are assumed to be independent of other jobs of the same task and those of other tasks.

Each task τ_i is a generalized sporadic task [1] and it is represented by a probabilistic worst case execution time (pWCET) denoted by $\mathcal{C}_i$[2] and by a probabilistic minimum inter-arrival time (pMIT) denoted by $\mathcal{T}_i$.

[2] In this paper, we use calligraphic typeface to denote random variables.

The probabilistic execution time (pET) of a job of a task describes the probability that the execution time of the job is equal to a given value. A safe pWCET $\mathcal{C}_i$ is an upper bound on the pETs $\mathcal{C}_i^j$, $\forall j$ and it may be described by the relation $\succeq$ as $\mathcal{C}_i \succeq \mathcal{C}_i^j$, $\forall j$. Graphically this means that the CDF of $\mathcal{C}_i$ stays under the CDF of $\mathcal{C}_i^j$, $\forall j$.

Following the same reasoning the probabilistic minimal inter-arrival time (pMIT) denoted by $\mathcal{T}_i$ describes the probabilistic minimal inter-arrival times of all jobs. The probabilistic inter-arrival time (pIT) of a job of a task describes the probability that the job's arrival time occurs at a given value. A safe pMIT $\mathcal{T}_i$ is a bound on the pITs $\mathcal{T}_i^j$, $\forall j$ and it may be described by the relation $\succeq$ as $\mathcal{T}_i^j \succeq \mathcal{T}_i$, $\forall j$. Graphically this means that the CDF of $\mathcal{T}_i$ stays below the CDF of $\mathcal{T}_i^j$, $\forall j$.

Hence, a task τ_i is represented by a tuple $(\mathcal{C}_i, \mathcal{T}_i)$. A job of a task must finish its execution before the arrival of the next job of the same task, i.e., the arrival of a new job represents the deadline of the current job. Thus, the task's deadline may also be represented by a random variable $\mathcal{D}_i$ which has the same distribution as its pMIT, $\mathcal{T}_i$. Alternatively, we can consider the deadline described by a distribution different from the distribution of its pMIT if the system under consideration calls for such model, or the simpler case when the deadline of a task is given as one value. The latter case is probably the most frequent in practice, nevertheless we prefer to propose an analysis as general as possible and in the rest of the paper, we consider tasks with implicit deadlines, i.e., having the same distribution as the pMIT.

Problem description: We address the problem of computing the response time distributions and, implicitly, Deadline Miss Probabilities (DMP) of tasks with pMIT and pWCET. The response time of a job is the elapsed time between its release and its completion. Since we consider jobs with probabilistic parameters, the response time of a job is also described by a random variable. The DMP of a job is obtained by comparing the response time distribution of said job and its deadline, be it a probabilistic deadline or a deterministic one. This is a novel problem, and the fact that the system under consideration has more than one task parameter given as a distribution makes it a complex one.

Probabilistic response time analysis

The probabilistic worst case response time (pWCRT) $\mathcal{R}_n$ of a task τ_n in the critical instance is computed by coalescing all the distributions $\mathcal{R}_n^{i,j}$ (called copies) resulted by iteratively solving the following equation (from [2]):

$$\mathcal{R}_n^{i,j} = (\mathcal{R}_n^{i-1,head} \oplus (\mathcal{R}_n^{i-1,tail} \otimes \mathcal{C}_m^{pr})) \otimes \mathcal{P}_{pr} \tag{1}$$

The iterations end when there are no more arrival of any job i of any higher priority task τ_m that occurs within the response time distribution at the current step. A stopping condition may be explicitly placed in order to stop the analysis after a desired response time accuracy has been reached. For example, the analysis can be terminated once an accuracy of 10^{-9} has been reached for the response time. In our case, the analysis stops when new arrivals of the preempting tasks are beyond the deadline of the task under analysis, i.e., the type of analysis required for systems where jobs are aborted once they reach their deadline.

Once the jobs' response time distribution can be computed, the Deadline Miss Probability can be obtained by comparing the response time distribution with that of the deadline, as follows:

$$\mathcal{B}_i = \mathcal{R}_i \ominus \mathcal{D}_i = \mathcal{R}_i \oplus (-\mathcal{D}_i), \tag{2}$$

where the $\ominus$ operator indicates that the values of the distribution are negated.

Note that the analysis can handle any combination of probabilistic and deterministic parameters, and in the case that all parameters are deterministic the returned result is the same as the one provided by the worst case response time analysis in [3]. More details about the analysis can be found in [2].

References

1 A. Ka-Lau Mok (1983). *Fundamental Design Problems of Distributed Systems for the Hard-Real-Time Environment.* Massachusetts Institute of Technology.

2 D. Maxim and L. Cucu-Grosjean. *Response Time Analysis for Fixed-Priority Tasks with Multiple Probabilistic Parameters.* In *Proceedings of the IEEE 34th Real-Time Systems Symposium, RTSS* 2013.

3 M. Joseph and P. K. Pandya. *Finding response times in a real-time system.* In *The Computer Journal* 29(5):390–395, (1986).

What is the meaning of mixed-criticality when time is the keyword?

4.12 Viewpoints on the Timing Aspect of Mixed Criticality Systems

David Broman (KTH Royal Institute of Technology, SE)
dbro@kth.se

Mixed criticality systems can be informally defined as systems where software components, with different levels of criticality, execute on the same hardware platform. Starting with the paper by Vestal [1] in 2007, a large body of research results has been presented within the real-time community. The common research problem can be seen as the challenge of reconciling the two requirements of partitioning for safety, and sharing resources [2]. There are, however, several different viewpoints on the timing aspects of mixed criticality systems; in particular of the meaning of criticality levels. We separate between two distinct viewpoints: i) the implementation view, and ii) the specification view.

In the *implementation view*, the model and meaning of criticality level also include aspects of the implementation. That is, consideration needs to be taken to the actual hardware platform and operating system (OS) that are used. Vestal's classic task model [1] falls into this category; different WCET estimate numbers are used for different criticality levels. To be able to get these numbers, programs need to be executed and measured on the real hardware platform, or accurate timing models of the hardware need to be used when computing safe bounds of the WCET. Other variants of Vestal's model, for instance Burns and Baruah's variant [3], can also be considered to fall into the same category. Both these examples are based on software scheduling for mixed criticality systems.

Another approach is to perform the scheduling in hardware. The FlexPRET [4] processor platform is an example of hardware-based scheduling using fine-grained multithreading. In this approach, several hardware threads are used, which either fall into the category of hard real-time threads or soft real-time threads. Hard real-time threads are guaranteed to have both temporal and spatial isolation, whereas soft real-time threads do not have temporal isolation, but can steal cycles from the hard real-time threads when they are not active. Tasks with different levels of criticality can then be scheduled on either hard real-time or soft real-time threads using only hardware scheduling or with a combination of hardware and traditional software scheduling. Clearly, this approach for ensuring the timing aspects of mixed criticality systems falls under the category of the implementation view.

An alternative is to use a *specification view* for the timing aspects. For such a viewpoint, nothing within the definition of the criticality levels should say anything about implementation aspects. One such approach is to define the criticality levels using frequency bounds. Yip et al. [5] proposes such approach, where tasks are divided into three different criticality levels. Each periodic task has two frequency parameters: f_{max} and f_{min}, meaning that the task is allowed to be executed with a frequency that falls within this interval. For life critical tasks, $f_{max} = f_{min}$ and for mission critical tasks $f_{max} > f_{min}$. For non-critical tasks, f_{max} is the goal frequency and $f_{min} = 0$. Note that this task model does not say anything about the implementation technique or WCET numbers for specific platforms. The different criticality levels are specified using constraints on timing.

The idea of using timing specifications can be taking further to make it more expressive. We call this approach *programming with time*, meaning that time and timing become part of a programming model. At the Dagstuhl seminar, I presented some work-in-progress about incorporating time in a small language, by formalizing the semantics using small-step operational semantics.

References

1 S. Vestal. Preemptive scheduling of multi-criticality systems with varying degrees of execution time assurance. In *Proc. of the IEEE Real-Time Systems Symposium (RTSS)*, pages 239–243, 2007.

2 A. Burns and R. I. Davis. Mixed Criticality Systems – A Review. *Department of Computer Science, University of York, Report. Fourth edition, July 31*, 2014.

3 A. Burns, S. Baruah, K. M. Phan, and I. Shin Towards a more practical model for mixed criticality systems In *WMC*, 2013.

4 M. Zimmer, D. Broman, C. Shaver, and E. A. Lee. FlexPRET: A Processor Platform for Mixed-Criticality Systems. In *Proceedings of the 20th IEEE Real-Time and Embedded Technology and Application Symposium (RTAS)*, pages 101–110. IEEE, 2014.

5 E. Yip, M. Kuo, D. Broman, and P. S. Roop. Relaxing the Synchronous Approach for Mixed-Criticality Systems. In *Proceedings of the 20th IEEE Real-Time and Embedded Technology and Application Symposium (RTAS)*, pages 89–100. IEEE, 2014.

4.13 Mapping the landscape of mixed criticality systems research

Sanjoy K. Baruah (University of North Carolina at Chapel Hill, US)
baruah@cs.unc.edu

There appears to be general agreement on the definition of mixed-criticality systems: a mixed-criticality system is a system in which functionalities of different specified criticalities are implemented upon a shared platform. Beyond this general definition, however, the situation parallels that described in John Godfrey Saxe's poem *The Blind Men and the Elephant*: different interpretations abound, each highlighting selected aspects of mixed-criticality systems while choosing to minimize (or ignore) other aspects. It is important to be cognizant of these different perspectives, and to better understand the different contexts and requirements that motivate the different interpretations; else, we end up with different sub-communities speaking across one another and misunderstanding each other – the same terms mean very different things to different people. Some of the different perspectives that I am aware of are listed below.

Perspective 1. A perspective that is found in the safety-critical systems industry holds that different criticality levels are user-specified attributes that have no additional semantic interpretation – different criticality levels are not really comparable to each other. Different requirements, such as correctness criteria, are specified for different criticality levels; each functionality is expected to satisfy the requirements for its specified criticality level. For instance, a presentation[3] advocating this perspective explicitly states:

> *What [a mixed-criticality system] is NOT: A system where system approach sacrifices lower criticality applications for whatever purpose.*

If the different criticality levels are assumed incomparable in this manner, then a reasonable objective of mixed-criticality research should be to devise mechanisms and policies that enable *isolation* amongst the different criticality levels. The research questions here then seek to determine how best to provide such isolation upon modern platforms (such as multicores), particularly as systems become increasingly more complex.

Perspective 2. In much of the mixed-criticality real-time scheduling literature, it is assumed that the different criticality levels correspond to different degrees of importance: a functionality that is semantically more important is assigned greater criticality. (For example, safety-critical functionalities may be accorded greater criticality than mission-critical ones.) The research objective here is to seek more resource-efficient implementations of such mixed-criticality systems. Such research may be classified into two broad categories according to the different approaches adopted:

- **Perspective 2A *(Run-time adaptation)*.** Under this approach a mixed- criticality system starts out executing functionalities of different criticalities, based upon optimistic assumptions regarding resource requirements. If these optimistic assumptions are observed during run-time to not hold, then the system adapts its run-time behavior to allocate additional resources to the more critical functionalities; less critical functionalities receive less resources and may experience a consequent degradation in performance.
- **Perspective 2B *(Pre-run-time verification)*.** Such research is based on the principle that resources must be provisioned under more conservative assumptions to more critical functionalities, in order to meet their more stringent correctness criteria – those typically include a requirement that such functionalities have their correctness validated to higher levels of assurance. Some of these provisioned resources may then be reclaimed during system design time itself, and used to make performance guarantees at lower levels of assurance to less critical functionalities. (This is the approach that is currently commonly referred to within the real-time scheduling theory community as the *Vestal approach*, in recognition of the fact that is was first proposed in a paper[4] by Vestal.) The research

[3] Michael Paulitsch (Thales) and Jan Nowotsch (Airbus Group). *Monitoring Techniques in COTS Multicore Processors in Mixed-Criticality Systems with Focus on Temporal Aspects*. Torrent Workshop, Toulouse, December 12, 2014. Slides available at http://www.irit.fr/torrents/seminars/20141212/20141212-paulitsch.pdf (Date accessed: 2015/02/18).

[4] Steve Vestal. *Preemptive Scheduling of Multi-criticality Systems with Varying Degrees of Execution Time Assurance*. Proceedings of the IEEE Real-Time Systems Symposium (RTSS), pp. 239–243. 2007.

activities within this category include seeking novel innovative ways of achieving such design-time resource reclamation.

Summary. Above, three different perspective to mixed-criticality systems research have been identified. Perspective 1 differs greatly from the two perspectives 2A and 2B, while the differences between 2A and 2B are somewhat more subtle. It is important to identify additional perspectives that may exist, and to study the relationships between these different perspectives in order to understand whether there are commonalities that allow for mutually beneficial interaction amongst advocates of the different perspectives, perhaps leading to the development of common research agendas.

4.14 Some Open Problems in Mixed-Criticality Scheduling

Pontus Ekberg (Uppsala University, SE)
pontus.ekberg@it.uu.se

I list a few open problems that I consider foundational for our understanding of mixed-criticality scheduling theory. Sprinkled among the questions are a few related claims, these are accompanied by much waving of hands. The questions concern the scheduling of mixed-criticality workload of the common "Vestal-type". In addition, they are restricted to systems with two criticality levels (LO and HI) running on a preemptive uniprocessor. Not because that is necessarily the most interesting case, but because we need to understand the basics first.

Let us start by considering static collections of mixed-criticality jobs, and denote with MC-JOB-SCHEDULABILITY the decision problem of whether a given collection of jobs has a correct online (i.e., non-clairvoyant) schedule. Further, let us slightly abuse established notation and denote with EDF-VD the *family* of schedulers that follow these rules:

1. Schedule all jobs J_i in EDF order, but according to their *virtual deadlines* v_i instead of absolute deadlines d_i.
2. In LO-criticality mode,
   a. $v_i = d_i$ for LO-jobs and
   b. $v_i \in [a_i, d_i]$ for HI-jobs.
3. In HI-criticality mode,
   a. $v_i = \infty$ for LO-jobs and
   b. $v_i = d_i$ for HI-jobs.

Baruah et al. [1] showed MC-JOB-SCHEDULABILITY to be strongly NP-complete for any constant number of criticality levels. The hardness part of their proof is a reduction from 3-PARTITION. It is fairly easy to see that all feasible job collections they construct are schedulable by some scheduler in the EDF-VD family, and given such a scheduler it can be verified whether it is correct in polynomial time. Therefore, it must be hard to identify which scheduler in the EDF-VD family to use.

Claim

Finding an optimal assignment of virtual deadlines for MC job collections when using EDF-VD is strongly NP-hard.

Why this focus on the EDF-VD family? It is because I believe the answer to the following question to be "yes".

Question

Given any collection of (2-level) MC jobs that is online schedulable, is there always a correct scheduler in the EDF-VD family for it

Now we turn to MC sporadic tasks instead. Let MC-Sporadic-Schedulability be the corresponding decision problem and let us abuse notation again and denote with EDF-VD the family of schedulers that behave as before, but with a static virtual deadline $v_i \in [0, d_i]$ per task τ_i, that is applied to all jobs of τ_i. Unfortunately, a strong link that we usually have between job collections and sporadic tasks does not exist for MC systems.

Claim

The synchronous arrival sequence is not a worst case for MC sporadic tasks.

If non-integer arrival times are allowed, the situation is even worse.

Claim

There are sporadic MC task systems which are unschedulable only when some arrival times are non-integer.

The ability to restrict attention to one or a few concrete cases is a very useful property for analysis. Can the SAS be replaced by some other case?

Question

For a given MC sporadic task set, can we efficiently identify some restricted set of job sequences that are the worst cases?

The lack of the SAS as a guaranteed worst case means that we can not trivially extend the hardness proof of Baruah et al. to sporadic MC tasks, but it seems fair to suspect that MC-Sporadic-Schedulability is also NP-hard. However, it is easy to see that it is coNP-hard via a reduction from the corresponding non-MC problem, and therefore it seems reasonable to suspect that it is neither in NP nor in coNP.

Question

What is the complexity of MC-Sporadic-Schedulability?

A closely related question is how to optimally schedule a set of sporadic MC tasks.

Question

What scheduling policies are optimal for sporadic MC tasks?

Unfortunately, the family of EDF-VD schedulers does not appear to be it, though hopefully there are some others that are also efficient at runtime.

Claim

When non-integer arrival times are allowed, there are (2-level) sporadic MC task sets that are online schedulable, but for which there are no correct schedulers in the EDF-VD family.

References

1 S. Baruah, V. Bonifaci, G. D'Angelo, H. Li, and A. Marchetti-Spaccamela. The Preemptive Uniprocessor Scheduling of Mixed-Criticality Implicit-Deadline Sporadic Task Systems. In *ECRTS*, 2012.

WCET – the central notion of mixed-criticality

4.15 Runtime monitoring of time-critical tasks in multi-core systems

Christine Rochange (Paul Sabatier University – Toulouse, FR)
rochange@irit.fr

Joint work of Kritikakou, Angeliki; Baldellon, Olivier; Pagetti, Claire; Rochange, Christine; Roy, Matthieu
Main reference A. Kritikakou, C. Pagetti, O. Baldellon, M. Roy, C. Rochange, "Run-Time Control to Increase Task Parallelism In Mixed-Critical Systems," in Proc. of the 26th Euromicro Conf. on Real-Time Systems (ECRTS'14), pp. 119–128, IEEE, 2014.
URL http://dx.doi.org/10.1109/ECRTS.2014.14

Existing WCET computation methods [1] consider pessimistic situations with permanent conflicts. This leads to WCET estimates that are safe but far from the frequent case, and then to over-provisioning time slots for tasks.

The objectives of the proposed approach are to relax constraints on scheduling so that it can consider less safe but more realistic predictions of execution times. A recovery mechanism monitors high-criticality tasks to check whether they can miss their deadlines: this is done based on the remaining WCET which is dynamically updated along the execution. If a hazard is detected, highly-criticality tasks are allowed to finish their execution in a contention-free mode.

We introduced a scheme based on extended control flow graphs and partial timing information that is computed offline and stored in a table looked up at runtime to update the remaining WCET.

References

1 R. Wilhelm et al. *The worst-case execution-time problem: overview of methods and survey of tools.* ACM Transactions on Embedded Computing Systems (TECS), 7(3), 2008.

4.16 Timing Analysis for Multi/Many-core Platforms

Jan Reineke (Universität des Saarlandes, DE)
reineke@cs.uni-saarland.de

Joint work of Reineke, Jan; Doerfert, Johannes; Wilhelm, Reinhard

Timing analysis seeks to answer the following question: Can a given task set be scheduled to meet all deadlines on a particular execution platform? If the execution platform is a single-core processor, timing analysis is a fairly well-understood problem. For such platforms timing analysis is commonly split into two phases:

1. *Worst-case execution time (WCET) analysis* determines for each task a bound on its execution time, independently of the other tasks.
2. *Schedulability analysis* determines whether all deadlines can be met based on these WCET bounds.

If the execution platform is a multi- or many-core processor such a clean separation into WCET and schedulability analysis is hard to maintain. Due to interference on shared resources, such as buses, caches, and DRAM-based main memory, the execution time of a task depends strongly on its execution context [1].

I discuss four approaches to timing analysis for multi- and many-core processors and their respective benefits and drawbacks:

1. The *Integrated-analysis approach*: Analyze the entire task set at once in a combined WCET and schedulability analysis. This is practically infeasible even for the analysis of two co-running tasks.
2. The *Murphy approach*: Determine a context-independent WCET bound. Perform schedulability analysis using these bounds. This can be extremely pessimistic: Radojkovic et al. [2] report a 14-fold slowdown due to interference on a shared L2 cache and memory controller, negating all performance benefits of using a multi-core processor.
3. The *Abstract interference approach*:
   1. Characterize the interference on shared resources generated by each task.
   2. Determine interference-aware WCET bounds, i.e., mappings from the amount of interference experienced to WCET bounds.
   3. Perform an extended schedulability analysis taking into account the information from 1 and 2.
4. The *Isolation approach*: Isolate tasks running on different cores by partitioning shared resources in time and space. This re-enables the single-core two-phase timing analysis approach. However, the question arises how to split the resources among the cores. An ingredient of an informed partitioning decision is *architecture-parametric timing analysis* [3]; a WCET analysis that determines how the execution time depends upon the amount of available resources.

References

1 A. Abel, F. Benz, J. Doerfert, B. Dörr, S. Hahn, F. Haupenthal, M. Jacobs, A. H. Moin, J. Reineke, B. Schommer, R. Wilhelm. *Impact of Resource Sharing on Performance and Performance Prediction: A Survey.* In *In CONCUR* 2013.

2 P. Radojkovic, S. Girbal, A. Grasset, E. Quinones, E., S. Yehia, F. J. Cazorla: *On the evaluation of the impact of shared resources in multithreaded COTS processors in time-critical environments.* ACM Transactions on Architecture and Code Optimization 8(4), January 2012.

3 J. Reineke, J. Doerfert: *Architecture-Parametric Timing Analysis.* In *In RTAS* 2014.

4.17 Analysis of pre-emptive systems with caches

Sebastian Altmeyer (University of Amsterdam, NL)
altmeyer@uva.nl

Main reference S. Altmeyer, "Analysis of Preemptively Scheduled Hard Real-time Systems," Dissertation, Universität des Saarlandes, 2013.
URL http://nbn-resolving.de/urn:nbn:de:bsz:291-scidok-52797

Proving timing correctness of an embedded system is traditionally a two-step approach: Timing analysis derives upper bounds on the execution times of tasks in isolation, called worst-case execution times (WCET). Scheduling analysis determines if each task complies with its timing constraints when scheduled according to a predefined scheduling policy. Timing constraints are typically defined by a task's period and a task's deadline, both determined by the physical environment. Hence, tasks are assumed to be fully characterized by a triple consisting of a period, a deadline and an execution time bound, i.e. the WCET of the task.

While this verification process provides a useful separation of concerns and a clean interface between the two steps, it fails to account for the complexity of modern embedded systems; already in the case of uniprocessor systems: History-sensitive hardware components, foremost caches, impact the system performance beyond the scope of a task. This is especially problematic in the case of pre-emptive scheduling, where the execution time of a pre-empted task strongly depends on whether previously cached data has been evicted during pre-emption or whether it is still resident in the cache. The additional execution time due to cache eviction is called cache-related pre-emption delay (CRPD). Consequently, a task's execution time can not be analyzed independently anymore.

There are three different solutions to this problem: (i) one can inflate the execution time bounds to account for the CRPD, (ii) one can avoid CRPD by using cache partitioning, or (iii) one can adapt the timing verification process to include the CRPD as part of the task model. Solution (i) and (ii) enable the reuse of the common task model, but potentially at the cost of substantial pessimism or degraded performance. Solution (iii) requires the highest effort, but allows us to compute safe and precise bounds. The timing analysis must provide metrics for the cache-reuse (the set of useful cache blocks) and the memory footprint (set of evicting cache blocks) of each task. The scheduling analysis then needs to correctly account for these metrics and needs to identify the worst-case pre-emption scenarios, which strongly depend on the metrics provided by the timing analysis.

The timing verification process for pre-emptively scheduled uni-processors with caches may serve as a blueprint for the multicore timing verification. In the case of multicore systems, the independence-assumption of the timing analysis is violated not only due to a common memory hierarchy, but also due to a shared bus system. This shared bus causes additional interference and creates a dependency not only between tasks scheduled on the same core, but also between tasks scheduled on all other cores.

Based on what we have learned for the analysis of pre-emptive systems, we can formulate the educated guesses that the notion of WCET alone is not sufficient to correctly represent the complex behaviour on multicore systems, that a precise analysis restricts the hardware components to be used and that the complete timing verification process needs to be addressed and revised instead of just one of the two sides.

Mixed-criticality models are the answer to adaptive time critical systems?

4.18 Using Mixed-Criticality to Reason about Temporal Correctness in Uncertain & Dynamic Environments

Nathan Fisher (Wayne State University, US)
fishern@cs.wayne.edu

Starting with the seminal paper by Steve Vestal at RTSS 2007 [1], avionics has been the most frequently-cited motivating application domain for the development of mixed-criticality scheduling theory (MCST). The reasons are quite clear: integrating multiple avionic subsystems with different criticalities and certification levels requires guarantees that lower-criticality subsystems do not have a negative effect upon the temporal correctness of higher-criticality subsystems. Given the initial progress of the MCST research community towards addressing these system- integration goals in avionics (and related application domains), this Dagstuhl seminar is an ideal setting to reflect upon the potential broader implications (beyond system integration) of the resulting MCST from the past eight years. Specifically, I would like to raise the question of *how can MCST results be leveraged in the design of adaptive real-time systems executing in dynamic and uncertain physical environments (esp., power-aware control systems)*?

Similar Notions of Uncertainty. One important insight that has been gained from MCST research is the ability to make formal timing guarantees in the presence of uncertain execution times. For instance, in a system with two criticality levels, HI and LO, the typical model specifies that when each job's total execution time does not exceed the LO-criticality bound, then the system is considered temporally correct if all jobs (both HI and LO criticality) meet their respective deadlines; however, whenever any job exceeds its LO-criticality execution bound, then the system must only guarantee that each HI-criticality job meets its deadline. Thus, with this model of mixed criticality, a system designer is able to reason about the temporal correctness of the system without knowing an exact execution time bound for some subset of the jobs.

The area of adaptive real-time control systems often requires reasoning about execution uncertainty from a similar, but slightly different perspective. Consider the problem of maintaining the CPU temperature below a specified threshold. Under typical environmental conditions, the processor can execute normally and not exceed its temperature threshold. However, if the environmental temperature increases, the CPU is unable to dissipate the heat generated from computation as efficiently. To guard against a temperature violation, modern CPUs often have dynamic voltage/frequency scaling (DVFS) capabilities to permit a reduction in the CPU heat generation. For real-time systems these adaptive DVFS changes present a challenge in reasoning about the temporal correctness of the system given that the thermal operating environment may be dynamic and unpredictable; using DVFS will create uncertainty in the execution time of the underlying jobs and may require some to be aborted or deferred.

Opportunities & Challenges. The similar notions of execution-time uncertainty present an opportunity to "port" some of the scheduling algorithms and associated analysis developed for MCST to the domain to adaptive real-time systems. Recent work on using mixed-criticality upon processors with varying execution speeds [2] may be one avenue to unify the notions of uncertainty used for MCST and power-aware real-time control systems. However, there are some fundamental differences in the settings that may present challenges in immediately applying MCST to such systems:

1. **Differing Mode-Change Semantics:** Traditional MCST appears to view changing modes from LO to HI as a rare event. Conversely, for systems executing in a dynamic environment, changing modes continuously to adapt is fundamental to their design. A recent talk by Alan Burns at WMC 2014 surveyed some adaptive criticality mode- change protocols that may prove useful for adaptive real-time system design [3]; the differences between these MCST-based protocols and multi-modal protocols developed specifically for power-aware control systems (e.g., [4]) warrant further discussion.
2. **Number of Operating Modes:** In power-aware systems, more operating modes (e.g., voltage/frequency levels) leads to more fine-grained control. Each of these operating modes can be viewed in MCST parlance as a "criticality level". Unfortunately, it seems that scaling the number of criticality levels beyond two is a non-trivial objective. Thus, it may be a worthwhile exercise to investigate whether the setting of discrete control (e.g., mode changes will occur at periodic intervals corresponding to the controller's sampling interval) can lead to some simplifications that permit an increased scaling of criticality levels.

References

1 S. Vestal. Preemptive scheduling of multi-criticality systems with varying degrees of execution time assurance. In *Proc. of the IEEE Real-Time Systems Symposium (RTSS)*, pages 239–243, 2007.

2 S. Baruah and Z. Guo. Scheduling Mixed-Criticality Implicit-Deadline Sporadic Task Systems upon a Varying-Speed Processor. *Proceedings of the IEEE Real-Time Systems Symposium (RTSS)*, pp. 31–40, 2014.

3 A. Burns. System Mode Changes – General and Criticality-Based. *Proceedings of the 2nd International Workshop on Mixed Criticality Systems*, pp. 3–8, 2014.

4 M. Ahmed and N. Fisher. Tractable Schedulability Analysis and Resource Allocation for Real-Time Multimodal Systems. *ACM Transactions on Embedded Computing Systems.* 13 (2s), January 2014.

4.19 Augmenting Criticality-Monotonic Scheduling with Dynamic Processor Affinities

Bjoern B. Brandenburg (MPI-SWS – Kaiserslautern, DE)
bbb@mpi-sws.org

Consider the problem of scheduling a dual-criticality workload consisting of high- and low-criticality sporadic real-time tasks on top of a fixed-priority (FP) scheduler. Each high-criticality (HC) task T_i has both a high- and a low- criticality WCET estimate, denoted e_i^L and e_i^H, resp., and low- criticality (LC) tasks are required to meet their deadlines only if no HC task exceeds its LC WCET estimate.

task	criticality	p_i	e_i^L	e_i^H
T_a	low	2	1	–
T_b	high	10	3	6
T_c	low	2	1	–
T_d	high	10	2	5

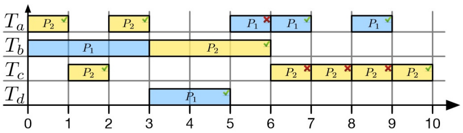


Figure 5 In this example, T_b exceeds e_b^L at time 3. Its affinity is then set to $\{P_1, P_2\}$, which allows T_b to finish on P_2. T_d is isolated; T_a and T_c miss one and three deadlines.

From a pragmatic point of view, FP scheduling with *criticality-monotonic* priorities [1], where HC tasks have higher priority than LC tasks, holds considerable appeal: it is simple, provides obvious isolation for HC tasks, and imposes no runtime overheads.

Unfortunately, as LC tasks may be more *urgent* than HC tasks (i.e., they may have shorter periods or more constraining deadlines), it is not always feasible to assign criticality-monotonic priorities [1]. For example, the task set $\tau_1 = \{T_a, T_b\}$ (as specified in Fig. 5), which consists of a LC task T_a that is urgent (i.e, it has a short period $p_a = 2$) and a HC task T_b that is less urgent ($p_b = 10$) but more costly ($e_b^L = 3$), cannot be scheduled on a uniprocessor with criticality-monotonic priorities: the LC task T_a, if given a lower priority than T_b, may miss deadlines even if no job of T_b exceeds e_b^L.

Similar urgency vs. criticality conflicts also arise on multiprocessors. For instance, the task set $\tau_2 = \{T_a, T_b, T_c, T_d\}$ cannot be scheduled with criticality-monotonic priorities on $m = 2$ cores using either *global* or *partitioned* FP scheduling: under global scheduling, the HC tasks T_b and T_d can cause the more-urgent LC tasks T_a and T_c to miss deadlines even with LC execution costs, and under partitioned scheduling, T_b and T_d need to be assigned to different partitions, but neither can be co- located with T_a or T_c. However, while scheduling τ_1 with criticality-monotonic priorities is infeasible on a uniprocessor, τ_2 *can* be scheduled with criticality- monotonic priorities on two processors—provided *processor affinities* are used to shield urgent tasks in the LC case.

Exploiting Arbitrary Processor Affinities (APAs)

Contemporary OSs such as Linux, Windows, QNX, or VxWorks provide flexible APIs to explicitly set a task's processor affinity, which is the set of processors on which it may execute. In particular, task affinities can be restricted to arbitrary processor sets and changed at arbitrary times during runtime. This can be exploited to render criticality-monotonic scheduling feasible.

Consider the following strategy for scheduling τ_2 on two processors P_1 and P_2: **(1)** Tasks are assigned criticality-monotonic priorities. **(2)** T_a and T_c may execute on both P_1 and P_2. **(3)** T_b and T_d may initially execute only on processor P_1. **(4)** When a HC job J_x of T_b (resp., T_d) fails to complete after e_b^L (resp., e_d^L) time units, it updates its processor affinity to include both P_1 and P_2. (The processor affinity of any other task is *not* changed.) **(5)** A HC task's affinity is reset when it completes its job.

A possible schedule is shown in Fig. 5: at time 3, when it becomes known that T_b's job requires more than $e_b^L = 3$ time units to complete, it relaxes its processor affinity to include P_1 and P_2. Consequently, under a FP scheduler with *strong APA semantics* [2] — which, intuitively, is an APA scheduler that *shifts* higher-priority tasks from one processor to another if that is required to enable lower-priority tasks with more- constraining affinities to be scheduled — T_b shifts to P_2, which enables T_d to be scheduled on P_1. As T_b handles its increased demand on P_2, T_d is protected from undue interference. LC tasks are not dropped, but may temporarily incur deadline misses.

Remarks and outlook

We have observed that an APA interface – readily available in current, already certified RTOSs – allows the timeliness requirements of urgent LC tasks to be reconciled with the desirable simplicity of criticality- monotonic scheduling. The sketched approach offers several practical benefits: HC tasks exceeding their LC WCET are effectively given a "dedicated" processor to cope with increased demand; only the currently executing task's affinity is adapted, which keeps runtime overheads low and independent of the number of tasks; there is no "mode change" and LC tasks are not abandoned, just temporarily delayed; and budget enforcement is not required.

Of course, the above example works only because of simplifying assumptions. We believe, however, that it is possible to generalize the approach to an arbitrary number of HC tasks and also to *weak* APA schedulers [2] such as those found in QNX and Linux.

References

1 S. K. Baruah, A. Burns, and R. I. Davis. Response-time analysis for mixed criticality systems. In *Proc. IEEE RTSS*, 2011, pages 34–43.

2 F. Cerqueira, A. Gujarati, and B. Brandenburg. Linux's processor affinity API, refined: *Shifting* real-time tasks towards higher schedulability. In *RTSS*, 2014.

4.20 Adaptive Uni-processor Fixed Priority Pre-emptive Probabilistic Mixed Criticality

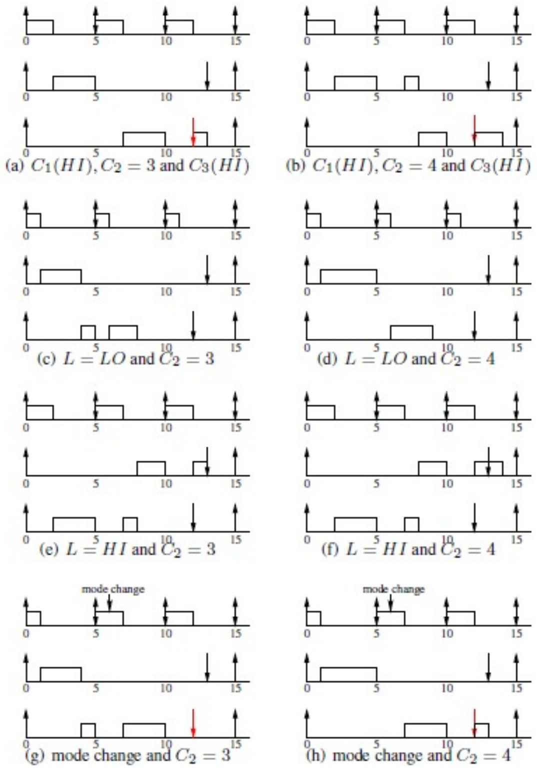
Yasmina Abdedda (Université Paris-Est, LIGM UMR CNRS 8049, ESIEE Paris, FR)
yasmina.abdeddaim@esiee.fr

Keywords: fixed priority; probabilistic scheduling; mixed criticality

Extended Abstract

According to [1], the most effective fixed priority approach for scheduling mixed criticality systems is the Adaptive Mixed Criticality (AMC) approach. This approach uses the assumption that no low criticality task is released when the system moves to high criticality. Our goal is to propose an adaptive approach for a model where low criticality tasks have a probabilistic computation time [2]. When the systems moves to high criticality level, the set of low criticality tasks are not ignored but their tolerated probability deadline miss is modified. More formally, we consider a system defined as a set of probabilistic periodic real-time tasks $\{\tau_1, \dots, \tau_n\}$ having a certain level of criticality: high (HI) or low (LO). Each task τ_i is defined as a tuple (L_i, T_i, D_i, C_i) with $L_i \in \{LO, HI\}$ the criticality of the task, T_i its period, D_i its constrained deadline and C_i is its worst-case execution time discrete random variable. We consider that the random variables $C_i, i = 1, \dots, n$ are independent such that for every task τ_i:

1. If $L_i = LO$, the sample space of C_i is $\{C_i(1), \dots C_i(m_i)\}$ and the probability distribution of C_i is the function $f_{C_i} : [1, m_i] \to \mathbb{N}^*$ with $\sum_{j=1}^{m_i} f_{C_i}(C_i(j)) = 1$.
2. If $L_i = HI$, the sample space of C_i is $\{C_i(LO), C_i(HI)\}$ with $0 < C_i(LO) \leq C_i(HI)$ and the probability distribution of C_i is a function $f_{C_i} : [LO, HI] \to \mathbb{N}$ with $f_{C_i}(x) = 1$ if $L = x$ and $f_{C_i}(x) = 0$ if $L \neq x$ where $L \in \{LO, HI\}$ is the criticality of the system.



(a) $C_1(HI), C_2 = 3$ and $C_3(HI)$

(b) $C_1(HI), C_2 = 4$ and $C_3(HI)$

(c) $L = LO$ and $C_2 = 3$

(d) $L = LO$ and $C_2 = 4$

(e) $L = HI$ and $C_2 = 3$

(f) $L = HI$ and $C_2 = 4$

(g) mode change and $C_2 = 3$

(h) mode change and $C_2 = 4$

Figure 6 $\tau_1 = (HI, 5, 5, C_1)$, $\tau_2 = (LO, 15, 13, C_2)$, $\tau_3 = (HI, 15, 12, C_3)$, $C_1 = \{C_1(LO) = 1, C_1(HI) = 2\}$, $C_2 = \{3, 4\}$, $f_{C_2}(C_2 = 3) = f_{C_2}(C_2 = 4) = 0.5$, $C_3 = \{C_3(LO) = 3, C_3(HI) = 4\}$, $P^{LO} = 0$ and $P^{HI} = 0.5$. Priority order: τ_1, τ_2, τ_3 not feasible but feasible if $D_3 = 14$.

The system behaves as described bellow (see Figure 6):

1. At the beginning of the execution, the criticality of the system is $L = LO$, and if a task τ_i with $L_i = HI$ does not notify its completion after the execution of $C_i(LO)$ time unit, the criticality of the system moves from $L = LO$ to $L = HI$,
2. When $L = LO$: **(a)** a task τ_i is executed if it is active and no higher priority task is active, **(b)** the probability of a deadline miss of all high criticality tasks is 0 and the probability miss of all low criticality task is less then a constant P^{LO}.
3. When $L = HI$: **(a)** for every task τ_i, if $L_i = LO$, τ_i is executed if it is active and no higher criticality or priority task is active, and if $L_i = HI$, τ_i is executed if it is active and no high criticality task of higher priority is active, **(b)** the probability of a deadline miss of all high criticality tasks is 0 and the probability of a deadline miss of all low criticality tasks is less then $P^{HI} \geq P^{LO}$.

References

1 S. K. Baruah, A. Burns, and R. I. Davis. Response-time analysis for mixed criticality systems. In *Proc. IEEE RTSS*, 2011, pages 34–43.

2 D. Maxim and L. Cucu-Grosjean. *Response Time Analysis for Fixed-Priority Tasks with Multiple Probabilistic Parameters.* In *Proceedings of the IEEE 34th Real-Time Systems Symposium, RTSS* 2013.

4.21 MC Scheduling on Varying-Speed Processors

Zhishan Guo (University of North Carolina at Chapel Hill, US)
zsguo@cs.unc.edu

Keywords: varying-speed processors, model combination

Introduction and Motivation

Most existing research on Mixed-Criticality (MC) scheduling (see [1] for a review) has focused on dealing with different WCET estimations of a single piece of code. This is typically a consequence of different tools for determining worst case execution time (WCET) bounds being more or less conservative than each other.

This narrative is now being repeated with respect to *processor speeds*. Modern powerful and energy-efficient processors are yielding innovations that result in varying speed during run-time. For example, [2] describes a mechanism such that late signals can be recovered by delaying the next clock tick, so that logical faults do not propagate to higher (i.e., the software) levels. In a Globally Asynchronous Locally Synchronous (GALS) circuit, local clocks can be affected by signals propagating between different synchronous modules in an asynchronous manner.

Research on such varying-speed platform may lead to better understanding of a wider range of problems. For example, in data communication of automobiles, aircrafts, or wireless sensor networks, time-sensitive data-streams must be transmitted over potentially faulty communication channels, where a high bandwidth is provided under most circumstances yet only guaranteeing a lower bandwidth.

Model and Existing work

A varying-speed processor is modeled as follows: under normal circumstances, it completes at least one unit of execution during each time unit, while it may fall into a degrade mode at any instant, during which it can only complete $x \in [s, 1)$ units of execution during each time unit, for some (known) threshold $s < 1$. It is not a priori known when, or whether, such degradation will occur. Similar to other MC scheduling problems, we seek a strategy that guarantees to complete all jobs by their deadlines under normal (LO-criticality) behaviors, while simultaneously guaranteeing to complete all HI-criticality jobs if either the platform (or the jobs) suffer from degradation (HI-criticality) behaviors. Note that here we are considering a combination of various aspects that MC may arise from, including periods, WCETs, processing speeds, etc.

Based upon the properties of the platform and the workload, we classify those problems into four categories:

1. Self-Monitoring: A self-monitoring (SM) processor immediately knows its execution speed during run-time[5] while non-monitored (NM) one may not.
2. Number of processors: Either uniprocessor, or multiprocessor.
3. Workload model: One shot job set, or sporadic/periodic task set.
4. Single(S)- or Multiple(M)- Worst case execution time (WCET) per job.

[5] Similar to Linux command `cpufreq-info`, SM platform has access to processor speeds, while NM processor may only identify degradation upon some job not signaling its finishing on time.

Table 1 Existing work on scheduling MC sets on varying-speed uniprocessor.

–	Jobs & S-WCET	Tasks & S-WCET	Jobs & M-WCET	Tasks & M-WCET
SM Uniproc.	[3] [4][6]	[3] [4][7] [8]	[5]	[8]
NM Uniproc.	[7]	[8]	[5]	[8]
SM Multiproc.	[6]	–	–	–
NM Multiproc.	–	–[8]	–	–[5]

Table 1 lists existing work on MC scheduling upon varying-speed platforms.

Further Directions

Most current work only deals with one-shot jobs or implicit-deadline sporadic tasks, and the generalization to constrained deadlines is not trivial. Also, as shown in Table 1, much remains to be done regarding multiprocessors – the degraded mode upon such platforms needs to be completely specified. If different processors are assumed to degrade to different speeds, the resulting degraded platform may become a heterogeneous one, for which the MC scheduling problem is totally open.

References

1 A. Burns and R. I. Davis. Mixed Criticality Systems – A Review. *Department of Computer Science, University of York, Report. Fourth edition, July 31*, 2014.

2 D. Bull, et al. A power-efficient 32b ARM ISA processor using timing-error detection and correction for transient-error tolerance and adaptation to PVT variation. In *IEEE ISSCC* 2010, pages 284-285.

3 S. Baruah and Z. Guo. Mixed-criticality scheduling upon varying-speed processors. IEEE RTSS 2013.

4 Z. Guo and S. Baruah. Mixed-criticality scheduling upon varying-speed multiprocessors. Leibniz Transactions on Embedded Systems, 1(2): 3:1–3:19, 2014.

5 Z. Guo and S. Baruah. The concurrent consideration of uncertainty in WCETs and processor speeds in mixed-criticality systems. Under submission.

6 Z. Guo and S. Baruah. Mixed-criticality scheduling upon varying-speed multiprocessors. IEEE DASC 2014, pp. 237–244.

7 Z. Guo and S. Baruah Mixed-criticality scheduling upon unmonitored unreliable processors. SIES 2013, pp. 161–167.

8 S. Baruah and Z. Guo. Scheduling Mixed-Criticality Implicit-Deadline Sporadic Task Systems upon a Varying-Speed Processor. *Proceedings of the IEEE Real-Time Systems Symposium (RTSS)*, pp. 31–40, 2014.

[6] The strong NP-hardness of non-preemption scheduling under such case is also shown in [3] and [4].

[7] Regarding scheduling tasks, [3] and [4] only provide necessary conditions and a sharing-based (fluid) scheduling scheme, which is not impractical due to too many preemptions.

[8] We may model a NM varying-speed processor with the multi-WCET MC model, and apply some existing MC scheduling work, while being somewhat pessimism, which is similar as [8].

Mixed-criticality systems: different models for scheduling problems (open or not)

4.22 Speedup bounds for multiprocessor scheduling

Suzanne van der Ster (Vrije Universiteit Amsterdam, NL)

Introduction

When studying mixed-criticality (MC) task systems, we are interested in worst-case behaviors and determining feasibility. Since determining feasibility exactly is hard, we design approximate feasibility tests. If such a test returns “feasible”, the task system is guaranteed to be feasible on a processor running at speed , while if it returns “infeasible”, the task set is guaranteed to be infeasible when processed on a unit-speed processor. The factor is also called the speedup factor (also for scheduling algorithms corresponding to the feasibility test).

Known results

There are two main paradigms for scheduling task systems on multiprocessors: global and partitioned scheduling. In the former, all tasks can use all machines, and jobs can even be migrated from one machine to another. In the partitioned scheduling approach, each task has to be assigned to one of the machines such that all its jobs have to be executed on this specific machine. For MC sporadic task sets, the only known results on multiple machines are for 2-level implicit-deadline task sets, i.e., for task sets such that the period equals the relative deadline for all tasks. Those results are based on an earlier result for implicit-deadline task systems on a single machine. For single-machine scheduling, the algorithm EDF-VD (introduced in [1]) is a modification of the well-known EDF policy, where higher-criticality tasks are assigned tighter deadlines (that are called virtual deadlines), in order to be able to meet all their deadlines, even in case of a criticality switch. It was shown [2] that any feasible 2-level MC task system can be scheduled successfully by EDF-VD on a processor running at speed 4/3.

This result is used in the partitioned scheduling policy in [3]. The algorithm given has a speedup for m machines of at most $8/3 - 4/3m$.

An alternative approach, only interesting from a theoretical point of view, is viewing the MC scheduling problem as a $VECTORSCHEDULING$ problem (see [4] for a definition), where each dimension corresponds to a criticality level. For this problem, a PTAS exists, when the number of dimensions is a constant. Combining the PTAS with EDF-VD yields that any task system that is feasible on m unit-speed machines can be scheduled on m machines of speed $4/3 + \epsilon$ For global scheduling, the EDF-VD scheduling policy is combined with the fpEDF scheduling policy, designed for non-MC task systems. For the resulting global scheduling algorithm it is proven [3] that any 2-level implicit-deadline MC task system that is feasible on m unit-speed machines, can be scheduled on m machines running at speed $\sqrt{5} + 1$.

Open problems

- Extending results to more than two criticality levels. For a single processor, schedulability conditions for EDF-VD are known [1] and the questions is how these can be incorporated into a partitioned or global scheduling algorithm for multiple processors.

- Extending results to different processor models, for instance unrelated machines. In [5], non-MC task systems are scheduled on unrelated machines with a speedup $8+2\sqrt{6} \approx 12.9$, via smart rounding of an integer linear program. An interesting question is if the ILP and the corresponding rounding procedure can be adjusted to accommodate schedulability conditions for MC task systems.

References

1 S. Baruah, V. Bonifaci, G. D'Angelo, A. Marchetti-Spaccamela, S. van der Ster, and L. Stougie. Mixed-criticality scheduling of sporadic task systems. *In Proceedings of 19th Annual European Symposium on Algorithms (ESA)*, pp. 555–566, 2011.

2 S. Baruah, V. Bonifaci, G. D'Angelo, H. Li, and A. Marchetti-Spaccamela. The Preemptive Uniprocessor Scheduling of Mixed-Criticality Implicit-Deadline Sporadic Task Systems. In *ECRTS*, 2012.

3 S. Baruah, B. Chattopadhyay, H. Li, and I. Shin. Mixed-criticality scheduling on multiprocessors. *Real-Time Systems* 50, 142–177, 2014.

4 C. Chekuri and S. Khanna. On multidimensional packing problems. *SIAM Journal on Computing* 33(4) 837–851, 2004.

5 A. Marchetti-Spaccamela, C. Rutten, S. van der Ster, and A. Wiese. Assigning sporadic tasks to unrelated machines. *Mathematical Programming.* DOI: 10.1007/s10107-014-0786-9

5 Working Groups

5.1 Report on Platforms and Experimental Evaluation

Robert I. Davis

Present: Sébastien Faucou, Leandro Indrusiak, Chris Gill, Gabe Parmer, Roman Obermaisser, Sebastian Stiller, Cristian Maxim, Jim Anderson, Albert Chen, Sophie Quinton, David Broman, Kai Lampka, Lothar Thiele

Benchmarks and workloads

Workloads

- Fudge factors relating measurements to execution time budget: Typically 20 to 50% for singlecore systems. Does this also make sense in multicore?
- How much bigger can C(HI) be than a 'well' measured C(LO) (that perhaps accounts for the paths through the code, but not variations due to HW)? Could we perhaps get an upper bound by turning the cache off?
- What type of systems offers a representative workload for MCS? Are UAVs a good candidate?

Benchmarks, WATERS workshop and Call to Action For RT Benchmarks

- Complaint: we need industrial benchmarks to design solutions to problems that would be of benefit to the industry. Using existing real code, even if it is not true level A code? Source code is fine, but should we also have benchmarks in the form of more abstract models.
- Do Mälardalen Benchmarks cover all the case-studies that we want? Is it possible to build realistic task sets from Mälardalen Benchmarks? Should we set up a set of different representative applications from the Mälardalen Benchmarks representative of cache access and memory footprint?

- A large goal here is to collect artifacts that are usable for experimental purposes by the community.
- If this is not possible, then perhaps we can create a set of these that might not functionally be interesting, but that maintain the interesting characteristics in terms of time/cache utilization/etc.
- What is the set of non-functional behaviors we care about? The top three are the cache usage, memory access patterns, and timing. Additional behaviors that would be nice down the line are synchronization/dependencies/system interactions.
- We want the benchmarks to be open and free.
- We need executable benchmarks: we want code that can be functionally irrelevant but which has realistic execution times, memory accesses, cache policies and ideally environment. Best case is that we have applications from industry. What about developping an obfuscation strategy? If this isn't possible, then we need a set of benchmarks that we can use to compare against each other, and seek industry blessing or modification afterwards.
- Papabench is a benchmark for the task models. Can we have a benchmark suite based on generating task models?
- Another idea, if we want more complicated tasks, perhaps we can can run a few of the Mälardalen benchmarks composed sequentially to make at least temporally more interesting tasks. This might not be reasonable, but it might be reasonable to go to industry and get feedback on what we *should* do. Of course, this will not work for cache footprints.
- How can we generate task models for MC? Vestal's original paper seemed to have the WCET "fudge factors" between *around* 20% to 50%. Importantly, there are concrete examples in his paper, so we should heed those.

What do we need to do as a community?

- A call for benchmarks/artifacts/code/task models from the community. We can take this to industry and get their feedback. See the call in http://waters2015.inria.fr/, though the call for benchmarks should be community-wide and go beyond this venue.
- Should we have a MCBench workshop devoted to creating this benchmarking suite? Or should we fold this into an existing workshop like WMC?
- We want exemplars of different application scenarios. These are the end-to-end suites of software you'd see running on a real system. For example, think the collection of software required to run UAVs. These should be emphasized in any call for benchmarks.

Other Questions

- Is complete isolation needed (or even possible) between criticality levels? Answer: No.
- Criticality level similar to memory hierarchy (by going down a level, you have less confidence but more tasks/work/utilization)?
- Tackling the whole complexity on a simple platform is too difficult today?

Links

- TACLeBench: http://tacle.knossosnet.gr/activities/taclebench
- Debie: http://www.irit.fr/wiki/doku.php?id=wtc:benchmarks:debie1
- Mälardalen Benchmarks: http://www.mrtc.mdh.se/projects/wcet/benchmarks.html
- Papabench: http://www.irit.fr/recherches/ARCHI/MARCH/rubrique.php3?id_rubrique=97

5.2 Report on WCET

Claire Maiza

Present: David Broman, Bjorn Lisper, Pontus Ekberg , Claire Maiza, Christine Rochange, Suzanne van der Ster, Liliana Cucu-grosjean, Jan Reineke, Pascal Richard, Sebastian Altmeyer.

In this subgroup, the idea was to discuss about worst-case execution time in the context of mixed-criticality. In the context of mixed-criticality systems, timing models at the scheduling phasis consider not only one guaranteed bound, but a set of execution time estimations. In this summary we first discuss where these different estimations come from, second we focus on mixed-criticality in multi-core systems and the specificities due to the timing interferences.

How to get different execution time estimations?

Note that as far as more than one estimation is considered, one can not name them "worst-case execution time". The notion of an estimation which is supposed to be closer to the real execution time but not an upper-bound on all possible execution time is clearly not a "worst-case".

We identified some sources of different execution time estimations:

- Due to the environment:
  Using a static analysis, one usually look for a bound on the execution time for a specific "execution context". The precision of this context may influence the execution time estimation. For instance, in automotive functionality may be developped for a large set of cars. However, once deployed, the specificities of the car in which the functionality is implemented could lead to a preciser estimation of the execution time.
- Due to the use of a margin:
  In some companies, the WCET is measured or estimated and a large margin (e.g., a factor of 100) is applied to get an upper-bound on the execution time. In this case, the upper-bound is largely over-estimated, but may give the feeling of a more trustful bound...
- Due to the WCET analysis:
  Timing analysis are based on three models: hardware, software and environement. Different tools or analysis method could get different estimations due to the precision of these models and/or the uncertainty involved. For instance, a measurement-based timing analysis considers a subset of the hardware model states.

Multi-core context

In case of multi-core, the large set of possible interferences of one task execution on the execution of other ones, leads to a more complex notion of execution time estimation. Due to the complexity of an exhaustive analysis that would take into account all possible interferences, there is a usual tradeof between precision of the estimation and complexity of the analysis.

Some approaches try to get more precision by adapting the architecture. These approaches may try to get a multi-core platform that suffers less from interferences (predictabe architecture) or to configure the architecture to get less interferences (e.g., partition). In this context, mixed-criticality is less an issue because the tasks with low-criticality should not influence the execution time of the high-criticality tasks.

When the platform is not designed to be predictable, execution time analysis may lead to a large set of different estimations. For instance, a bus analysis may consider a very large guaranteed bound on the interferences or model precisely all possible accesses to the bus. In the first case, the bound should be over-estimated. In the second case, the complexity of the analysis might not scale real application size. That may be a reason for the need of different execution time bounds in the scheduling analysis. This lead to two open-questions: should WCET and scheduling analysis be one common analysis in the case of multi-core? Should the uncertainty in the multi-core hardware model lead to a new execution time analysis method?

5.3 Report on Criticality

Sanjoy K. Baruah

Present: Zoë Stephenson, Vincent Nelis, Joël Goossens, Sophie Quinton, Leen Stougie, Dorin Maxim, Alberto Marchetti-Spaccamela, Enrico Bini, Wang Yi, Marko Bertogna, Nathan Fisher, Gerhard Fohler, Emmanuel Grolleau, Zhishan Guo, Pengcheng Huang, Sanjoy K. Baruah.

Agenda

This subgroup was spawned off with a mandate to explore an agenda that includes the following issues

- Obtain a better understanding of the safety background that motivates consideration of criticality levels. Why do we even have criticality levels, what need do they address?
- Identify the role that the WCET concept plays in safety considerations in mixed-criticality systems. We should distinguish between WCET budgets used for runtime enforcement and mode changes, and WCET estimates which approximate the WCET with different levels of confidence. Distinguish between budgets and estimates.
- The above issues may help determine what characterizes a system as being a mixed-criticality one. Is it about where WCET values are usable or is it about having isolation/lack of interference or both?

Discussions

1. Notions of criticality, as used in the research community, come from safety standards, e.g., IEC61508 and ISO26262. However, the use of some of the criticality-related terminology in the mixed-criticality systems (MCS) research community is not always consistent with their use in the standards (see Figure 7). It is incumbent on the research community to make an effort to familiarize practitioners with their research findings. Some possible avenues for achieving this were discussed:

- Issue is maybe of widening the scope of who gets involved with this work.
- We should speak of graceful degradation and fault tolerance rather than changing criticality.
- Mixed criticality in industry is currently mainly about isolation and separation; a significant portion of the research efforts are aimed at ensuring more efficient utilization of computational resources.

Where do criticality levels come from?

- In several application domains, criticality levels they are defined by standards
- The research community could think that criticality levels in MCS are related to those standards

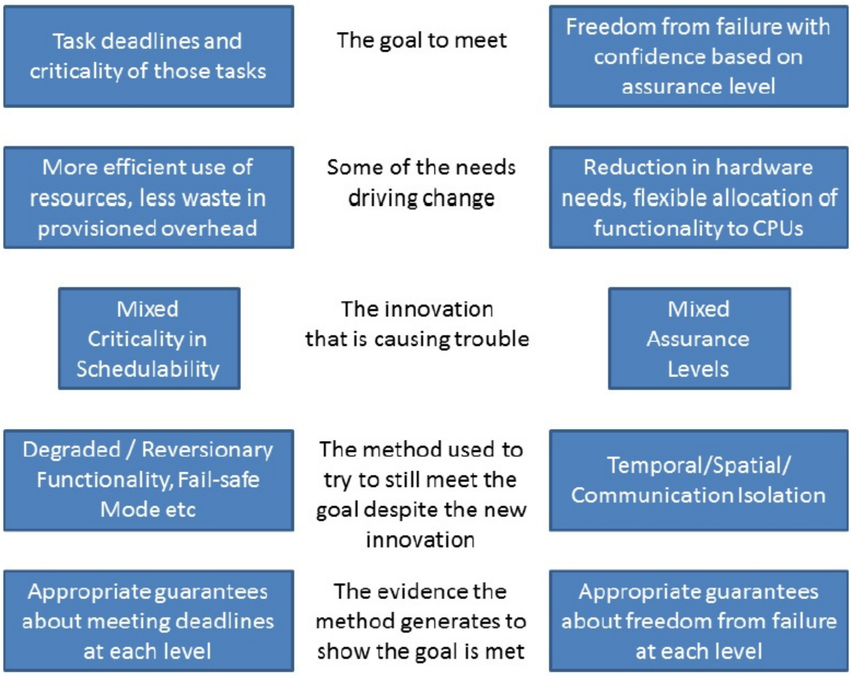

Task deadlines and criticality of those tasks
The goal to meet
Freedom from failure with confidence based on assurance level
More efficient use of resources, less waste in provisioned overhead
Some of the needs driving change
Reduction in hardware needs, flexible allocation of functionality to CPUs
Mixed Criticality in Schedulability
The innovation that is causing trouble
Mixed Assurance Levels
Degraded / Reversionary Functionality, Fail-safe Mode etc
The method used to try to still meet the goal despite the new innovation
Temporal/Spatial/ Communication Isolation
Appropriate guarantees about meeting deadlines at each level
The evidence the method generates to show the goal is met
Appropriate guarantees about freedom from failure at each level

Figure 7 Mapping Research Concepts to Industrial Concerns.

2. ALARP – "As Low As Reasonably Practicable" – is a widely-adopted guiding principle in safety analysis for evaluating success in risk reduction. It is not expected that risk can be reduced to zero; nor it is desirable (cost-effective) to over-engineer for no tangible benefit. Neither is it desirable to miss out some risk area from analysis and mitigation.

It is important to be aware of these distinctions:

- *Safety* relates to inadvertent harm that a system can do. It is sufficiently safe if the risk of causing a hazard is reduced as low as reasonably practical.
- *Security* relates to intentional violation of access control – exposure of data through overt and covert channels, for example.
- *Surety* is not a term that is often used, it relates to having 100% confidence or 0% residual risk. Since there is always risk in the environment and in hardware failures, this induces a limit on the level of risk reduction for software that will ever be acceptable in practice. However, it is still important to reduce uncertainty in what the risks even are.

3. Industry often uses an isolation/ separation based approach to partition software of different assurance levels so that it can be known that there is sufficient freedom from interference with sufficient confidence. In order to justify dropping this approach, there would have to be a good reason to suffer the pain of arguing about why there is still sufficient freedom from interference with sufficient confidence. What would the gains be? – flexibility? Would it be possible to use current hardware for longer? Would it be feasible to reduce the confidence level with which one has to assess some kinds of interference?

4. Arguments were made in favor of the mixed-criticality approach advocated in the MCS research community vs. an isolation-based approach:

- Today there is a gap between actual and worst-case execution times requiring, especially in the case of isolation-based approaches, significant over-provisioning the computing resources.

- We can expect future architecture to increase this gap, will it be increased to the point of being unbearable?
- If so, is the cost of loosening isolation worth the gain of computing resource utilization?
- Other gains of mixed-criticality: dealing with different cases of uncertainty (not only WCET, but also periods, thermal aspects, etc.)

5. When we try to reason about uncertainty we need to be clear about how the standards relate aleatory uncertainty (e.g. MTBF of a hardware component) and epistemic uncertainty (e.g. I'm not 100% sure I got enough coverage in my testing).

6. MCS research and the certification process. Currently, correct by construction is the common way to demonstrate correctness for the purposes of obtaining certification. Evidence can be provided by analysis, but it is challenging to make this acceptable to certification authorities. The question was discussed: Can MCS research be used in certification? The following points were made:

- any new theory takes time to be accepted
- perhaps we should be working on developing a theory that is ready to be applied whenever industry is ready
- There was a discussion about how mixed-criticality is applicable or could be in the future in the industry: in mixed-criticality systems research there is room for every aspect: theory, operating systems, practical research more certification standard oriented, etc.

5.4 Report on Probabilistic Approaches

Liliana Cucu-Grosjean

Present: Arvind Easwaran, Zhishan Guo , Adriana Gogonel, Dorin Maxim, Sebastian Altmeyer, Yasmina Abdeddaim, Rob I. Davis, Liliana Cucu-Grosjean.

The discussions on probabilistic approaches took place during two time slots:

1. Following the presentations on probabilistic approaches, the first slot of discussions within this group has been dedicated to the application of Extreme Value Theory (EVT). This theory is used to solve the problem of estimating a probabilistic bound on all possible execution times, this bound is usually denoted by pWCET.
   We underline four different threads of discussions related to the utilization of pWCET to estimate WCET in the context of mixed criticality systems. Each thread had identified one or several open problems detailed below.
   - Currently the static analysis is extensively used to estimate the WCET. The users of static analysis need the understand the assumptions of EVT in order to use it while WCET estimating, but also to compare against state of the art approaches.
   - An important effort of popularization is necessary in order to increase the understanding of the steps of EVT when applied to the problem of estimating the WCET.
   - Today the differences between functional independence, probabilistic independence and statistical independence are not well understood by the community and this has a direct impact on the overall understanding of this method.
   - Once a pWCET is estimated, how do we calculate the probability of more than one overrun of C(LO) in a given time? It is generally admitted that an overrun never appears alone and that it is usually related to other possible overruns.

2. The second slot of discussions has been concentrated on the understanding how a probability distribution of a WCET defines different criticalities ?

   Three different models have been identified as follows.
   - A first model that associates to each level of criticality a pair (value for the WCET, probability of appearance of that value). For instance in the Vestal model this could correspond to a random variable with three possible values $C(LO), C(HI)$ and ∞.
   - A second model that associates to each level of criticality a random variable describing the pWCET. For instance in the Vestal model this could correspond to $\mathcal{C}(LO)$[9] and $\mathcal{C}(HI)$ where $\mathcal{C}(HI) = \mathcal{C}(LO) + \mathit{constant}$.
   - A third model that associates to the highest level of criticality an unique WCET (that could be obtained using static analysis for instance) and to the lowest level of criticality a pWCET.

6 Open Problems

6.1 Unification of mixed criticalities, WCET, and probabilistic execution time

Enrico Bini (Scuola Superiore Sant'Anna, Pisa, IT)

- I have no experience with mixed-criticality systems
- I believe that some concepts we have been listening about
  - mixed-criticality
  - probabilistic exec time
  - mode change

  do overlap significantly
- This presentation is an attempt to relate them with each other
- It may well be something very obvious to you (especially timing analysis people).

Execution time

What does the sequence of job execution times[10] depend on?

- Let Ω be the sample space (input data, machine type, cache status, alpha particles flipping bits, etc.)
- $\omega \in \Omega$ is an event
- execution time is $c : \Omega \to \mathbb{R}$

- "Worst-case execution time"

$$C_{\mathsf{WCET}}(\Omega) = \sup_{\omega \in \Omega} c(\omega)$$

[9] We use calligraphic letters to denote random variables.

[10] next arguments are valid for any task parameter

Criticality

- Sanjoy: criticality is a desired "level of assurance"
- It seems that "criticality" are then just subsets of Ω

$$\mathsf{LO} \subseteq \mathsf{HI} \subseteq \Omega \tag{3}$$

- Then, for any criticality level $\mathcal{L} \subseteq \Omega$, the corresponding $\mathcal{L}$-WCET is

$$C_{\mathsf{WCET}}(\mathcal{L}) = \sup_{\omega \in \mathcal{L}} c(\omega)$$

- Notice that (3) implies

$$C_{\mathsf{WCET}}(\mathsf{HI}) \geq C_{\mathsf{WCET}}(\mathsf{LO})$$

- The partial ordering of set inclusion over Ω also induces a partial ordering of the criticalities

Property of criticality

One possible property of criticality:

- Let us have a chain of criticality levels

$$\mathcal{L}_1 \subseteq \mathcal{L}_2 \subseteq \ldots \subseteq \mathcal{L}_n$$

- Is an event $\omega' \in \mathcal{L}_{i+1} \setminus \mathcal{L}_i$ "worse" than any event $\omega \in \mathcal{L}_i$? This is reasonable to expect

▶ Property 1 (monotonicity of $c(\cdot)$ over crit). If an event $\omega' \in \mathcal{L}_{i+1} \setminus \mathcal{L}_i$ "worse" than any event $\omega \in \mathcal{L}_i$?, then

$$\forall i = 1, \ldots, n-1, \ \forall \omega' \in \mathcal{L}_{i+1} \setminus \mathcal{L}_i, \ \forall \omega \in \mathcal{L}_i, \quad c(\omega') \geq c(\omega)$$

Probability

- If Ω is equipped with a probability measure P, then $c : \Omega \to \mathbb{R}$ becomes a random variable
  - its cumulative distribution function ($\mathsf{cdf}(x)$) is

$$\mathsf{cdf}(x) = P(c \leq x) = P(\{\omega \in \Omega : c(\omega) \leq x\})$$

- $P(\mathsf{HI})$ is then the probability that the system belongs to the criticality $\mathsf{HI} \subseteq \Omega$;
- $P(\mathsf{HI}) \geq P(\mathsf{LO})$
- How is the measure P defined? I don't know
  - it has to do with the probability of the input values, probability of being in some status, etc.

Criticality & Probability

- Given
  - a probability measure P over Ω,
  - a criticality $\mathcal{L} \subseteq \Omega$ with $P(\mathcal{L}) \neq 0$, and
  - the computation time function $c : \Omega \to \mathbb{R}$
- we can define the *conditional probability* any event $A \subseteq \Omega$ given $\mathcal{L}$ as

$$P(A|\mathcal{L}) = \frac{P(A \cap \mathcal{L})}{P(\mathcal{L})}$$

- the conditional random variable $c : \Omega \to \mathbb{R}$ give $\mathcal{L}$, has

$$\mathsf{cdf}(x) = \frac{\{c(w) \leq x\} \cap \mathcal{L}}{P(\mathcal{L})}$$

  Average execution time, with criticality $\mathcal{L}$

- the average execution time is

$$C_{\mathsf{avg}}(\mathcal{L}) = E[c \,|\, \mathcal{L}] = \frac{1}{P(\mathcal{L})} \int_{\mathcal{L}} c(\omega)\, dP(\omega)$$

- nice property (maybe proved on the blackboard) is:
  given $\mathsf{LO} \subseteq \mathsf{HI} \subseteq \Omega$, then

$$C_{\mathsf{avg}}(\mathsf{HI}) < C_{\mathsf{avg}}(\mathsf{LO}) \quad \Leftrightarrow \quad C_{\mathsf{avg}}(\mathsf{HI} \setminus \mathsf{LO}) < C_{\mathsf{avg}}(\mathsf{LO})$$

- however of Property “monotonicity over crit” holds, then

$$\forall\, \mathsf{LO} \subseteq \mathsf{HI} \subseteq \Omega, \quad C_{\mathsf{avg}}(\mathsf{HI}) \geq C_{\mathsf{avg}}(\mathsf{LO})$$

7 New collaborations

7.1 Providing Weakly-Hard Guarantees for Mixed-Criticality Systems

Robert I. Davis (Real-Time Systems Research Group, Department of Computer Science, University of York, UK and AOSTE team, Inria Paris-Rocquencourt, FR)
Sophie Quinton (SPADES team, Inria Grenoble – Rhône-Alpes, FR)

Mixed Criticality Systems are systems running applications of different criticality levels [1]. Often only two criticality levels are considered, denoted LO-criticality and HI-criticality respectively. According to the definition most widely accepted by the research community, usually called the Vestal model, tasks are expected to run in normal mode as specified by their LO-criticality model (which is based on somewhat optimistic parameters) so that all task requirements are satisfied. In addition, one must consider the possibility for tasks to run out of the bounds defined by their LO-criticality parameters in degraded mode, following their HI-criticality model. In that case requirements for the HI-criticality tasks must remain satisfied but requirements for the LO-criticality tasks are dropped.

One criticism that is often made of this approach is that it is not realistic to consider that LO-criticality tasks may be dropped, even in a context where the safety of HI-criticality tasks may be at risk. We are interested here in how weakly-hard guarantees [2] (i.e. having to meet m out of k deadlines rather than all of them) can be used to avoid dropping LO-criticality tasks entirely. The simplest scenario that can be envisioned is that LO-criticality tasks have to meet all deadlines in normal mode, but have weakly-hard constraints in degraded mode, while HI-criticality tasks have to meet all deadlines (i.e. hard constraints) in both modes. The rationale behind this is that control algorithms can often tolerate some jobs missing their deadlines or not executing, but then need to guarantee that a number of jobs will meet their deadlines so that the system returns to a stable state [4]. A key consequence of using weakly-hard constraints is that this may allow postponement of the change in scheduling policy resulting from a switch from normal to degraded mode.

We believe that introducing weakly-hard constraints into the mixed criticality model might help increase the acceptance of the latter in industry. Note that various other scenarios are interesting as well. For example we could consider that HI-criticality tasks have hard deadlines while LO-criticality tasks always have weakly-hard constraints (maybe weaker ones in degraded mode). Alternatively, all tasks could have weakly-hard constraints in both modes. Again in that case weakly-hard constraints may allow postponement of the change in scheduling policy: a HI-criticality task could be aborted rather than exceed its LO-criticality execution time.

We aim to collaborate on research integrating the concept of weakly-hard constraints into Mixed Criticality Systems. In the first instance, we will explore how these constraints can be incorporated into the Adaptive Mixed Criticality scheduling policy and analysis proposed by Baruah et al. [1].

References

1 Sanjoy K Baruah, Alan Burns, and Robert I Davis. Response-time analysis for mixed criticality systems. In *Real-Time Systems Symposium (RTSS), 2011 IEEE 32nd*, pages 34–43. IEEE, 2011.

2 Guillem Bernat, Alan Burns, and Albert Llamosí. Weakly hard real-time systems. *IEEE Trans. Computers*, 50(4):308–321, 2001.

3 Alan Burns and Robert I. Davis. Mixed criticality systems – a review. http://www-users.cs.york.ac.uk/burns/review.pdf.

4 Goran Frehse, Arne Hamann, Sophie Quinton, and Matthias Woehrle. Formal analysis of timing effects on closed-loop properties of control software. In *Proceedings of the IEEE 35th IEEE Real-Time Systems Symposium, RTSS 2014, Rome, Italy, December 2-5, 2014*, pages 53–62, 2014.

7.2 A Multicore Response Time Analysis Framework

Sebastian Altmeyer (University of Amsterdam, NL)
Robert I. Davis (Real-Time Systems Research Group, Department of Computer Science, University of York, UK and AOSTE team, Inria Paris-Rocquencourt, France)
Leandro Indrusiak (University of York, GB)
Claire Maiza (VERIMAG – Gières, FR)
Vincent Nelis (The Polytechnic Institute of Porto, PT)
Jan Reineke (Universität des Saarlandes, DE)

In this paper, we introduce a Multicore Response Time Analysis (MRTA) framework. This framework is extensible to different multicore architectures, with various types and arrangements of local memory, and different arbitration policies for the common interconnects. We instantiate the framework for single level local data and instruction memories (cache or scratchpads), for a variety of memory bus arbitration policies, including: Round-Robin, FIFO, Fixed Priority, Processor Priority, and TDMA, and account for DRAM refreshes. The MRTA framework provides a general approach to timing verification for multicore systems that is parametric in the hardware configuration and so can be used at the architectural design stage to compare the guaranteed levels of performance that can be obtained with different hardware configurations. The MRTA framework decouples response time analysis from a reliance on context independent WCET values. Instead the analysis formulates response times directly from the demands on different hardware resources.

7.3 Mixed criticality support for automotive embedded systems

Yasmina Abdeddaim (Université Paris-Est, LIGM UMR CNRS 8049, ESIEE Paris, FR)
Sébastien Faucou (University of Nantes, FR)
Emmanuel Grolleau (ENSMA – Chasseneuil, FR)

On the subject of probabilistic analysis on mixed criticality systems when some criticality levels have deterministic constrains and parameters descriptions, while other criticality levels allow for a certain probability of failure and hence can be modeled and analyzed probabilistically.

Participants

- Yasmina Abdeddaim
ESIEE – Noisy le Grand, FR
- Sebastian Altmeyer
University of Amsterdam, NL
- James H. Anderson
University of North Carolina – Chapel Hill, US
- Sanjoy K. Baruah
University of North Carolina – Chapel Hill, US
- Marko Bertogna
University of Modena, IT
- Enrico Bini
Scuola Superiore Sant'Anna – Pisa, IT
- Björn B. Brandenburg
MPI-SWS – Kaiserslautern, DE
- David Broman
KTH Royal Institute of Technology, SE
- Alan Burns
University of York, GB
- Albert Cohen
ENS – Paris, FR
- Liliana Cucu-Grosjean
INRIA – Le Chesnay, FR
- Robert I. Davis
University of York, GB
- Arvind Easwaran
Nanyang TU – Singapore, SG
- Pontus Ekberg
Uppsala University, SE
- Rolf Ernst
TU Braunschweig, DE
- Sébastien Faucou
University of Nantes, FR
- Nathan Fisher
Wayne State University, US
- Gerhard Fohler
TU Kaiserslautern, DE
- Christopher D. Gill
Washington University – St. Louis, US
- Adriana Gogonel
INRIA – Le Chesnay, FR
- Joel Goossens
Free University of Brussels, BE
- Emmanuel Grolleau
ENSMA – Chasseneuil, FR
- Zhishan Guo
University of North Carolina – Chapel Hill, US
- Pengcheng Huang
ETH Zürich, CH
- Leandro Soares Indrusiak
University of York, GB
- Kai Lampka
Uppsala University, SE
- Björn Lisper
Mälardalen University – Västeras, SE
- Claire Maiza
VERIMAG – Giè res, FR
- Alberto Marchetti-Spaccamela
University of Rome
"La Sapienza" IT
- Cristian Maxim
Airbus S.A.S. – Toulouse, FR
- Dorin Maxim
The Polytechnic Institute of Porto, PT
- Vincent Nelis
The Polytechnic Institute of Porto, PT
- Roman Obermaisser
Universität Siegen, DE
- Gabriel Parmer
George Washington University – Washington, US
- Sophie Quinton
INRIA - -Grenoble, FR
- Jan Reineke
Universität des Saarlandes, DE
- Pascal Richard
ENSMA – Chasseneuil, FR
- Christine Rochange
Paul Sabatier University – Toulouse, FR
- Zoe Stephenson
Rapita Systems Ltd. – York, GB
- Sebastian Stiller
TU Berlin, DE
- Leen Stougie
CWI – Amsterdam, NL
- Lothar Thiele
ETH Zürich, CH
- Suzanne van der Ster
VU University of Amsterdam, NL
- Wang Yi
Uppsala University, SE

Report from Dagstuhl Seminar 15122

Formal Models of Graph Transformation in Natural Language Processing

Edited by

Frank Drewes[1], Kevin Knight[2], and Marco Kuhlmann[3]

1 **Umeå University, SE,** drewes@cs.umu.se
2 **University of Southern California, US,** knight@isi.edu
3 **Linköping University, SE,** marco.kuhlmann@liu.se

Abstract

In natural language processing (NLP) there is an increasing interest in formal models for processing *graphs* rather than more restricted structures such as strings or trees. Such models of graph transformation have previously been studied and applied in various other areas of computer science, including formal language theory, term rewriting, theory and implementation of programming languages, concurrent processes, and software engineering. However, few researchers from NLP are familiar with this work, and at the same time, few researchers from the theory of graph transformation are aware of the specific desiderata, possibilities and challenges that one faces when applying the theory of graph transformation to NLP problems. The Dagstuhl Seminar 15122 "Formal Models of Graph Transformation in Natural Language Processing" brought researchers from the two areas together. It initiated an interdisciplinary exchange about existing work, open problems, and interesting applications.

Seminar March 15–20, 2015 – http://www.dagstuhl.de/15122
1998 ACM Subject Classification F.4.2 Grammars and Other Rewriting Systems, F.4.3 Formal Languages, G.2.2 Graph Theory, I.2.7 Natural Language Processing
Keywords and phrases Automata theory, Graph transformation, Natural language processing
Digital Object Identifier 10.4230/DagRep.5.3.143
Edited in cooperation with Petter Ericson (pettter@cs.umu.se)

1 Executive Summary

Frank Drewes
Kevin Knight
Marco Kuhlmann

Strings are fundamental data structures in natural language processing (NLP). Weighted finite-state string acceptors and transducers, first introduced as theoretical constructs, have proven their worth in speech recognition, part-of-speech tagging, transliteration, and many other applications. The string automaton framework provides efficient generic algorithms for composition, bidirectional application, k-best extraction, determinization, minimization, parameter tuning, etc. These algorithms have been packaged in software toolkits that form the core of many state-of-the-art systems.

Tree automata go further in permitting large-scale, syntactically-motivated re-ordering of subtrees. They were originally devised to help formalize Chomsky's linguistic theories, but their subsequent development was largely disconnected from NLP practice. In 2005, tree automata theorists and machine translation (MT) practitioners began working together to come up with a new kind of statistical MT system based on tree automata. This led to some of the best practical results in common evaluations of MT quality, and syntactic methods are now used in industrial MT systems. This work at the intersection of tree automata and NLP created vibrant new research directions for both areas.

Nowadays, *graphs* are becoming an even more general fundamental data structure in practical NLP. Classic feature structures can be seen as rooted, directed, edge- and leaf-labeled graphs. Recent work in dependency parsing produces graphs rather than trees. New work in deep semantic annotation organizes logical meanings into directed graph structures, and several efforts are now being made that in the near future will yield large amounts of linguistic data annotated with these representations. Formal models of *graph transformation* are therefore of fundamental importance for the development of practical systems for these tasks. The situation is familiar: there exists a formal theory of graph transformation, but this theory is largely disconnected from research and practice in NLP.

The theory of graph transformation studies rule-based mechanisms for the manipulation of graphs. A particularly well-studied subject within the area of graph transformation, and one that has received quite some attention recently within the NLP community, are *context-free graph grammars*. These grammars have many nice properties in common with context-free phrase structure grammars, but are considerably more powerful and versatile; in particular, they can be used to generate context-sensitive string languages (when strings are represented as chain graphs). The price of this expressiveness is a higher computational complexity; in particular, there are context-free graph languages for which parsing is NP-complete. This has triggered research on specialized, more efficient algorithms for restricted classes of graphs. A well-known result in this area is that many in general intractable problems on graphs become solvable in polynomial time when restricted to graphs of bounded tree-width.

With the number of interesting applications and the amount of available data quickly increasing, there is a clear need for the NLP community to acquire knowledge about formal models of graph processing, as such models can greatly simplify practical systems, by providing a uniform knowledge representation and efficient, generic algorithms for inference. Unfortunately, most NLP researchers are unaware of the rich literature on graph transformation, and even those who are find it hard to connect it to their own work. Conversely, few researchers in graph transformation are aware of the new applications of their research within natural language processing, the characteristic properties of the available data, the specific desiderata of these applications, and the research problems that are posed by them.

The overall goal of the seminar was to bring the various research communities together to assess the state of the art, identify areas of common interest, and pave the way for future collaborations. We think that this goal was reached to a very high degree, which will be a major factor in the creation of a new interdisciplinary research community.

Organization of the Seminar

The seminar was attended by 29 participants from 9 countries in North America, Europe, and Africa. It was held from March 15 to March 20, 2015. Since the intention was to foster the creation of a new research community, it was decided to organize the seminar in the form of a self-organized workshop with many informal discussion meetings on topics suggested by the participants themselves. For this, the seminar roughly followed the idea of Open Space Technology. This worked very well and gave rise to many insightful discussions. (See Section 5 for the list of topics discussed.)

2 Table of Contents

Executive Summary

Overview of Talks

Reports from Working Groups

3 Overview of Talks

3.1 Tutorial: Introduction to Graph Transformation

Frank Drewes (University of Umeå, SE)

URL http://www8.cs.umu.se/~drewes/gratra_tutorial.pdf

The theory of graph transformation studies formal rule-based models for the manipulation of graphs. In particular, this includes graph grammars that generate graph languages and graph rewrite systems that turn input graphs into output graphs. The tutorial gave an introduction to some of the most well-studied aspects of graph transformation, focusing on those which seem to be of particular interest for NLP.

General Graph Transformation Systems

These are graph transformation systems which are Turing complete. They usually consist of finitely many rules (possibly enhanced by control structures or application conditions) that replace a subgraph (the left-hand side) in a host graph by another graph (the right-hand side). The most well-known approaches are the single and double pushout approaches, which belong to the so-called algebraic approaches.

Context-Free Graph Grammars

Context-free graph grammars are grammars based on rules that either replace single nodes or single (hyper)edges by other subgraphs. This makes them context-free and results in many desirable properties, e.g., there exist algorithms for various tasks.

Parsing Hyperedge Replacement Languages

One of the most important algorithmic tasks in connection with context-free graph grammars, such as hyperedge replacement grammars, is parsing. There are easy and very versatile proofs showing that this problem is NP-hard even in the non-uniform case, i.e., where the grammar is fixed. In other words, there are NP-complete hyperedge replacement languages. However, in special cases polynomial-time parsing is known to be possible.

Monadic Second-Order Logic

There are well-known and very useful connections between monadic second-order logic on strings or trees on the one hand, and regular string and tree languages on the other hand. Similarly useful connections relate monadic second-order logic on graphs with context-free graph languages. For example, the restriction of a context-free graph language by a logical sentence is again context-free, and it can be decided whether all graphs/finitely many graphs/no graphs of a given context-free graph language satisfy a given sentence.

Term Graphs

Acyclic directed graphs can represent terms with sharing, so-called term graphs. This has been used in order to implement functional programming languages efficiently. Since directed acyclic graphs are important in meaning representation, the techniques and results of term graph rewriting may turn out to be useful in NLP.

4 Reports from Working Groups

4.1 Convolution Kernels for Graphs

Giorgio Satta

In machine learning, *kernels* are a class of functions that measure the "similarity" between two objects or structures. Many algorithms implementing kernel functions use an underlying feature vector representation for the input structures in a space with very high or even infinite dimension, but only implicitly represent this feature space. The advantage is that kernel algorithms avoid explicit computation of the feature map in such large space, and are thus much more efficient than direct algorithms.

Convolution kernels use a decomposition of the input structures into overlapping substructures or patterns, and are based on the computation of a census function for these substructures. In natural language processing, convolution kernel functions have been introduced for strings and trees. This group has been exploring existing convolution kernel methods for graph structures, in view of their potential application to semantic analysis of natural language based on directed acyclic graph structures.

Several convolution kernels for graphs have been proposed in the literature, based on simple path or tree-like substructures; see [1] and references therein.

This group has been exploring the idea of using so-called elastic kernels for directed acyclic graphs. The idea is borrowed from tree-based kernels as used in syntactic parsing (Alessandro Moschitti, personal communication). In an elastic kernel, substructure matching is allowed by stretching an arc of the pattern graph over a path of several arcs in the graph under analysis. This allows "jumping" over portions of the semantic representation that might correspond to some adjunct or modifier. Linguistic relevance of this idea has been discussed, without reaching a consensus among the participants on its effectiveness in semantic representations.

This group has also looked into existing similarity measures for Abstract Meaning Representation (AMR) such as smatch [2]. It has been observed that the smatch similarity measure does not satisfy the standard kernel function conditions and therefore can not be considered a well-formed kernel. This is so because convolution kernels make use of summations over substructure matching, while smatch make use of a max operator over the same counts.

References

1 Nino Shervashidze and Karsten M. Borgwardt. Fast subtree kernels on graphs. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conf. on Neural Information Processing Systems 2009. Proc. of a meeting held 7–10 December 2009, Vancouver, British Columbia, Canada*, pp. 1660–1668. Curran Associates, Inc., 2009.

2 Shu Cai and Kevin Knight. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

4.2 Efficient HRG-Parsing for the NLP Domain

Christoph Teichmann and Frank Drewes

Hyperedge Replacement Grammars (HRGs) are one of the candidate formalisms for the generation and parsing of graph based semantic representations [1]. In general, HRGs can generate NP-complete languages [2, 3]. Thus, one cannot hope to develop efficient parsing algorithms that work for all HRGs.

While it has been known for a long time that parsing based on HRGs can theoretically be implemented efficiently if the graphs that are being processed exhibit certain properties [4], researchers are currently trying to solve many of the details of actual implementation. Parsing algorithms often require a number of very complex design decisions in order to be efficient.

One approach to more efficient parsing has been the proposal to make use of tree decompositions of rules and input graphs, and work on representations based on the "boundary" of subgraphs that have already been processed [5]. This approach is asymptotically more efficient than storing copies of complete subgraphs and can be used to establish upper bounds on the number of items that need to be considered. On the comparatively small graphs that are used as input data in natural language processing, however, it may actually be more efficient just to store the complete subgraphs.

Once the questions of representation and look-up have been settled, it is then possible to design graph parsers in a way that is very similar to well known approaches in string parsing for natural language processing.

Another approach that is currently being worked on is to generalize techniques known from compiler construction, based on restrictions that make string parsing more efficient than the usual $O(n^3)$ obtained by CKY or Earley parsing. A first approach in this direction is *predictive top-down parsing*, which extends SLL(1) string grammars to the HRG case. A predictively top-down (PTD) parsable HRG yields a quadratic, and in many cases linear parsing algorithm. However, the analysis of a grammar needed to establish PTD parsability is complicated and cannot usually be done by hand [6]. It is currently still unclear whether PTD parsable HRGs are suitable for NLP applications. It may be worthwhile to check whether predictive bottom-up parsing is possible as well, generalizing the well-known notion of LR(1) string parsing.

Another idea, which may be especially useful for parsing structures such as Abstract Meaning Representations, is to extend Lautemann's almost forgotten concept of componentwise derivations [4]. Roughly speaking, the intuition behind componentwise derivations is that, if a nonterminal generates a graph that consists of several connected components, then the derivations of the individual components are independent of each other. This corresponds well to the intuition that, if several modifiers are attached to a concept in an AMR, then the sub-DAGs corresponding to those modifiers can usually be generated independently.

References

1 Bevan Keeley Jones, Sharon Goldwater, and Mark Johnson. Modeling graph languages with grammars extracted via tree decompositions. In *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*, page 54–62, 2013.

2 I. J. Aalbersberg, A. Ehrenfeucht, and G. Rozenberg. On the membership problem for regular DNLC grammars. *Discrete Applied Mathematics*, 13:79–85, 1986.

3 Klaus-Jörn Lange and Emo Welzl. String grammars with disconnecting or a basic root of the difficulty in graph grammar parsing. *Discrete Applied Mathematics*, 16:17–30, 1987.

4 Clemens Lautemann. The complexity of graph languages generated by hyperedge replacemen. *Acta Informatica*, 27:399–421, 1990.

5 David Chiang, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, Bevan Jones, and Kevin Knight. Parsing graphs with hyperedge replacement grammars. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, page 924–932, 2013.

6 Frank Drewes, Berthold Hoffmann, and Mark Minas. Predictive top-down parsing for hyperedge replacement grammars. In F. Parisi-Presicce and B. Westfechtel, editors, *Proc. 8th Int'l Conf. on Graph Transformation (ICGT'15)*, Lecture Notes in Computer Science. Springer, 2015.

4.3 Grammarless Approaches

Joakim Nivre, Christoph Teichmann, and Giorgio Satta

The term *grammarless approaches* is used informally to describe parsing algorithms where no hard constraint is imposed on the search for a syntactic analysis of the input sentence. In other words, the parser never rejects any input sentence. Typically, in a grammarless approach the parser uses soft constraints (weights) to choose some syntactic analysis in a space that includes all candidate structures that are compatible with the input sentence. More specifically, some search process is carried out that applies soft constraints either globally or locally to optimize possible choices.

In dependency parsing a number of transition systems and weighting schemes have been proposed that can generate projective and/or non-projective dependency trees given an input sentence. It is possible to use these to efficiently generate analyses for input sentences, when they are paired with a classifier that selects one out of a list of possible transitions given the current state of the parsing process and the input data, or by finding maximum weight substructures. It seems natural to extend this approach to structures that do not obey the treeness constraint.

In this group we discussed transition systems and weighting schemes capable of selecting different graph structures and their comparative benefits. In this context it is important to consider the problem of learning the weight function that guides the selection of the final analysis and whether it is possible to efficiently reconstruct the operations that generated a graph. It was observed that there is a simple generalization of Covington's algorithm for dependency tree parsing [1] that can derive any graph in time quadratic in the length of the input sentence. This algorithm could possibly be restricted to interesting special cases in order to improve accuracy or efficiency or both.

Generation/Decomposition of Graphs with Page Number Bounded by Two

In graph theory, a book embedding of a graph is an embedding into a collection of half-planes, all having the same line as their boundary. The vertices of the graph are required to lie on this boundary line, and the edges are required to stay within a single half-plane. The *page number* of a graph is the smallest possible number of half-planes for any book embedding of the graph [2].

This group has focused on dependency graphs with page number of two, where the vertices of the graph lying on the book boundary line is fixed and specified by the input sentence. The group has then discussed the problem of computing the highest score dependency graph with page number of two given an input sentence. Note that here we are considering dependency structures that are graphs in the strict sense, that is, these structures are not tree-like graphs.

Discussion has focused on how to apply dynamic programming techniques to solve this problem efficiently, therefore showing that the problem is in PTIME. There are known dynamic programming algorithms for the case of dependency graphs with page number of one; see [3]. Shortly after the meeting in Dagstuhl, two participants to this discussion group came up with a proof that the problem at hand is NP-hard; see [4].

Grammarless Generation of Strings from DAGs

We also discussed the problem of generating strings from DAGs, which can be viewed as a kind of graph transduction problem. The question then becomes: is there a grammarless approach to transducing graphs? Informally, the answer seems to be yes, and it is helpful to first think about how grammarless parsing algorithms apply to strings and trees. In the string case, the set of outputs can be defined as the set of all trees over a set of input words. In some cases we define it more carefully as the set of projective trees, or the set of non-projective trees meeting particular criteria [5]. We can then use combinatorial optimization algorithms or transition systems to search over the set of output trees.

Before considering generation of strings from graphs, we can consider generation of strings from trees, which is a special case (this is sometimes called realization in the generation community). Suppose that we have a labeled input tree. Then the output might be defined as follows: the string obtained by any tree traversal and substitution operation on the nodes of the tree. In other words, at each node of the tree, we must visit the node and its children (recursively) once, and when we visit the node, we output a word. This can be thought of as linearization of an (unordered) dependency tree, which always results in a projective structure. It should be possible to produce non-projective structures by encoding them into the traversal [6]. To extend this idea to DAGs, it suffices to observe that some nodes can be visited more than once, but we always know the number of visits since we know the number of parents. Hence, we can split the (recursive) visit of a node into as many parts as there are parents, and execute each part in order as the node is visited from its parents. Algorithmically, this could be accomplished many ways: by a transition system that outputs words as it visits nodes, or a global or local model that predicts the visit order for each node.

References

1 Michael A Covington. A fundamental algorithm for dependency parsing. In *Proceedings of the 39th annual ACM southeast conference*, pages 95–102. Citeseer, 2001.

2 Frank Bernhart and Paul C Kainen. The book thickness of a graph. *Journal of Combinatorial Theory, Series B*, 27(3):320–331, 1979.

3 Marco Kuhlmann. Tabulation of noncrossing acyclic digraphs. CoRR abs/1504.04993, Linköping University, 2015.

4 Peter Jonsson and Marco Kuhlmann. Maximum pagenumber-k subgraph is NP-complete. CoRR abs/1504.05908, Linköping University, 2015.

5 Emily Pitler, Sampath Kannan, and Mitchell Marcus. Finding optimal 1-endpoint-crossing trees. *Transactions of the Association for Computational Linguistics*, 1:13–24, 2013.

6 Marco Kuhlmann and Matthias Möhl. Mildly context-sensitive dependency languages. In *Proc. Association for Computational Linguistics*, pages 160–167, 2007.

4.4 HRG-Grammar Induction for NLP

Christoph Teichmann

Recently there has been interest in generating graph based semantic representations for input sentences and/or generating sentences that correspond to the meaning represented by a graph. This task can be solved by using synchronous grammars that generate a string and a graph in parallel. One can then parse an input with one side of the synchronous grammar and use any result that would have been generated by to other side as output. This leads naturally to the problem of inferring these synchronous grammars from input data. Since we will usually only have access to sentences and graphs and not to any information about an underlying grammar, we are forced to consider a potentially large set of parallel derivation steps and then extract a – preferably small – grammar. We assume that the sentence has been generated by some context-free derivation tree and that the graph was generated by a hyperedge replacement derivation. Unfortunately there are a large number of many potential pairings of sentence and graph parses, even when they are only considered in some packed representation. Therefore it is necessary to employ alignments that are the result of some simpler pre-processing step.

4.5 Probabilities for DAG Automata and Other Non-Context-Free Graph Rewriting Systems

Adam Lopez

The last few decades of work in natural language processing have confirmed that probabilistic systems learned from data are indispensable. So graph rewriting systems used for natural language processing must be probabilistic. Consider semantics-based machine translation, in which the goal is to explicitly convert a source string to its semantic representation, and then convert this representation to a target string, as in the following example from Jones et al. [1].

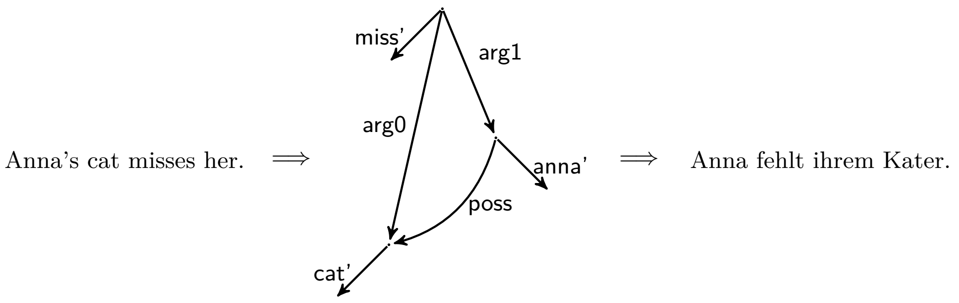


Figure 1 Example translation using semantics.

In a probabilistic setting, our goal is first to predict a graph g from a source string s, and then to predict a target string t from g, giving us a model $p(t, g|s) = p(t|g)p(g|s)$. Jones et al. [1] suggest solving this with a pair of synchronous grammars, each defining a probabilistic relation on string/ graph pairs. Given a language of source strings $\mathcal{L}_s$, a language of source graphs $\mathcal{L}_g$, a language of target graphs $\mathcal{L}_{g'}$ and a language of target strings $\mathcal{L}_t$, these grammars define relations $R \subseteq \mathcal{L}_s \times \mathcal{L}_g$ and $R' \subseteq \mathcal{L}_{g'} \times \mathcal{L}_t$. From these relations, we can formally define translations as the sets of semantically equivalent strings: the set of all translation pairs is $\{s, t | \exists g : s, g \in R \wedge g, t \in R'\}$. Hence we must be able to efficiently compute the intersection of the graph languages, $\mathcal{L}_g \cup \mathcal{L}_{g'}$. Compositions of this kind are widely used on string data in current speech and machine translation models, where they can be implemented using compositions of finite-state transducers [2, 3].

Our goal is to define similarly composable probabilistic graph languages. If our graph grammar is context-free (in the sense that its productions are associative and commutative), then we can attach normalized weights to each production to define a probability distribution over the set of graphs that it generates. We can also define its productions to be isomorphic to a (string) context-free grammar, enabling us to define probabilistic relations on strings and graphs, as desired. Although grammars and automata on graphs are much less well-studied than they are on strings and trees, two candidate formalisms have recently been identified: *hyperedge replacement grammar* (HRG), described by Drewes et al. [4] and studied by Chiang et al. [5]; and *DAG automata*, introduced by Kamimura and Slutzki [6] to model type-0 derivations and studied by Quernheim and Knight [7].

Unfortunately, neither HRG nor DAG automata satisfy our desired criteria. Although HRGs are context-free and can easily be made probabilistic, they are not closed under intersection – the emptiness of intersection is undecidable, as are many other useful questions on HRGs. In contrast, DAG automata are closed under intersection but are not context-free –so it is unknown how to make them probabilistic in a way that would yield practical algorithms, although weighted algorithms have been developed that do not define proper probability distributions. The session addressed the question of whether a fully probabilistic treatment of DAG automata is possible. Informally, the fundamental problem stems from a confluence of properties:

1. DAG automata were initially designed to model type-0 derivations, so they are non context-free. This means that rewriting operations are not commutative: applying a single rewrite to the frontier states of a DAG automaton may change the set of rewrites that are applicable to remaining states.
2. One plausible way to define probabilistic rewriting systems is to define a probability distribution over all possible rewriting steps that can be applied to a particular configuration of the system. However, property 1 implies that this probability distribution cannot factor over the frontier states of the automaton: the rewriting of any particular state is not independent of rewriting other states. Most likely, this means that a probability distribution over derivations of a DAG automaton cannot factor over its productions, as it can in context-free formalisms. So, probability distributions must depend on the complete configuration of the automaton.
3. Since probability depends on the state of the automaton, steps of a probabilistic parsing algorithm must also know the state of the automaton before and after application of a particular rewrite. Unfortunately, the naïve algorithm for this is to define a parsing state for every cut of an input graph, and the number of cuts is exponential.

Hence, it appears that a natural definition of probabilistic graph automata leads to exponential probabilistic recognition algorithms, which is actually worse than in the weighted case. It is an open problem whether a better solution is possible.

References

1 Bevan Jones, Jacob Andreas, Daniel Bauer, Karl Mortiz Hermann, and Kevin Knight. Semantics-based machine translation with hyperedge replacement grammars. In *Proceedings of COLING*, 2012.

2 Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. Speech recognition with weighted finite-state transducers. In Larry Rabiner and Fred Juang, editors, *Handbook on Speech Processing and Speech Communication, Part E: Speech recognition.* Springer, 2008.

3 Cyril Allauzen, William Byrne, Adria de Gispert, Gonzalo Iglesias, and Michael Riley. Pushdown automata in statistical machine translation. *Computational Linguistics*, 2014.

4 Frank Drewes, Hans-Jörg Kreowski, and Annegret Habel. Hyperedge replacement graph grammars. In Grzegorz Rozenberg, editor, *Handbook of Graph Grammars and Computing by Graph Transformation*, pages 95–162. World Scientific, 1997.

5 David Chiang, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, Bevan Jones, and Kevin Knight. Parsing graphs with hyperedge replacement grammars. *Proceedings of ACL*, 2013.

6 Tsutomu Kamimura and Giora Slutzki. Parallel and two-way automata on directed ordered acyclic graphs. *Information and Control*, 49(1):10–51, 1981.

7 Daniel Quernheim and Kevin Knight. Towards probabilistic acceptors and transducers for feature structures. In *Proceedings of the Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2012.

4.6 Properties of (Various Types of) DAG Automata

Frank Drewes

Graph structures in NLP are often directed acyclic graphs (DAGs). In particular, representations of meaning such as Abstract Meaning Representations (AMRs) may safely be assumed to be DAGs. Therefore, automata working on DAGs, similar to the well-known concept of tree automata, could be of great usefulness if they (a) allow to recognize DAG languages that are of interest from an NLP point of view and (b) exhibit useful closure properties as well as algorithmic properties. Ideally, such a theory should also provide corresponding notions of transducers, i.e., automata with output. Unfortunately, not much work seems to have been done in this field. The oldest approach [1, 2] was explicitly invented to capture the nature of derivations in type-0 Chomsky grammars and is thus much to powerful. The approach of [3, 4] considers only DAGs that are trees with maximally shared subtrees. This is clearly inappropriate for processing AMRs or other representations of meaning because such DAGs cannot contain isomorphic sub-DAGs. An interesting approach from the NLP point of view seems to be the one proposed in [5], but it has some disadvantages:

- It is capable of generating NP-complete DAG languages, and thus parsing is too inefficient.
- The recognized DAG languages have non-context-free path languages in the worst case, whereas it is reasonable to assume that the set of all valid meaning representations (in any reasonable formalism such as AMRs) have regular path languages.
- The class is not closed under complementation.

It is unclear how important the last point is, but from a formal point of view closedness under complementation would certainly be a positive property. Hence, it seems that [5] may be used as a starting point and source of inspiration, but does not provide a satisfactory solution in itself.

Properties such as those mentioned above left aside, there are properties of the DAGs being worked on that are of interest:

- Incoming and outgoing edges of a node in a DAGs may be ordered or unordered. In the ordered case, this order may be defined globally by placing an order on the nodes of a DAG. This type of DAGs is considered in [1, 2], and it seems that the global order (and the fact that it defines the local order at each node) is responsible for the power of these automata. From the point of view of NLP, and in particular from the point of view of AMRs, it seems that outgoing edges of a node should at least be partially ordered (or labeled, which is equivalent) whereas incoming edges should be unordered. Thus, an ideal model should be able to capture both.
- DAGs can be ranked or unranked, where ranked means that every node has a predefined number of incoming and outgoing edges determined by its label. Meaning-representing DAGs are usually unranked in the sense that there is no a priori bound on the number of incoming and outgoing edges of a node. From this point of view, unranked seem to be more appropriate. However, given the fact that it is difficult to devise a model that is both general and has nice properties, it may be reasonable to consider the ranked case, anyway, at least as a first step towards a more general model.

More recently, Quernheim and Knight [6] proposed a new notion of DAG automata based loosely on the approach by Kamimura and Slutzki. Inspired by that approach, Chiang, Drewes, Gildea, Lopez, and Satta have started to work on a restricted model of ranked DAG automata that has been discussed during the Dagstuhl Seminar. These automata have a decidable emptiness problem and regular path languages, the latter being an insight obtained during the discussions of the Dagstuhl Seminar (with considerable help of J. Björklund and A. Maletti). Unfortunately, even these rather restricted DAG automata can recognize NP-complete DAG languages.

References

1 T. Kamimura and G. Slutzki. Parallel and two-way automata on directed ordered acyclic graphs. *Information and Control*, 49:10–51, 1981.

2 T. Kamimura and G. Slutzki. Transductions of dags and trees. *Mathematical Systems Theory*, 15:225–249, 1982.

3 W. Charatonik. Automata on DAG representations of finite trees. Research Report MPI-I-1999-2-001, MPI Saarbrücken, 1999.

4 S. Anantharaman, Narendran P., and M. Rusinowitch. Closure properties and decision problems of dag automata. *Information Processing Letters*, 94:231–240, 2005.

5 L. Priese. Finite automata on unranked and unordered DAGs. In *Proc. 11th Int'l Conf. on Developments in Language Theory (DLT 2007)*, volume 4588 of *Lecture Notes in Computer Science*, pages 346–360, 2007.

6 Daniel Quernheim and Kevin Knight. Towards probabilistic acceptors and transducers for feature structures. In *Proc. 6th Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 76–85. Association for Computational Linguistics, 2012.

4.7 Typical or Desirable Features of Graphs in NLP

Stephan Oepen, Mark Steedman, Frank Drewes, Laura Kallmeyer, and Daniel Bauer

GraphaLogue: Surveying Graph Banks

With a growing community interested in graph processing, it would seem worthwhile to compile a survey of existing graph banks, i.e. collections that pair natural language data with graph-structured representations of linguistic analysis (syntactic, semantic, or otherwise). Initially at least, such resources will primarily be annotations of meaning (in various interpretations), but in principle other types of linguistic annotations that transcend tree-structured analyses should be included. For this survey, one should define and quantify relevant structural properties such as:

- reentrancy
- edge density
- connectedness
- rootedness
- acyclicity
- functionality of edge labels
- treewidth
- page number

Likewise, it would be helpful to try and characterize the (purpose and) contents of the annotations, sentence and token counts, and licensing.

Besides providing a catalogue of available resources, this survey could also develop into a quantitative and qualitative comparison of representations. To the extent that we can tease apart different layers of semantic construction – for example differentiate grammatical control from anaphoric binding – it would also seem useful to characterize annotated resources in terms of which of these phenomena they target.

Marco Kuhlmann and Stephan Oepen would be happy to try and coordinate an initial catalogue (or “graphalogue”, if you will) construction. With a bit of luck, this could evolve into a community resource (e.g. on the ACL Wiki) and enjoy collective maintenance over time. Obvious existing resources to look at include:

- AMR Bank
- SemEval 2014 and 2015 Semantic Dependency Parsing (SDP) graphs
- Universal Dependencies
- Semantic Dependencies in CCGBank

Semantic Representations Adapted to Inference

A central problem in using semantic parsing for tasks like open-domain question answering and (more obviously) machine translation is that the sentence in unseen text or the target language may take a form that is not the one most directly suggested by the question or source. (Thus the answer to the question “Did Google buy YouTube” may or may not be answered by “Google bought every company”, “Google’s purchase of YouTube”, “Google subsidiary YouTube,” “L’acquisition de YouTube par Google”, etc.) Most semantic parsers are too specific to the original form of language to allow the question to be settled without lengthy inference of a kind that is not usually affordable – hence the habit of search engines of returning multiple pages containing such phrases in the hope that the user can work out the answer by reading them. Often the user can do this, but sometimes they cannot. (Try "What are Miles Davis Recordings without Fender-Rhodes piano?").

Since the Generative Semantics and Conceptual Dependency Semantics of the '70s, there have been many attempts to produce a Universal or "Natural" Semantics underlying paraphrase and common-sense entailment relations between expressions, including attempts to use such a representation or "Interlingua" for Machine Translation. However, none of them have got very far beyond language specificity, even when multiple languages have been considered, and there has recently been a move to recast the problem in machine language terms as that of learning a "hidden" set of semantic relations from large amounts of unannotated text.

Two main approaches were distinguished. The most radical is the "pure distributional" approach, based on collocations of content words represented as dimensionally reduced vectors, with linear algebraic operations like vector addition and multiplication substituting for traditional semantic composition in forming meanings for larger structures, often under the control of dependency parsers [1]. Such representations have some striking advantages, such as being able to simultaneously represent multiple ambiguous readings, which may be disambiguated by linear algebraic composition.

Such representations are capable of representing the similarity of concepts as closeness in the multidimensional vector space, and hence of detecting the similarity between paraphrases in source and target. However, it is hard to see how they can be interfaced with logical semantics. In particular, there does not seem to be a vector or linear algebraic representation for operators such as negation. A second kind of distributional semantic seeks to identify relations of paraphrase and entailment directly in unseen text, using parsing or "machine reading", and to build such logical relations into natural language semantics directly, treating paraphrases as clusters and entailment as logical conjunction [2].

The latter approach has been shown to to be capable of capturing linguistically significant entailments, such as that "McCain regrets that he wasn't nominated" entails that "McCain wanted to be nominated", which could be used to acquire the information that the semantics of verbs like "want" includes an implicit controlled subject of the complement "to be nominated".

There was further discussion of the vector based alternative, and whether recent developments using "Deep Learning", or multi-layer perceptrons or Boltzmann machines using backpropagation training at a vast scale would render interaction with logicist semantic unnecessary. It was generally felt that the radical approach was probably incompatible with any form of structured representation such as AMR, but that the paraphrase and entailment based clustering approaches were entirely compatible and might even be helpful.

Logical Operators and Quantification

Currently logical operators and quantification are not represented explicitly in most graph based meaning representations like AMR, although they have been used traditionally to represent linguistic meaning. We discussed if we can and should develop a graph-based representation that has a translation into some logical form. Such a representation would allow for semantic inference but can be difficult to annotate, especially when annotators have to resolve possible scope orders.

In AMR, some vertices represent existentially quantified variables (instances of concepts, such as events and objects), which also interferes with the scope order. Universal quantification and negation is expressed using additional edges. Negation attaches to the root of the sub-DAG it takes scope over. Disjunction is represented using an additional "or" node. Conjunction is sometimes represented this way, but only if it is mentioned explicitly in the sentence (for instance, conjunction of two events). Unfortunately scope is not adequately represented in the structure of the AMR graph. Logical operators are generally only represented if they are mentioned explicitly in the sentence described by the AMR.

Existentially quantifying all event nodes allows us to account for the reading with narrow scope of the existentially quantified event in the following sentence: *"Most of the students have read the book."* In this reading there is one *reading* event for each student. Without an explicit representation of scope the reading in which the existential quantifier takes broad scope is lost. According to this reading there is a single *reading* event of the book, for instance if the students take turns.

A better solution might be to represent scope outside the graph structure, on a separate representation level. Scope could then be left underspecified if it is ambiguous. Annotators could either annotate the specific order of quantifiers, for instance

> *"Most people know two languages"*: most people > exists knowing > two languages

or they could specify dependencies in Skolem terms, such as

> *"most people know two languages"*: language(people), know(most)

The second option appears to be more intuitive for the annotator and it would allows annotators to leave out dependencies they are not sure of (or that are actually ambiguous or independent of each other). The result would be an underspecified representation of quantifier scope that allows for reasoning.

We leave the specification of a graph-based representation that addresses these issues and an annotation scheme for future work.

Linguistic Phenomena that "Cause" Reentrancies in AMRs

The following "causes" of reentrancies could be discovered by looking at a variety of AMRs:

- anaphora such as pronouns
- control (-like) structures such as in *"John promised me to paint the wall."* (John will be the $\texttt{arg}_0$ of `paint`.)
- multiple participles (`NB` "front-loading" assigns $\texttt{arg}_1$)
- VP coordination / shared conjuncts
- implicit arguments ("he" is $\texttt{arg}_1$ of "hospital treatment")
- relative clauses

Argument Sharing Exemplars

Besides inspecting concrete example annotations in the AMR bank, a more general inventory of phenomena that cause reentrancies in semantic graphs due to argument sharing is of interest. We aim at creating such a collection of argument sharing exemplars that will be a useful resource for linguistic analysis and grammar developing. The following list is a first collection of such phenomena.

- *Grammatical Control*
  Kim wants to sing. ; subject-equi
  Kim wants Sandy to sing. ; raising-to-object (no reentrancy)
  Kim persuaded Sandy to sing. ; object-equi
  Sandy seemed to sing. ; raising-to-subject (no reentrancy)
  Kim promised to seem to be competent.

- *Passives*
  Kim wants to be heard.
- *Nominalizations*
  Kim made a promise to sing.
  Kim has a desire to sing.
  Kim's plan is to sleep more.
  Kim showed signs of recovery.
- *Modification*
  The drying and washing machine broke.
  The washing machine is expensive.
- *Coordination*
  Kim drank wine and ate pizza.
  Kim showed Sandy and sold Tony the wine.
  Kim showed and Sandy sold the wine.
  Kim sold the wine and Tony the pizza.
  Kim and Sandy sang.
  his arms and feet. ; interaction with possessive determiner
  Kim sang on Monday and on Tuesday.
  Kim wanted and expected to sing.
- *Reflexive Pronouns and reciprocals*
  Kim saw herself.
  Kim and Sandy admired each other.
- *Relative Clauses*
  Kim ate the pizza that Tony had sold.
  Kim saw the boy whose father sold the pizza.
  Kim arrived on the day that Sandy arrived.
- *Secondary Predicates*
  Kim placed the book on the table.
  Kim wiped the table clean.
  Kim left Sandy without paying.
  Kim met Sandy singing.
  Kim met Sandy drunk.

Stephan Oepen and Laura Kallmeyer plan to extend this initial list of examples in the near future to a more complete resource called *SemSharE* – *Sem*antic Argument *Shar*ing *E*xemplars.

References

1 Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

2 Mike Lewis and Mark Steedman. Combining distributional and logical semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192, 2013.

5 Seminar Program

Introductory Presentations

The seminar started by two introductory presentations on the use of graphs for representing meaning:

1. Kevin Knight. *Mapping English Strings to Reentrant Semantic Structures*
2. Marco Kuhlmann. *Properties of the SemEval-2015 Data Sets*

Tutorial on Graph Transformation

In addition, a longer tutorial on the theory of graph transformation, divided into several parts and spread out over 3 days, was given by Frank Drewes.

Working Groups

Open Space group discussions were held on Monday, Tuesday, and Thursday. The following topics were discussed:

Monday

Session 1

- Node replacement grammar for semantics. What does treewidth mean (in NLP)?
- (Greedy) parsing algorithms for graphs. Restricted Graph formalisms that allow efficient parsing. Grammarless parsing
- DAG automata. Generating and recognizing AMRs
- Universal dependencies. Syntax–semantics interface. Integrating logical operators into semantic graphs. Comparison of graph banks

Session 2

- Characterizing graphs produced by various grammar formalisms. Tree-to-DAG transformations
- What happens to HRG when we impose a linear order on the nodes?
- Identifying a “good” generator set of graph operations
- Multitape graph transducers (insufficient output)
- Convolution kernels for directed acyclic graphs and other similarity measures

Tuesday

Session 1

- DAG automata. Generating and recognizing AMRs
- Grammarless approaches to graph parsing
- Dress up an inventory of the graphs we want to have; what are the consequences for required HRG?
- Hyperedge Unification Grammars

Session 2

- Characterizing graphs produced by various grammar formalisms. Tree-to-DAG transformations
- Practical parsing of HRG
- Grammarless generation

Thursday

Session 1

- Types and causes of reentrance. Linguistically relevant graph grammars. An inventory of the graphs that we want to have; consequences regarding constraints on HRG
- An argmax-algorithm for pagenumber-2 graphs. Relationship to dependency parsing. Eisner-like algorithms and treewidth (or other notions of x-width)

Session 2

- How can semantic representations support inference?
- Restricted but fast HRG parsing

Session 3

- Graphalog – a catalogue of graph banks/Example-based comparisons of graph representations (AMR, SDP etc.)
- Inducing synchronous context-free string ↔ graph transformations. Graph-string alignment algorithms

Evening session

- How do we assign probabilities to graphs? Probabilistic non-context-free graph rewriting

Note that the remainder of this report is not structured according to the list above. Instead, we have tried to structure the major outcomes of the discussions and present them in an appropriate way in order to serve as a reference for future work.

Closing Session

Friday morning was devoted to a general recap and an evaluation of the seminar. The result of the evaluation was very positive; it was decided to consider the possibility of applying for a follow-up workshop after a couple of years when the community and its research area had taken shape, which to a significant extent would be thanks to this Dagstuhl Seminar.

Participants

- Daniel Bauer
Columbia University, US
- Suna Bensch
University of Umeå, SE
- Henrik Björklund
University of Umeå, SE
- Johanna Björklund
University of Umeå, SE
- David Chiang
Univ. of Notre Dame, US
- Frank Drewes
University of Umeå, SE
- Petter Ericson
University of Umeå, SE
- Daniel Gildea
University of Rochester, US
- Karl Moritz Hermann
Google DeepMind – London, GB
- Berthold Hoffmann
Universität Bremen, DE
- Peter Jonsson
Linköping University, SE
- Laura Kallmeyer
Heinrich-Heine-Universität Düsseldorf, DE
- Kevin Knight
USC – Marina del Rey, US
- Alexander Koller
Universität Potsdam, DE
- Marco Kuhlmann
Linköping University, SE
- Adam Lopez
University of Edinburgh, GB
- Andreas Maletti
Universität Stuttgart, DE
- Jonathan May
USC – Marina del Rey, US
- Mark Minas
Universität der Bundeswehr – München, DE
- Joakim Nivre
Uppsala University, SE
- Stephan Oepen
University of Oslo, NO
- Detlef Plump
University of York, GB
- Giorgio Satta
University of Padova, IT
- Natalie Schluter
University of Copenhagen, DK
- Mark Steedman
University of Edinburgh, GB
- Christoph Teichmann
Universität Potsdam, DE
- Brink van der Merwe
University of Stellenbosch, ZA
- Heiko Vogler
TU Dresden, DE
- Daniel Zeman
Charles University – Prague, CZ

Report from Dagstuhl Seminar 15131

Normative Multi-Agent Systems

Edited by

Amit K. Chopra[1], Leon van der Torre[2], Harko Verhagen[3], and Serena Villata[4]

1 **Lancaster University, GB,** a.chopra1@lancaster.ac.uk
2 **University of Luxembourg, LU,** vivianetorressilva@gmail.com
3 **Stockholm University, SE,** verhagen@dsv.su.se
4 **INRIA Sophia Antipolis – Méditerranée, FR,** serena.villata@inria.fr

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15131 "Normative Multi-Agent Systems". Normative systems are systems in the behavior of which norms play a role and which need normative concepts in order to be described or specified. A normative multi-agent system combines models for normative systems (dealing for example with obligations, permissions and prohibitions) with models for multi-agent systems. Normative multi-agent systems provide a promising model for human and artificial agent coordination because they integrate norms and individual intelligence. They are a prime example of the use of sociological theories in multi-agent systems, and therefore of the relation between agent theory—both multi-agent systems and autonomous agents—and the social sciences—sociology, philosophy, economics, legal science, etc. The aim of this Dagstuhl Seminar was to feature two fresh themes in broader computing and software engineering: social computing and governance. These themes are highly interdisciplinary, bringing together research strands from computing, information sciences, economics, sociology, and psychology. Further there is considerable excitement about these areas in academia, industry, and public policy organizations. Our third theme was agreement technologies, a more traditional topic but nonetheless relevant for the NorMAS community. A norm is a fundamental social construct. Norms define the essential fabric of a society. Our purpose in this seminar was to explore the connections of norms to each of the themes, especially from a computational perspective. Moreover, the seminar has been conceived for the writing of a volume titled "Handbook of Normative Multi Agent Systems" aimed to become a standard reference in the field and to provide guidelines for future research in normative multi-agent systems.

Seminar March 22–27, 2015 – http://www.dagstuhl.de/15131
1998 ACM Subject Classification I.2.11 Distributed Artificial Intelligence
Keywords and phrases Normative Multiagent Systems, Autonomous agents and Multiagent systems, Social Computing, Governance, Agreement Technologies, Norms
Digital Object Identifier 10.4230/DagRep.5.3.162

Normative Multi-Agent Systems, *Dagstuhl Reports*, Vol. 5, Issue 3, pp. 162–176
Editors: Amit K. Chopra, Leon van der Torre, Harko Verhagen, and Serena Villata
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Amit K. Chopra
Leon van der Torre
Harko Verhagen
Serena Villata

The multi-disciplinary workshop on Normative Multi Agents attracted leading international scholars from different research fields (e.g., theoretical computer science, programming languages, cognitive sciences and social sciences). The workshop was organized as follows: the organizers identified three relevant themes of research covering a wide and comprehensive spectrum of topics in the filed of Normative Agents, namely Social Computing, Governance, and Agreement Technologies. In the months preceding the workshop the chairs collected material from the participants. During the first day each participant present herself to the audience, and the chairs introduced the goal of the seminar, i.e., writing an handbook of Normative Multiagent Systems based on the roadmap produced during the previous edition of the Seminar, and the discussions during the current one. The participants were divided in groups corresponding to the areas identified as relevant in the field of Normative Multiagent Systems. Four invited talks have been proposed by scholars from different areas in the field, targeting in particular the three themes of the Seminar and an overview about Normative Multiagent Systems. The format was well received by the participants and conducive to discussion. It gave them the opportunity to give very focused presentations while keeping the audience attention. During the morning sessions, we started with an invited talk and we continued with short presentations by the Seminar participants about their personal contribution to Normative Multi-Agents (plus some time for QA). The afternoon sessions, other the contrary, were dedicated to group work and group discussions. The aim of these sessions was to build consensus material of the specific topics and to identify fundamental research directions. The material is expected to be refined and to be articulated in chapters intended as a first step for the development for the handbook for this emerging area of computer-science with close interactions with other disciplines.

Results

During the seminar, participants split in different working groups, centered around discussion themes relevant to NorMAS. Each working group was further divided into smaller working groups, each of which worked on specific topics. In the following paragraphs there is a summary of the discussion held by each working group.

Logic and reasoning. This theme included subgroups on topics such as *deontic logic*, *argumentation*, *computation approaches*, *motivational attitudes*, *social games*, and *emotions*.

Modeling. This theme included subgroups on issues such as *taxonomies*, *law*, *conflicts*, and *norm dynamics*.

Engineering. This theme included subgroups on themes such as *interactions*, *agent programming*, *agent architecture*, *data-driven norms*, *institutions and technology*, and *reference architectures*.

Simulation. This theme discussed issues of simulating multiagent systems to understand norm dynamics such as *emergence* and *diffusion*.

Applications. This theme included subgroups on killer applications for norms. Identified applications included *governance*, *audit control*, *cybersecurity*, *jurisinformatics*, and *sociotechnical systems.*

Each subgroup presented its findings twice to the entire seminar. Each subgroup identified past work, connections to other subgroups, and future work. Based on their presentations, we decided that each subgroup should write a chapter on its topic. This chapter will become part of a Handbook of Normative Multiagent Systems. This is in line with the roadmap produced during the previous edition of the Seminar and the discussions held during the present Seminar. The handbook will be an authoritative and detailed introduction for anyone seeking information on normative multiagent systems. The handbook will give a historical overview, present a survey of established techniques and open challenges, and discuss applications and directions. Our aim is to have to handbook sent for publication in a year's time. We already have a publisher lined up (College Publications).

2 Table of Contents

Executive Summary

Invited talks

Overview of Talks

3 Invited talks

3.1 Distributed epistemic agency, responsibility and trust in socio-technical systems

Judith Simon (IT University of Copenhagen, DK)

Contemporary practices of knowing take place in increasingly complex and dynamic socio-technical systems consisting of human and artificial agents, of people, technologies and infrastructures embedded in socio-economic environments. Given this distribution of epistemic agency, how can we ensure that agents act responsible in such knowledge practices, that trust is placed only in trustworthy agents and informational resources? In my talk, I will first outline a conception of distributed agency and relate this to notions of epistemic trust, i.e. the necessity to trust other entities or agents in our processes of knowing as well as to notions of distributed epistemic responsibility, i.e. the responsibilities of various entangled agents as recipients and providers of information. By using examples related to social computing and big data practices, I will show how individualized understandings of agency, responsibility or even knowledge are not only inadequate, but potentially harmful due to their neglect of issues of power and injustice. I will end my talk with some considerations of how such socio-technical epistemic systems could be governed to support trustworthiness, fair distributions of responsibility and responsible action.

3.2 An Overview on Normative Conflicts Detection and Resolution

Viviane Torres da Silva (IBM Research – Rio de Janeiro, BR)

A conflict between two norms occurs when the fulfillment of one norm violates another norm. When a conflict takes place the agent is unable to fulfill all norms that are active without violating at least one of them. The detection of normative conflicts is one of the main challenges in the specification of normative systems. In this talk I will present several approaches used to detect normative conflicts and, also, several techniques used to solve these conflicts.

3.3 Juris-Informatics and PROlog-based LEGal reasoning system: PROLEG

Ken Satoh (National Institute of Informatics – Tokyo, JP)

We have been doing research on "juris-informatics" which is application of informatics to legal domain. The name is made from a hope that we will make a similar impact to "bio-informatics" and show some related results to "juris-informatics" As a part of research of "juris-informatics", we implement "Japanese Ultimate Fact (JUF) theory" to simulate judge's reasoning at a civil court. JUF theory is a tool for judges to make a judgment based on burden of proof under incomplete information. We show correspondence of burden of proof and negation as failure in logic programming and we introduce a system called PROLEG which we developed using the correspondence. PROLEG consists of general rules and exceptions which directly reflect lawyers' knowledge structure in legal reasoning. Then, we show that the representation power of PROLEG is same as Answer Set Programming and that PROLEG could be applied to any other legal domains where general rules and exceptions co-exist.

3.4 Governance and accountability

Joris Hulstijn (TU Delft, NL)

In NORMAS we study conceptualizations of norm following. Ultimately these should inform the development of software tools. But such tools only work when they are embedded in the right organizational context. That is what governance is all about: the arrangements of governing. Governance structures indicate who have power over whom. However, those who are in power should be accountable for their deeds. How can we ensure accountability in a normative multi-agent system? In this lecture, I would like to tell you stories – based on research I have done or supervised – about ill-fitting governance structures that caused failure of some sort. The lesson we can draw from these stories is that it is in fact possible to institutionalize "opposition" into a governance structure to ensure a basic level of accountability.

4 Overview of Talks

4.1 Towards Distributed Support of Distributed Software Development Processes

Daniel Moldt (Universität Hamburg, DE)

Processes and structures of distributed teams are of special interest for the support by tools. Considering these teams as multi-agent systems or as multi-organization systems requires to provide an adequate conceptual modeling perspective. Within this perspective the flexible support of development processes is difficult, due to the heterogeneous requirements and hence somehow unstructured processes. The unstructuredness is however quite well structured when observing professional developers at work. My talk will give insight into a multi-agent and multi-organization based modeling perspective and how we support software development process based on social metaphors and still with the formal background of high-level Petri nets.

4.2 The Rationale behind the Concept of Goal

Guido Governatori (NICTA – Brisbane, AU) and Antonino Rotolo (University of Bologna, IT)

The paper proposes a fresh look at the concept of goal and it advances that motivational attitudes like desire, goal and intention are just facets of the broader notion of (acceptable) outcome. We propose to encode the preferences of an agent as sequences of "alternative acceptable outcomes". We study how the agent's beliefs and norms can be used to filter the mental attitudes out of the sequences of alternative acceptable outcomes. We formalize such intuitions in a novel Modal Defeasible Logic and we prove that the resulting formalization is computationally feasible.

4.3 The Complexity of Strategic Argumentation under Grounded Semantics

Antonino Rotolo (University of Bologna, IT) and Guido Governatori (NICTA – Brisbane, AU)

We study the complexity of the Strategic Argumentation Problem for 2-player dialogue games where a player should decide what move to play at each turn in order to prove (disprove) a given claim. We shall prove that this is an NP-complete problem. The proof covers Dung (1995)'s grounded semantics with structured and abstract arguments.

4.4 Reasoning with Group Norms in Software Agent Organisations

Huib Aldewereld, Virginia Dignum, and Wamberto Vasconcelos

Norms have been used to represent desirable behaviours that software agents should exhibit in sophisticated multi-agent solutions. Existing research has mostly focused on the study of norms that affect a single individual. An important open research issue refers to group norms, i.e. norms that govern groups of agents. Depending on the interpretation, group norms may be intended to affect the group as a whole, each member of a group, or some members of the group. Moreover, upholding group norms may require coordination among the members of the group. We have identified three sets of agents affected by group norms, namely, i) the addressees of the norm, ii) those that will act on it, and iii) those that are responsible to ensure norm compliance. We present a formalism to represent these, connecting it to a minimalist agent organisation model. We use our formalism to develop a reasoning mechanism which enables agents to identify their position with respect to a group norm, so as to further support agent autonomy and coordination when deciding on possible courses of action.

4.5 Indirect Normative Conflict: Conflict that Depends on the Application Domain

Viviane Torres da Silva (IBM Research – Rio de Janeiro, BR)

Joint work of Christiano Braga, Viviane Torres da Silva, and Jean Zahn

Norms are being used as a mechanism to regulate the behavior of autonomous, heterogeneous and independently designed agents. Norms describe what can be performed, what must be performed, and what cannot be performed in the multi-agent systems. Due to the number of norms specified to govern a multi-agent system, one important issue that has been considered by several approaches is the checking for normative conflicts. Two norms are said to be in conflict when the fulfillment of one norm violates the other and vice-versa. In this paper, we formally define the concept of an indirect normative conflict as a conflict between two norms that not necessarily have contradictory or contrary deontic modalities and that may govern (different but) related behaviors of (different but) related entities on (different but) related contexts. Finally, we present an ontology-based indirect norm conflict checker that automatically identifies direct and indirect norm conflicts on an ontology describing a set of norms and a set of relationships between the elements identified in the norms (behavior, entity and context).

4.6 Toward a Norms-Based Theory of Sociotechnical Systems

Amit K. Chopra (Lancaster University, GB)

Researchers and practitioners are increasingly concerned with the challenge of engineering sociotechnical systems. Healthcare, emergency response, and smart cities are examples of sociotechnical systems, and experience bears out that these systems are not easy to build and maintain. In the present paper, I discuss of some of the challenges of engineering sociotechnical systems and their potential solutions. In particular, I focus on challenges related to software engineering, distributed computing, and information and programming models. I also discuss the governance of sociotechnical systems. My proposal to address these challenges centers around the concept of norms, thereby constituting an outline of a coherent theory of sociotechnical systems. Research on norms and organizations is a strength of the multiagent systems community, which gives us a leg up in addressing the challenges of engineering complex sociotechnical systems.

4.7 Generating Legal Reasoning Structure by Answer Set Programming

Ken Satoh (National Institute of Informatics – Tokyo, JP)

Joint work of Duangtida Athakravi, Ken Satoh, Krysia Broda, and Alessandra Russo

In legal reasoning, different assumptions are often considered when reaching a final verdict and judgment outcomes strictly depend on these assumptions. In this paper, we propose an approach for generating a declarative model of judgments from past legal cases, that expresses a legal reasoning structure in terms of principle rules and exceptions. Using a logic-based reasoning technique, we are able to identify from given past cases different underlying defaults (legal assumptions) and compute judgments that (i) cover all possible cases (including past cases) within a given set of relevant factors, and (ii) can make deterministic predictions on final verdicts for unseen cases. The extracted declarative model of judgments can then be used to make automated inference of future judgments, and generate explanations of legal decisions. The rules generated by our approach can also be automatically translated into a representation compatible with the legal reasoning system PROLEG, so making our method a useful computational mechanism for generating PROLEG models from past cases.

4.8 Social Computing with 2COMM4JASON

Matteo Baldoni (University of Turin, IT)

Joint work of Matteo Baldoni, Cristina Baroglio, Federico Capuzzimati, and Roberto Micalizio

Social Computing (SC) requires agents to reason seamlessly both on their social relationships and on their goals, beliefs. We claim the need to explicitly represent the social state and social relationships as resources, available to agents. We built a framework, based on JaCaMo, where this vision is realized and SC is implemented through social commitments and commitment protocols.

4.9 Collaboration Pattern Modeling in Support of Norm Specification, Monitoring, and Preservation

Christoph Dorn (TU Wien, AT)

Collaboration-intensive environments call for technical systems that permit flexible user interactions. Rigid workflows are no suitable collaboration paradigm. As users apply various patterns such as shared artifact, social networks, client/principal, or publish/subscribe for interaction, their cooperative behavior becomes largely determined by norms. In this paper, we make the case for explicit modeling of collaboration patterns as the substrate for specifying, monitoring, and preserving norms. Describing collaboration patterns in the form of human-centric component and connector architecture views provides a means for reasoning on collaboration control, flexibility, and ultimately adaptability. We report on recent work targeting executable collaboration patterns and outline resulting synergies with norms.

4.10 Compatibility of Licenses in the Web of Data

Ho-Pun Lam and Guido Governatori (NICTA – Brisbane, AU)

While several proposal have been offered to represent licensing information through ah-hoc ontologies and patterns, only few approaches have addressed the problem of reasoning over such information. In this paper, we propose and evaluate a deontic logic semantics which allows us to define the deontic modalities of licenses, i.e., permission, obligation and prohibition, to verify the compatibilities among the deontic components of different licenses, and can compose them into a single theory if they are compatible. Based on this, heuristics for composing different deontic components of licenses are proposed, and an extension based on the SPINdle defeasible reasoner has been developed to evaluate our framework. Our result show that our approach provide a flexible and efficient solution to the problem.

4.11 Norms in criminal organizations: inside the evolution of social order

Martin Neumann (Universität Koblenz-Landau, DE)

This paper presents two simulation models about internal conflict resolution within criminal organizations. Securing compliance in the absence of state monopoly of violence makes criminal organizations a test bed for studying evolution of social order. Target systems are briefly described: One case is the Sicilian Mafia and temporary Mafia wars. The other case describes the breakdown of a criminal group in its infancy. While the Mafia has a strict hierarchical organization, the contrasting case had a flat structure. This difference corresponds to cognitive trust in the organization in case of the Mafia and affective trust in interpersonal relations in the contrasting case. This enables Mafiosi to cognitively trust the organization while affectively mistrusting other Mafiosi. This stabilizes organizational endurance. The paper ends with remarks about the insights for evolution of social order from investigating criminal organizations.

4.12 Norms and Collectives – Between Narratives, Simulations and Games

Corinna Elsenbroich (University of Surrey – Guildford, GB) and Harko Verhagen (Stockholm University, SE)

In this paper we describe a narrative of a civic resistance movement to defeat the Italian Mafia, a model comparing strategic and normative modes of reasoning in an individual and collective interpretation of an extortion racket situation and a serious game through which to collect data on the four types of behaviours used in the simulation. These three elements will be used to discuss the reflexive and iterative nature of simulation research, in particular in a field as elusive as changing motivations of agents. Finally we will describe how online games can be used to calibrate the model parameters and to accomplish social change.

4.13 The Role of Power in Legal Compliance

Robert Muthuri and Llio Humphreys (University of Turin, IT)

Joint work of Sepideh Ghanavati, Robert Muthuri, Andre Rifaut, Llio Humphreys, and Guido Boella

Powers constitute a significant foundation for the law as we know it yet their role has largely been neglected in requirements engineering in favour of more familiar deontic notions. We therefore explore the different conceptualizations of legal power to facilitate their incorporation in modelling the legal requirements. We apply our analysis to the legal-urn framework.

4.14 Distributed Rule-Based Agents in Rule Responde

Adrian Paschke (FU Berlin, DE)

Rule Responder is a rule-based multi-agent framework in which agents run platform-specific rule engines as distributed inference services. An important aspect for the agent communication is the use of common standardized rule interchange format. In this paper we introduce core capabilities of Reaction RuleML 1.0 for rule interchange and agent communication, supporting functionalities such as knowledge interface declarations with signatures, modes, and scopes; distributed knowledge modules with static and dynamic scopes enabling imports and scoped reasoning within metadata-based scopes (closed constructive views) on the knowledge base; messaging reaction rules enabling conversation-scope based interactions between agents interchanging queries, answers, and rulebases; and evaluation and testing of interchanged knowledge bases with intended semantic profiles and self-validating test suites. We demonstrate these Reaction RuleML 1.0 capabilities with our proof-of-concept implementation, the Rule Responder agent architecture and the Prova 3.0 rule engine.

4.15 Expressing Access Policies and Regulations for Linked Data using ODRL 2.1

Axel Polleres (Wirtschaftsuniversität Wien, AT)

Joint work of Simon Steyskal and Axel Polleres

Together with the latest efforts in publishing Linked (Open) Data, legal issues around publishing and consuming such data are gaining increased interest. Particular areas of interest include (i) how to define more expressive access policies which go beyond common licenses, (ii) how to introduce pricing models for online datasets (for non-open data) and (iii) how to realize (i)+(ii) while providing descriptions of respective meta data that is both human readable and machine processable. In this paper, we show based on different examples that the Open Digital Rights Language (ODRL) Ontology 2.1 is able to address all previous mentioned issues, i.e. is suitable to express a large variety of different access policies for Linked Data. By defining policies as ODRL in RDF we aim for (i) higher flexibility and simplicity in usage, (ii) machine/human readability and (iii) fine-grained policy expressions for Linked (Open) Data.

4.16 From Anarchy to Monopoly: How Competition and Protection Shaped Mafia's Behavior

Luis Gustavo Nardin (LABSS – ISTC – CNR – Rome, IT)

Joint work of Luis Gustavo Nardin, Giulia Andrighetto, and Rosaria Conte

Mafia-like organizations are highly dynamic and organized criminal groups characterized by their extortive activities that impact societies and economies in different modes and magnitudes. This renders the understanding of how these organizations evolved an objective of both scientific and application-oriented interests. We propose an agent-based simulation model – the Extortion Racket System model – aimed at understanding the factors and processes explaining the successful settlement of the Sicilian Mafia in Southern Italy, and which may more generally account for the transition from an anarchical situation of uncoordinated extortion to a monopolistic social order. Our results show that in situations of anarchy, these organizations do not last long. This indicates that a monopolistic situation shall be preferred over anarchical ones. Competition is a necessary and sufficient condition for the emergence of a monopolistic situation. However, when competition is combined with protection, the resulting monopolistic regime presents features that make it even more preferable and sustainable for the targets.

4.17 An Abstract Formal Model for Normative Multiagent Systems

Munindar P. Singh (North Carolina State University – Raleigh, US)

Norms provide an elegant basis for modeling and realizing interactions between autonomous parties. The subtle interplay between norms and the structure of a normative multiagent system (MAS) is not adequately understood. We propose a formal model that synthesizes key factors including identity, credentials, naming, autonomy, authority, privileges and liabilities, and forming and disbanding collaborations. This model is abstract and independent of specific norm languages. We demonstrate its power by capturing a variety of real-life cases.

4.18 Friday Dropin Talks

A number of participants gave shorter dropin talks on Friday. Dov Gabbay and Victor Rodriguez Doncel gave a talk on licenses and reasoning; Simon Caton gave a talk on his work on identifying user sentiments in social media; Julian Padget gave two talk on policies and institutions; Robert Muthuri gave a talk on modeling legal concepts; Pablo Noriega gave a talk on institutions and technology.

Participants

- Huib Aldewereld
  TU Delft, NL
- Diego Agustin Ambrossio
  University of Luxembourg, LU
- Matteo Baldoni
  University of Turin, IT
- Simon Caton
  KIT – Karlsruher Institut für Technologie, DE
- Amit K. Chopra
  Lancaster University, GB
- Rob Christiaanse
  TU Delft, NL
- Silvano Colombo Tosatto
  University of Luxembourg, LU
- Célia da Costa Pereira
  University of Nice, FR
- Christoph Dorn
  TU Wien, AT
- Hein Duijf
  Utrecht University, NL
- Corinna Elsenbroich
  Univ. of Surrey – Guildford, GB
- Dov M. Gabbay
  King's College London, GB
- Aditya K. Ghose
  University of Wollongong, AU
- Guido Governatori
  NICTA – Brisbane, AU
- Joris Hulstijn
  TU Delft, NL
- Llio Humphreys
  University of Turin, IT
- Franziska Klügl
  University of örebro, SE
- Ho-Pun Lam
  NICTA – Brisbane, AU
- Beishui Liao
  Zhejiang University, CN
- Daniel Moldt
  Universität Hamburg, DE
- Robert Muthuri
  University of Turin, IT
- Luis Gustavo Nardin
  LABSS – ISTC – CNR – Rome, IT
- Martin Neumann
  Universität Koblenz-Landau, DE
- Pablo Noriega
  IIIA – CSIC – Barcelona, ES
- Julian Padget
  University of Bath, GB
- Adrian Paschke
  FU Berlin, DE
- Gabriella Pigozzi
  University Paris-Dauphine, FR
- Axel Polleres
  Wirtschaftsuniversität Wien, AT
- Livio Robaldo
  University of Turin, IT
- Victor Rodriguez Doncel
  Technical Univ. of Madrid, ES
- Antonino Rotolo
  University of Bologna, IT
- Ken Satoh
  National Institute of Informatics – Tokyo, JP
- Judith Simon
  IT Univ. of Copenhagen, DK
- Munindar P. Singh
  North Carolina State University – Raleigh, US
- Xin Sun
  University of Luxembourg, LU
- Viviane Torres da Silva
  IBM Research – Rio de Janeiro, BR
- Leon van der Torre
  University of Luxembourg, LU
- Wamberto Vasconcelos
  University of Aberdeen, GB
- Harko Verhagen
  Stockholm University, SE
- Serena Villata
  INRIA Sophia Antipolis – Méditerranée, FR